U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Ciências
ULisboa

A genome-wide perspective on the evolutionary trajectory of wolves in human-dominated landscapes

Diana Carina da Costa Lobo

Tese de Doutoramento apresentada à
Faculdade de Ciências da Universidade do Porto

Biodiversidade, Genética e Evolução

2023



A genome-wide perspective on the evolutionary trajectory of wolves in human-dominated landscapes

Diana Carina da Costa Lobo

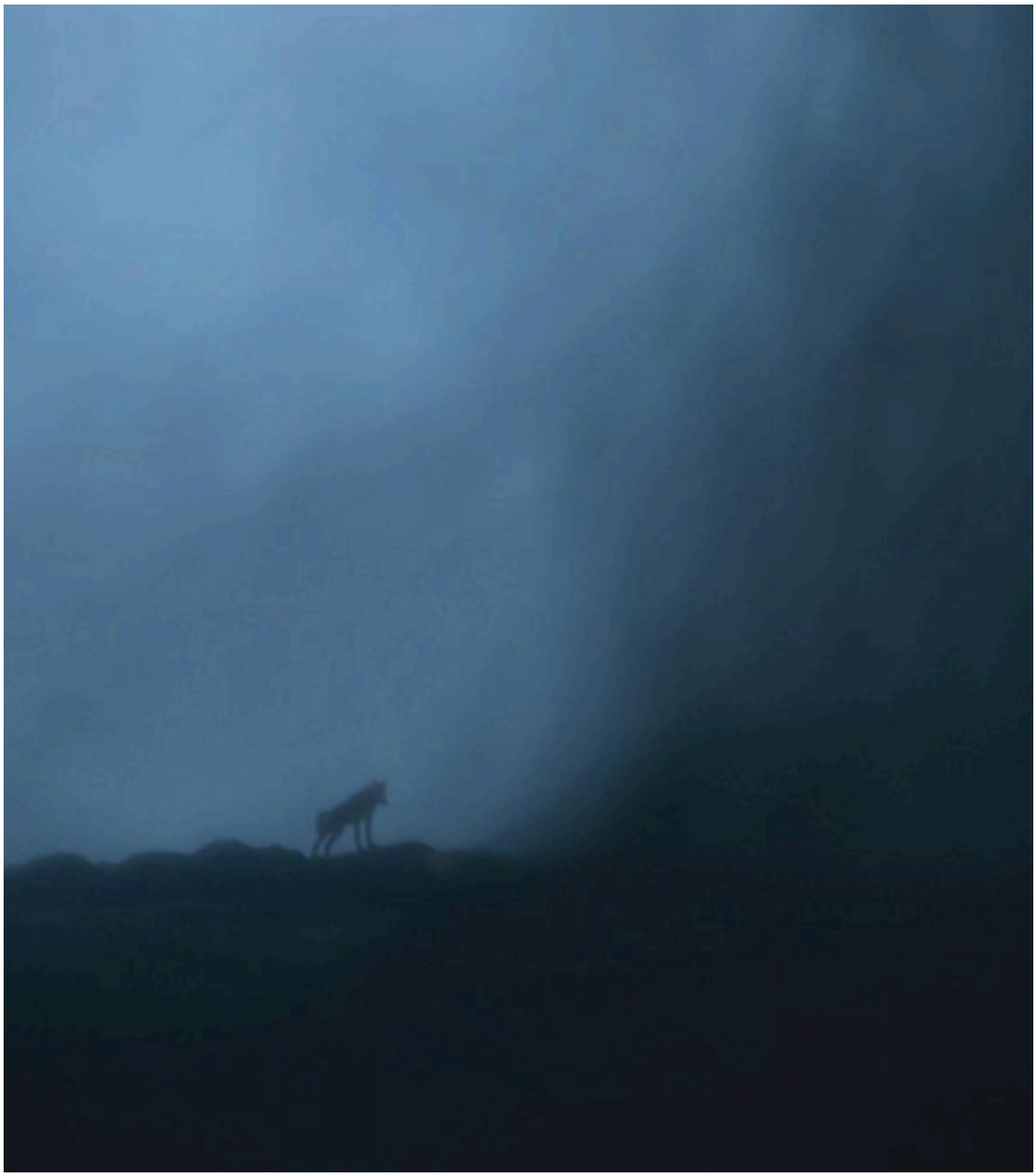
Programa Doutoral em Biodiversidade, Genética e Evolução

Departamento de Biologia, Faculdade de Ciências da Universidade do Porto

2023

Orientador:
Dr. Raquel Godinho, Investigadora Principal, CIBIO-InBIO/BIOPOLIS
Professora Auxiliar Convidada, Universidade do Porto

Coorientador:
Dr. John Patrick Archer, Investigador Principal, CIBIO-InBIO/BIOPOLIS



Iberian wolf in Peneda-Gerês National Park, Portugal. Photo credits: Carlos Pontes.

Ao meu avô.

Nota Prévia

A Faculdade de Ciências da Universidade do Porto foi a instituição de origem da candidata, tendo o trabalho apresentado nesta dissertação sido orientado pela Doutora Raquel Godinho, Professora Auxiliar Convidada do Departamento de Biologia da Faculdade de Ciências da Universidade do Porto e Investigadora do Centro de Investigação em Biodiversidade e Recursos Genéticos (CIBIO-InBIO)/BIOPOLIS, Program in Genomics, Biodiversity and Land Planning, e coorientado pelo Doutor John Patrick Archer, Investigador do CIBIO-InBIO/BIOPOLIS. O trabalho foi realizado na íntegra no CIBIO-InBIO/BIOPOLIS, no âmbito do programa doutoral em Biodiversidade, Genética e Evolução (BIODIV).

A candidata foi financiada pela Fundação para a Ciência e a Tecnologia (FCT) através da atribuição de uma bolsa de doutoramento (PD/BD/132403/2017) no âmbito do programa doutoral BIODIV. Todo o trabalho apresentado foi financiado pela FCT, através do projeto de investigação "WOLFDOG" (PTDC/BIA-EVF/2460/2014), e parcialmente pelo projeto de investigação PTDC/BIA-EVL/29115/2017.

Agradecimentos / Acknowledgments

Estes últimos seis anos em que estive a desenvolver o meu projeto de doutoramento não poderiam ter sido mais enriquecedores. Creio que a beleza do doutoramento reside na oportunidade de aprender algo novo todos os dias, de percebermos que não podíamos estar mais errados e ainda assim, ter o privilégio de descobrir algo pela primeira vez.

Esta jornada não teria ficado completa sem o precioso apoio de diversas pessoas e entidades, por isso, quero deixar aqui a minha gratidão para convosco.

À minha orientadora, Doutora Raquel Godinho, vou ser sempre grata por confiar em mim um projeto que lhe era tão especial, e mais importante de tudo, o ter partilhado comigo de forma tão generosa. Obrigada pelo teu contínuo incentivo, apoio e orientação, acredita que foram fundamentais durante este percurso. Agradeço-te ainda por sempre me dares liberdade de pensamento e apoiares as minhas ideias. Aprendi muito contigo, e isso vai para além da ciência, obrigada.

To my supervisor, Dr. John Archer, I would like to thank you for your guidance, for introducing me to the world of bioinformatics, and for always being so kind. Thanks for all the time you spent teaching me java and for being so patient, I learned a lot from you. I will always be grateful for your effort and perseverance, especially during this past year, thank you.

I am also deeply grateful to all the excellent collaborators with whom I had the pleasure to work and who contributed to make this work possible. In particular, I would like to thank prof. Cock Van Oosterhout and Dr. Hernán Morales for dedicating so much of their time to this project and for teaching me so much about modeling. To Dr. Pedro Silva, I thank all the help with the analyses and for being so generous and sharing your knowledge with me, I will always admire your way of thinking. To Dr. José Vicente López-Bao, I thank you for being our eyes in the field, your immense expertise in the ecology and behavior of wolves was fundamental in this work.

Quero agradecer também a todos os membros do grupo ECOGEN, incluindo àqueles que atualmente já não fazem parte do grupo, pelo companheirismo, ajuda e discussões

científicas. Nestes últimos anos, principalmente naqueles em que nos vimos confinados, as nossas reuniões foram momentos muito importantes para mim, obrigada a todos.

Quero também deixar o meu agradecimento a todas as técnicas do Centro de Testagem Molecular (CTM), que foram incansáveis e sempre prestaram auxílio durante o trabalho laboratorial. À Diana Castro e à Sandra Afonso devo um agradecimento especial por toda a ajuda. A todos os funcionários e staff administrativo do CIBIO, o meu agradecimento por facilitarem a nossa vida diária.

Tenho também de agradecer à minha instituição de acolhimento, CIBIO-InBIO/BIOPOLIS, bem como o apoio financeiro por parte da FCT. A todas as entidades que forneceram amostras de lobo ibérico – Instituto de Conservação da Natureza e da Biodiversidade (Portugal), Consejerías de Medio Ambiente del Principado de Asturias, de la Junta de Galicia e de la Junta de Castilla y León, Gobierno de Cantabria (Espanha), Museo Nacional de Ciencias Naturales (Madrid, Espanha), Estación Biológica de Doñana (Sevilha, Espanha) e Museu Nacional de História Natural e da Ciência (Lisboa, Portugal) – e a todos os técnicos envolvidos na colheita das amostras, o meu sincero agradecimento.

Gostaria ainda de agradecer à direção do meu programa doutoral, pelo esforço contínuo em fazer do BIODIV uma referência. Uma palavra de apreço ao Prof. Paulo Célio, pela sua dedicação e boa vontade.

A todos os meus colegas do CIBIO quero agradecer por tornarem esta jornada mais bonita. Particularmente àqueles com quem partilhei grande parte dos meus dias (e dos meus almoços), a Carolina, a Marisa, o João, o Nuno e a todos os outros que, entretanto, já terminaram a sua passagem pelo CIBIO.

Muito obrigada aos meus amigos, por serem sempre uma ponte para o equilíbrio. Especialmente à Margarida, por ser a melhor e mais presente amiga. À Vanessa, Marina, Nádia, Diana e Ana, porque apesar da ausência, quando nos vemos é como se a última vez tivesse sido ontem.

Por fim, o mais importante, à minha família. Muito obrigada por serem o meu porto seguro. Aos meus pais devo tudo aquilo que sou hoje. Obrigada pelo vosso amor incondicional e por sempre confiarem que eu podia ser tudo aquilo que quisesse. Esta conquista também é vossa! À minha irmã, por ser a melhor companheira, ouvinte e fonte de alegria. Obrigada por trazeres ao mundo o sorriso mais bonito. Ao meu querido avô

Domingos, a quem dedico este trabalho, porque sempre acreditou que eu seria capaz e sei que ficaria orgulhoso.

Ao Ivo, por (ser) tudo. Sem ti ao meu lado não teria conseguido.

Abstract

The development of human agricultural societies at least 10 thousand years ago (ka), fostered by the domestication of plants and animals, triggered extensive and enduring consequences for species as a result of dramatically altered landscapes. Some species ended extinct; others have persisted, apparently finding evolutive solutions for a rapid adaptation. Wolves in the Iberian Peninsula, *Canis lupus signatus,* are among the latter. Despite the intense human persecution over the last centuries, and contrary to most of the Central European wolf populations, Iberian wolves have never been on the brink of extinction. However, the population suffered a severe bottleneck in the 1970s, which significantly reduced its population size and distribution range. The Iberian wolf exhibits a unique set of behavioral phenotypes, such as low dispersal ability, marked consumption of human-related resources, and high tolerance towards humans and their activities, which possibly have facilitated their persistence in the highly human-dominated landscape of the Iberian Peninsula. However, up until today, no study has tried to understand whether an evolutionary mechanism underlies the persistence of wolves in human-modified landscapes. The main aim of this dissertation was to advance our understanding of the evolutionary mechanisms underlying the behavioral traits of Iberian wolves that possibly allowed their adaptation to human-dominated landscapes.

The integration of historical and contemporary data offers a unique opportunity to understand how distinct evolutionary mechanisms have affected species' history. However, clear guidelines on how to sample natural history collections (NHCs) are often not available, and the genotyping of the same set of genetic markers in contemporary and historical samples still poses several challenges. In Chapter II, we addressed these limitations by i) comparing the performance of distinct tissues – skins, nasal bones, and bones – from hundreds of Iberian wolf specimens in genome-wide assessments; and ii) developing a capture array to sequence 100,000 of the genome-wide single nucleotide polymorphisms (SNPs) present in the commercial Illumina Canine HD BeadChip, in historical samples. We showed that skin tissues should be prioritized, given the increased amount of endogenous DNA they provided, and the minimal damage that their collection inflicts on specimens. The hybrid approach for genotyping the same set of SNPs in historical and contemporary samples of Iberian wolf revealed to be fundamental for the population genomic analysis that we performed in this dissertation (Chapters III and V), where the genomic composition of both populations was directly compared.

Our first hypothesis for the apparent adaptation of Iberian wolves to human-dominated landscapes was the “introgression-fueled adaptation” (Chapter III). The underlying idea was that genetic variants from domestic dogs introgressed into the Iberian wolf genome enhancing their adaptive potential to co-exist with humans. We found that the Iberian wolf genome carries only ~1% of dog ancestry, confirming that their genetic legacy is not threatened by admixture. Still, we found a 90 kb dog haplotype in chromosome 2 of ~30% of the Iberian wolf population (historical and contemporary), which displayed signatures compatible with adaptive introgression. The time of introgression was estimated to be ~3.6 ka, which is a period concordant with the settlement of human communities during the Bronze age. The introgressed haplotype contain the MAST4 gene, which is a possible candidate for some of the behavioral responses known in Iberian wolves given its involvement in several neurological processes and its association with NMDA glutamate receptors. The latter have been extensively associated with mammal domestication for regulating levels of anxiety and stress responses. Therefore, we propose that dog adaptive introgression may have contributed, at least partially, to the persistence of Iberian wolves in human-dominated landscapes, by increasing their tolerance towards humans and human activities.

In addition to the latter hypothesis, we also proposed that adaptation could result from a natural evolutionary process towards traits favoring persistence, which could resemble some of the behavioral phenotypes selected during dog domestication. This would lead to an ongoing process of self-domestication in the Iberian wolf (Chapter IV). We addressed this hypothesis by generating complete brain transcriptomes of four Iberian wolves living in the wild and subsequently comparing their gene expression profiles with those of non-Iberian wolves and dogs. Among the genes differentially expressed between Iberian and non-Iberian wolves, there were genes associated with key pathways for neurological processing, such as serotonin and glutamate genes, and genes involved in locomotory behavior. The direction of expression in some of these genes was similar in Iberian wolves and dogs and pointed towards reduced anxiety, increased cognitive abilities, and reduced exploratory tendencies. In addition, we also detected overlapping signatures of allele frequency changes and expressional variation in genes implicated in the same group of functions. These results suggest an important role of neurotransmission in shaping the behavioral traits found in Iberian wolves, which might be allowing their adaptation to human-dominated landscapes through a process of self-domestication.

In Chapter V, we aimed to uncover the effects of the population bottleneck on the genomic composition of the Iberian wolf, by directly comparing historical and contemporary samples. Particularly, we focused on the levels of genetic diversity and dynamics of interspecific hybridization with domestic dogs. This comparative approach revealed a rapid and sudden loss of genetic diversity following the period of the most dramatic decline in the population. Nevertheless, the loss of genetic diversity did not occur randomly in the genome, as evidence of active maintenance of variation in genes associated with immune response was found in the contemporary population. Population decline did not promote an increase in the rate of wolf-dog hybridization, nor in the amount of dog ancestry found in the genome of historical individuals, which was estimated to be < 2%. These findings are particularly relevant to characterize aspects of the evolutionary trajectory of wolves in the past decades in the Iberian Peninsula but can also provide valuable insights for other populations of large carnivores facing similar declines.

To aid in the analysis of RNA-seq data, we developed a new software tool, TVscript (Chapter VI), that estimates per transcript read count intra-condition variation and allows the removal of hyper-variable transcripts before differential expression analysis. We demonstrate the effectiveness of TVscript in a simulated dataset and in a real case study, involving publicly available RNA-seq data from the brain of wolves, dogs, tame and aggressive foxes. The removal of hyper-variable transcripts from the reference set increased the sensitivity of differential expression analysis. Additionally, by comparing the expression profiles of wild/aggressive vs. domestic/tame animals we identified a set of genes commonly expressed in the brain of dogs and tame foxes that could be potential candidates for tame behavior.

Overall, the work developed in this dissertation will contribute to a better understanding of the evolutionary trajectory of wolves living in human-dominated landscapes and suggest that distinct evolutionary mechanisms, involving adaptive introgression and self-domestication, might have been key drivers for the behavioral adaptation of Iberian wolves. Emerging techniques of genome-editing have the potential to revolutionize our understanding of genotype-phenotype associations and will greatly contribute to fully characterize the functional impact of our findings.

Keywords: adaptation, dogs, domestication, genomics, hybridization, introgression, transcriptomics, wolves.

Resumo

O desenvolvimento da agricultura pelos humanos há pelo menos 10 mil anos, fomentado pela domesticação de plantas e animais, desencadeou uma série de consequências extensas e contínuas para as espécies em resultado de paisagens altamente modificadas. Algumas espécies acabaram por se extinguir; outras persistiram, aparentemente encontrando soluções evolutivas para uma adaptação rápida. Os lobos na Península Ibérica, *Canis lupus signatus*, encontram-se entre o último caso. Apesar da intensa perseguição humana que ocorreu nos últimos séculos, e contrariamente à maior parte das populações de lobo na Europa Central, o lobo ibérico nunca esteve perto da extinção. Contudo, perto dos anos 1970s a população sofreu uma significativa redução no seu tamanho populacional bem como na sua área de distribuição. O lobo ibérico exibe um conjunto único de fenótipos comportamentais, como baixa propensão para dispersar, dieta baseada em recursos pecuários, e alta tolerância para com os humanos e as suas atividades, que poderá estar a permitir a sua persistência em paisagens altamente modificadas pelo Homem, como a Península Ibérica. No entanto, até à data, nenhum estudo abordou a hipótese de a persistência dos lobos neste tipo de paisagens estar associada a um mecanismo evolutivo. O objetivo principal desta dissertação foi avançar o conhecimento sobre os mecanismos evolutivos que podem estar associados aos fenótipos comportamentais do lobo ibérico, que possivelmente permitiram a sua adaptação a paisagens modificadas pelo Homem.

A integração de dados históricos e contemporâneos oferece uma oportunidade única para compreender como diferentes mecanismos evolutivos afetaram a história das espécies. Contudo, existe uma falta de diretrizes claras de como amostrar coleções de história natural, e a genotipagem do mesmo conjunto de marcadores genéticos em amostras contemporâneas e históricas ainda apresenta vários desafios. No Capítulo II, abordamos estas limitações ao i) comparar o desempenho de diferentes tecidos – peles, ossos nasais e ossos – de centenas de espécimes de lobo ibérico em estudos genómicos; e ii) desenvolver um *array* para capturar e sequenciar 100,000 dos polimorfismos de nucleótido único (SNPs) que estão presentes no chip comercializado pela Illumina, Canine HD BeadChip, em amostras históricas. Demonstramos que os tecidos de pele devem ser priorizados dado o maior conteúdo de ADN endógeno que providenciam, bem como o baixo grau de danificação que a sua coleção inflige aos espécimenes. A abordagem híbrida para genotipar o mesmo conjunto de SNPs em

amostras históricas e contemporâneas de lobo ibérico foi fundamental nas análises de genómica populacional que realizamos nesta dissertação (Capítulos III e V), onde a composição genómica de ambas as populações foi diretamente comparada.

A nossa primeira hipótese para a aparente adaptação do lobo ibérico a paisagens modificadas pelo Homem foi a "adaptação por introgressão" (Capítulo III). A ideia subjacente era que variantes genéticas do cão doméstico foram incorporadas no genoma do lobo ibérico por introgressão, potenciando a sua capacidade adaptativa para coexistir com os humanos. Mostramos que o genoma do lobo ibérico possui apenas ~1% de ancestralidade de cão, o que confirma que o seu legado genético não está ameaçado por miscigenação. Não obstante, encontramos um haplótipo de cão com 90 kb no cromossoma 2 de ~30% da população de lobo ibérico (histórica e contemporânea), que apresentava padrões compatíveis com introgressão adaptativa. O tempo de ocorrência da introgressão foi estimado em ~3.6 milhares de anos atrás, um período concordante com o estabelecimento das comunidades humanas durante a idade do Bronze. O haplótipo resultante da introgressão contém o gene MAST4, cuja associação com diversos processos neurológicos e com recetores de glutamato NMDA, fazem dele um possível candidato para alguns dos fenótipos comportamentais observados no lobo ibérico. Os recetores NMDA têm vindo a ser amplamente associados com a domesticação em mamíferos por regularem níveis de ansiedade e respostas ao stress. Desta forma, propomos que a introgressão adaptativa de cão poderá ter contribuído, ainda que parcialmente, para a persistência do lobo ibérico em paisagens modificadas pelo Homem, ao aumentar a sua tolerância aos humanos e às suas atividades.

Adicionalmente à hipótese supramencionada, propusemos que a adaptação poderia resultar de um processo evolutivo natural para fenótipos que favorecem a persistência, e que se poderiam assemelhar àqueles que foram selecionados durante a domesticação do cão. Este processo levaria à "auto-domesticação" no lobo ibérico (Capítulo IV). Testamos esta hipótese através da sequenciação do transcriptoma cerebral de quatro lobos ibéricos a viver em estado selvagem, e, posteriormente, comparando os seus perfis de expressão génica com os de lobos não ibéricos e cães. Entre os genes diferencialmente expressos entre lobos ibéricos e não ibéricos, foram encontrados genes envolvidos em vias-chave para o processamento neurológico, como os genes da serotonina e do glutamato, e genes associados no comportamento locomotor. A direção de expressão de alguns desses genes era similar em lobos ibéricos e em cães, e apontava para uma redução da ansiedade, capacidade cognitiva aumentada, e uma

redução das tendências exploratórias. Adicionalmente, foi também detetada uma sobreposição entre alterações das frequências alélicas e variação na expressão génica em genes implicados nesse mesmo grupo de funções. Estes resultados sugerem um papel importante da neurotransmissão nos fenótipos comportamentais do lobo ibérico, o que poderá estar a permitir a sua adaptação a paisagens modificadas pelo Homem através de um processo de "auto-domesticação".

No Capítulo V integramos amostras históricas e contemporâneas de lobo ibérico, para compreender a magnitude dos efeitos da redução populacional na composição genómica do lobo ibérico. Focamo-nos particularmente nos níveis de diversidade genética e nas dinâmicas de hibridação interespecífica com cães domésticos. Esta abordagem comparativa revelou uma perda rápida e repentina da diversidade genética após o período de declínio populacional mais acentuado. Porém, evidência de uma manutenção ativa de diversidade em genes associados com resposta imunitária na população contemporânea, revelou que a perda de diversidade não decorreu de forma aleatória pelo genoma. O declínio populacional não promoveu um aumento nos eventos de hibridação lobo-cão, nem na proporção de ancestralidade de cão encontrada no genoma da população histórica, a qual foi estimada em < 2%. Estes resultados são particularmente relevantes para compreender aspetos da trajetória evolutiva dos lobos na Península Ibérica nas últimas décadas, mas podem inclusivamente ser úteis para outras populações de grandes carnívoros que estão a enfrentar declínios semelhantes.

Para auxiliar na análise de dados de *RNA-seq*, desenvolvemos uma nova ferramenta de software, TVscript (Capítulo VI), que estima por transcrito níveis de variação na abundância de expressão génica dentro das condições e permite remover transcritos altamente variáveis antes da análise diferencial de expressão. A eficácia e utilidade do TVscript foi demonstrada em dados simulados e num caso de estudo real, envolvendo dados públicos de *RNA-seq* do cérebro de lobos, cães, raposas dóceis e agressivas. A sensibilidade da análise de expressão génica mostrou-se aumentada após a remoção de transcritos altamente variáveis do conjunto de referência. Adicionalmente, a comparação dos perfis de expressão génica de animais selvagens/agressivos vs. domésticos/dóceis permitiu identificar um conjunto de genes comumente expressos no cérebros de cães e raposas dóceis, que podem ser possíveis candidatos para o comportamento dócil.

De modo geral, o trabalho desenvolvido nesta dissertação irá contribuir para uma melhor compreensão da trajetória evolutiva de lobos a viver em paisagens modificadas pelo

Homem, sugerindo que mecanismos evolutivos distintos, como introgressão adaptativa e “auto-domesticação”, podem ter impulsionado a adaptação comportamental do lobo ibérico. Técnicas emergentes de edição do genoma têm o potencial de revolucionar a nossa compreensão sobre as associações genótipo-fenótipo e irão contribuir de forma significativa para caracterizar o impacto funcional dos nossos resultados.

Palavras-chave: adaptação, cão, domesticação, genómica, hibridação, introgressão, lobo, transcriptómica.

Table of contents

Nota Prévia 4
Acknowledgments / Agradecimentos 5
Abstract 8
Resumo 11
Table of contents 15
List of figures 20
List of tables 29
List of abbreviations 30

Chapter I. General Introduction 33

1.1. Evolutionary adaptation in an anthropogenic world 34
1.1.1. Genetic diversity and signatures of neutral and selective events 35
1.1.2. Gene flow as a source of novel genetic variation 37
1.2. Genomics of animal domestication 39
1.2.1. Origin of the common traits shared by domestic animals 40
1.2.2. Genetic basis of the domestic behavioral phenotype 42
1.3. Hybridization between wild and domestic animals 44
1.3.1. The case of wolf-dog hybridization 46
1.4. Insights into the evolutionary history of gray wolves 50
1.4.1. The Iberian wolf as a remarkable case of persistence 54
1.5. Objectives and dissertation outline 58
1.6. References 61

Chapter II. Assessing the performance of historical skins and bones for museomics using wolf specimens as a case study 75

Abstract 76
2.1. Introduction 77
2.2. Materials and methods 79
2.2.1. Sampling, DNA extraction, and shotgun sequencing 79
2.2.2. Bait design 80
2.2.3. Capture enrichment 81
2.2.4. Processing of sequencing reads and single nucleotide polymorphisms calling 81
2.2.5. Statistical analysis 83

2.3. Results …… 83
2.3.1. Endogenous DNA content …… 83
2.3.2. Capture efficiency across sample types …….. 85
2.3.3. Genotyping errors and coverage effect ……... 85
2.4. Discussion …….. 87
2.4.1. Reducing sample damage without compromising data recovery 89
2.4.2. Limitations and opportunities for further development ……. 90
2.5. Concluding remarks ……... 91
2.6. References …… 92
2.7. Supplementary Materials …… 97

Chapter III. Dog introgression into the Iberian wolf genome may have facilitated adaptation to human-dominated landscapes …… 109

Abstract …… 110
3.1. Introduction …… 111
3.2. Results and Discussion …… 112
3.2.1. Presence of a dog block in chromosome 2 of Iberian wolves …. 112
3.2.2. Evidence that the dog block is an introgressed variant …… 117
3.2.3. Simulations of selection support adaptive introgression …….. 119
3.2.4. Potential functional role of the introgressed haplotype ……. 121
3.2.5. The origin and evolution of Δblock ……... 122
3.3. Conclusions …… 123
3.4. Materials and Methods …… 124
3.4.1. Canine HD SNP BeadChip data ……... 124
3.4.2. Genetic diversity analysis ……..... 125
3.4.3. Global ancestry estimation ……..... 125
3.4.4. Local ancestry analysis ……..... 126
3.4.5. SNP-specific Δstatistics and validation of introgression …….. 126
3.4.6. Historical samples and in-solution capture enrichment …… 127
3.4.7. Local ancestry analysis in historical samples ……. 128
3.4.8. Whole-genome sequences and SNP calling …….. 129
3.4.9. Introgression analysis using WGS data …….. 130
3.4.10. Genetic sequence similarity ……..... 131
3.4.11. Common introgressed block and shared haplotype …… 131
3.4.12. Probability of maintaining the 90 kb haplotype from shared ancestral variation ……..... 133
3.4.13. Testing for adaptive introgression and SLiM simulations ……... 133

3.4.14. Estimating time of dog introgression 134
3.5. References 136
3.6. Supplementary Materials 143

Chapter IV. Brain transcriptomes uncover a possible genetic basis for the behavioral responses of Iberian wolves to human-dominated landscapes 181

Abstract 182
4.1. Introduction 183
4.2. Results and Discussion 184
4.2.1. Gene expression profiles in the brain of wolves and dogs 184
4.2.2. Expressional changes in candidate genes for the domestic behavioral phenotype 188
4.2.3. Overlap between DEGs and allele frequency changes in Iberian wolves 190
4.3. Conclusions 193
4.4. Materials and Methods 193
4.4.1. Brain tissue collection 193
4.4.2. RNA-seq experiments 194
4.4.3. Differential expression analysis 195
4.4.4. Exome-wide SNP calling and allele frequency estimate 196
4.5. References 198
4.6. Supplementary Materials 202

Chapter V. Linking effects of a wolf population bottleneck to historical and contemporary genomic signatures of diversity and admixture with domestic dogs 211

Abstract 212
5.1. Introduction 213
5.2. Materials and Methods 215
5.2.1. Historical and contemporary SNP dataset 215
5.2.2. Global ancestry estimation and hybrid identification 216
5.2.3. Local ancestry analysis across hybrid genomes 217
5.2.4. Local ancestry analysis across the historical population genomes 218
5.2.5. Genetic diversity analysis 218
5.3. Results 219
5.3.1. Levels of admixture and frequency of hybridization 219
5.3.2. Local ancestry analysis of hybrids 221

5.3.3. Proportion of dog introgression in the historical population ……. 221
5.3.4. Patterns of genetic diversity …….. 222
5.4. Discussion …….. 225
5.4.1. Dynamics of wolf-dog hybridization across a population bottleneck ……. 225
5.4.2. Dog introgression as a conservation threat to wolf populations .. 226
5.4.3. The non-random loss of genetic diversity ……... 227
5.5. Conclusions …… 229
5.6. References …… 231
5.7. Supplementary Materials ……. 237

Chapter VI. On taming the effect of transcript level intra-condition count variation during differential expression analysis: a story of dogs, foxes and wolves ……... 248

Abstract …… 249
6.1. Introduction …… 250
6.2. Materials and Methods …… 253
6.2.1. RNA-seq datasets ……..... 253
6.2.2. Software …… 254
6.2.3. Controlled intra-condition variation within simulated data …… 255
6.2.4. Exploring the removal of transcripts associated with high intra-condition variation within real data ……... 257
6.2.5. Gene annotation and gene family analysis ……. 259
6.3. Results …… 259
6.3.1. Mapping success of RNA-seq data …….. 259
6.3.2. Controlled intra-condition variation within simulated data …… 260
6.3.3. Exploring the removal of transcripts associated with high intra-condition variation within real data …… 263
6.3.4. Differential expression analysis ……... 265
6.3.5. Evaluation metrics ……..... 265
6.3.6. Candidate genes and gene families ……. 268
6.4. Discussion …….. 270
6.5. References …… 275
6.6. Supplementary Materials ……. 281

Chapter VII. General Discussion ……. 319

7.1. Introgression as a driver of adaptation of wolves to human-dominated landscapes ……. 320
7.2. Evidence for self-domestication in Iberian wolves ……... 324

7.3. Genomic signatures following a population bottleneck in Iberian wolves ……. 327
7.4. Workflow to detect hybridization and introgression ……. 329
7.4.1. Population-based methods …… 330
7.4.2. Individual-based methods …….. 332
7.5. Concluding remarks and perspectives …….. 335
7.6. References ……. 337
Appendix …… 343
Appendix I …….. 343
Appendix II ……. 355

List of figures

Chapter I.

Figure 1.1. Pictures of several animal domesticated species and tame foxes from the "Belyaev Fox Farm experiment", showing similar phenotypic traits, such as changes in coat coloration, floppy ears, and increased tameness (shown in different columns). This common set of biological traits is among those that constitute the "domestication syndrome". Figure adapted from Trut et al. 2009.

Figure 1.2. Picture of an Iberian wolf (*Canis lupus signatus*; top) and two wolf-dog hybrids (middle and bottom) found within the same pack in Sierra de Penouta, Western Asturias, Spain. Individual assignments to the wolf cluster estimated through Bayesian admixture analyses are presented (Godinho et al. 2011). Phenotypic differentiation between the three individuals can be seen particularly in the coat color, nose elongation, and body size. Figure adapted from Godinho et al. 2011.

Figure 1.3. Map showing the IUCN geographical range of gray wolves in the world (2018). Orange areas represent the extant distribution, green represents areas where the species was reintroduced, and red represents areas where the gray wolf went extinct (in this map only Japan is considered as an extinct area, but gray wolves disappeared also from Mexico, most of the USA, and Western Europe). The dashed box highlights the Iberian Peninsula, which includes Portugal and Spain. Map retrieved from the IUCN website, last accessed on 23rd of August 2022.

Figure 1.4. Phylogenetic relationships within the wolf-like clade. (A) Bayesian phylogenetic tree of the family Canidae based on ~15 kb DNA sequence from exons and introns across the genome. Different clades are shown with different colors: the red-fox-like clade in red, the South-American clade in green, the wolf-like clade in purple, and the gray and island fox clade in orange. Values above and below the internodes represent bootstrap values and Bayesian posterior probability values, respectively (dashes indicate values below 95%). Underlined species names have the corresponding illustration at the right. Figure adapted from Lindblad-Toh et al. 2005. (B) Increased resolution of the relationships within the wolf-like clade, including estimates of divergence times in millions of years before the present (horizontal gray bars represent 95% of highest posterior densities) using nuclear sequence alignments of 70 kb to 28

Mb. Note that in this case, both black-backed and side-striped jackals are positioned outside the wolf-like clade. Given their high divergence and distant phylogenetic position, it was recently proposed that their status should be revised and that they should be placed in a distinct genus (Álvares, 2019). An extinct species, the Dire wolf (*Canis dirus*), is also represented in the tree, found to have split from living wolf-like canids around 5.7 Ma. Figure adapted from Perri et al. 2021.

Figure 1.5. Gray wolf phylogeny and distribution across Europe. (A) Genome-wide phylogenetic tree based on autosomal SNPs. Estimates of divergence time between New and Old-World wolves are presented. Different colors represent different wolf populations in Eurasia and dogs. Values represent the node support from bootstrapping (dashes indicate values below 95%). Figure adapted from Fan et al. 2016. (B) The current distribution of wolves in Europe. Dark green patches indicate areas of permanent wolf occurrence, light green patches indicate areas of sporadic wolf occurrence, and orange dots indicate reports of recent sightings. Yellow lines denote boundaries between the 10 defined wolf populations (Chapron et al. 2014). Figure adapted from BBC/WWF.

Figure 1.6. Distribution of wolves in the Iberian Peninsula. (A) Picture of an Iberian wolf, where is possible to see the distinctive dark marks on its front legs. This characteristic, together with dark marks on the tail (not entirely visible in this picture), were responsible for the designation *signatus*, meaning “marked” or “signed”, by Ángel Cabrera in 1907, who proposed that Iberian wolves represent a distinct subspecies. Photo credits: Raquel Godinho. (B) Estimated wolf range in the Iberian Peninsula, including Portugal and Spain, between the 19th and 21st-century. Figure adapted from Sastre 2011.

Figure 1.7. Genetic structure and low dispersal in Iberian wolves. (A) Spatial overlap between the movement behavior of 85 collared wolves (between 1982 and 2015) is represented by the hollow polygons, and the genetic clusters identified at K = 11 (in yellow). Red polygons represent the movement of the two wolves overlapping more than one genetic group. (B) Ancestry proportion of wolves to each of the genetic clusters (K = 4 and K = 11). Each circle represents a single individual and the polygons a cluster. The white dotted area identifies regions within the cluster with only non-admixed individuals. The shaded green area represents the wolf distribution in the Northwestern region of the Iberian Peninsula. Figures adapted from Silva et al. 2018.

Chapter II.

Figure 2.1. (A) Photographs illustrating the specimens used for the collection of each tissue type used in this study (Image credits Raquel Godinho). (B/C) Correlation (r^2) between DNA yield (ng) following extraction and the proportion of endogenous DNA content with the original collection date of the specimens. Each dot represents a historical sample and inset boxplots display the median (central line) and distribution per tissue type. Significant differences ($p < 0.05$) between tissues are identified with an asterisk. Colors depict the three different tissue types used in this study: orange - bones (jaw and femur); yellow - maxilloturbinal (nasal) bones; and, green - skins.

Figure 2.2. Genomic metrics assessed from capture enrichment for each tissue type: bones (orange), maxilloturbinal (nasal) bones (yellow) and skins (green). Boxplots display the proportion of reads mapping after capture enrichment and reads mapping on target region in relation to all or mapped reads for each tissue. Within boxplots, dots represent historical samples, and the central line indicates the median value. Significant differences ($p < 0.05$) between tissues for each metric are identified with an asterisk.

Figure 2.3. Genotyping performance of hDNA from three sample types. (A) Interval of potential genotyping error rate estimates (%) for the three hDNA sample types, and average empirical genotyping error rate for control samples, considering the three minimum depth levels. (B) Percentage of genotyped SNPs (N = 99,982) considering a minimum depth of 3x, 4x or 5x per SNP across 103 samples. Each dot represents a sample.

Chapter III.

Figure 3.1. Sample location. (A) Geographical location of the analyzed canid samples across the world. Yellow, dark yellow and black dots denote samples from Eurasian wolves, North American wolves, and a golden jackal, respectively. (B) Zoomed-in view of the sampling area in the Iberian Peninsula, delimited by the dashed line in (A). Iberian wolves with and without the ∆block are represented in red and pink, respectively. Circles and triangles denote contemporary and historical samples. Further details about samples can be found in Supplementary Tables 3.1, 3.2, and 3.4. Dog samples are not represented. (C) Picture of an Iberian wolf (*Canis lupus signatus*); credits Raquel Godinho.

Figure 3.2. Evidence of a dog block in Iberian wolf chromosome 2. (A) ∆ancestry scores indicating the excess of dog ancestry across the Iberian wolf genome (based on ~85K SNPs across 95 contemporary samples). The ∆block on chromosome 2 is shadowed in gray. The inset displays a zoomed-in view of chromosome 2, with the dashed line indicating the cut-off used for detecting outliers. (B) Representation of the dog block found in chromosome 2 of Iberian wolves using ELAI local ancestry analysis (considering 10 mixture generations). Each row represents an individual (same order as in Supplementary Table 3.1) and different colors denote the attributed local ancestry: pink for homozygous wolf, red for wolf/dog (heterozygous), and purple for homozygous dog. Most individuals are heterozygous for the dog block. (C) Network of ∆block haplotypes (78 SNPs) for contemporary Iberian wolf samples and village dogs from Iberia. The size of the circles is scaled relative to the frequency of each haplotype. (D) NJ tree for the ∆block using contemporary and historical (subset > 60K SNPs; see Materials and Methods) samples of Iberian wolves, and village dogs from Iberia. In panels C and D Iberian wolves with and without the ∆block are colored in red and pink, respectively, and dogs in purple. Contemporary and historical samples are represented by circles and triangles, respectively.

Figure 3.3. Signatures of adaptive introgression using WGS data. (A) Representation of the ∆block on chromosome 2 of five Iberian wolf genomes using local ancestry analysis. Different colors denote the attributed local ancestry: pink for homozygous wolf, red for wolf/dog (heterozygous), and purple for homozygous dog. (B) Structure of the haplotype (~90 kb; 470 SNPs) found in the ∆100 kb region in common to all Iberian wolves with the ∆block. In the top is the introgressed haplotype found among Iberian wolves; in the middle are the dog haplotypes (village dogs from Iberia appear first) and then the haplotypes found among wolves without the ∆block (Iberian, Eurasian, and North American wolves); in the bottom are the golden jackal haplotypes. Each row depicts a haplotype, and each position is colored according to whether it carries the introgressed allele (∆ in red) or the alternative allele (in pink). Haplotypes marked with an asterisk represent the alternative haplotype of Iberian wolves with the ∆block. Numbers within brackets indicate the total number of haplotypes per group.

Figure 3.4. Increased allele sharing between Iberian wolves and dogs measured with WGS data. (A) Fraction of introgression (f_d) between Iberian wolves with the ∆block and dogs across chromosome 2 in non-overlapping 100 kb windows. Red dots indicate the highest 1% outlier windows. The test was performed following the phylogenetic conformation illustrated on the right (P1– pink: Iberian wolves without the ∆block, n = 9;

P2 – red: Iberian wolves with the ∆block, n = 5; P3 – purple: dogs, n = 13; outgroup – black: golden jackal, n = 1. (B) Pairwise F_{ST} between different groups (represented by distinct colors) in non-overlapping 50 kb windows within ∆block. The position of the ∆100 kb region is represented by a black bar, and the protein coding genes located within the region and in the vicinity are illustrated below. (C) Distribution of d_{XY} between haplotypes of wolves and dogs within the ∆100 kb region and in the 6 Mb flanking region (each comparison is color coded). (D) The probability of maintaining a haplotype of 0-200 kb length due to ancestral shared variation, assuming different divergence times between Eurasian gray wolves and dogs. The gray star indicates the length of the introgressed haplotype (90 kb) and the corresponding p-value.

Figure 3.5. Individual-based simulations of genetic introgression. Mean and standard deviation of 100 simulated replicates in SLiM for average admixture proportion (across the entire ∆100 kb region) and mean frequency of introgressed alleles, across adaptive models with one migrant and one migration event at generation 800 (plots with results for distinct timings of introgression can be found in Supplementary Figure 3.27). Empirical estimates of admixture (0.53-0.60; 95% CI of ELAI average estimate) and mean frequency of introgressed alleles (see Supplementary Table 3.6) in ∆100 kb region, are shown in dashed lines. Simulations explored neutral evolution, as well as an additive model (h = 0.5), dominant model (h = 1) and overdominant model (h = 2), with a range of selection coefficients (s = 0-0.1) for 1, 3 or 6 introgressed SNPs (denoted with distinct colors), resulting in a total amount of selection ranging from 0 to 0.6 (x-axis).

Chapter IV.

Figure 4.1. PCA of gene expression profiles and exome-wide NJ tree. (A) PCA plots for each of the three comparisons performed for differential expression analysis, constructed using normalized counts (rlog transformed) of all the genes in the reference set. Each dot represents a sample (for each of the four Iberian wolves three replicates/individual libraries were performed), and each group is represented by a different color. (B) NJ tree constructed with genetic distances estimated between all individuals using 79,267 exome-wide SNPs. Individuals used for the RNA-seq data are indicated with “RNA” in their sample ID (for Iberian wolf samples the three replicates were merged for genotyping). Each species/subspecies is represented by a different color.

Figure 4.2. Levels of differential expression. MA plots showing the number of differentially expressed genes in each comparison (A) non-Iberian vs. Iberian wolves; (B) non-Iberian wolves vs. dogs; (C) Iberian wolves vs. dogs. The direction of expression is presented always in relation to the second condition (indicated within each plot by the dark silhouette). Up and down-regulated genes are colored in red and blue, respectively, and the corresponding number of genes in each category is presented.

Figure 4.3. Expression profiles in the set of candidate genes. (A) Heatmap of rlog transformed mean centered counts reflecting levels of expression for genes associated with the domestic behavioral phenotype (rows) between Iberian and non-Iberian wolves. Genes that presented the same significant direction of expression in dogs as in Iberian wolves are identified with a star. Hierarchical clustering of Pearson's correlation coefficient between samples is represented in the dendrogram. (B) Expression profiles (average of rlog normalized counts) of genes marked with a star in (A) across the three groups. Boxplots display the mean value and standard errors, and each dot represents a single sample. The tree replicates per individual in Iberian wolves are presented in both figures.

Chapter V.

Figure 5.1. Sampling location and admixture levels. (A) Geographical location of the analyzed wolf samples in the Iberian Peninsula (delimited by the dashed line in the left map). Historical (N = 54) and contemporary (N = 68) samples are represented in red and orange, respectively (the same color scheme was followed in panels B and C). Circles and triangles denote pure wolves and hybrids, respectively. Further details about samples can be assessed in Supplementary Table 5.1. (B) Principal component analysis (PCA) for historical and contemporary wolves and dogs (N = 43; represented in blue). Each dot denotes an individual. Principal components 1 and 2 are displayed with the corresponding variance values. (C) Ancestry fractions of all individuals inferred in ADMIXTURE for historical/contemporary wolves and dogs in K = 2 clusters. Historical and contemporary wolf samples were analyzed independently. Individuals are represented by a single bar and different colors indicate the ancestry proportion attributed to each cluster. White dotted lines separate wolves from dogs.

Figure 5.2. Local ancestry analysis and introgression levels. (A) Fraction of the global dog ancestry estimated for the 15 hybrids using LAMP-ANC (full dots) and ADMIXTURE (light dots). Historical and contemporary samples are represented in red and orange, respectively. Sample IDs are presented on the x-axis. (B) Δ ancestry scores indicating

the excess of dog ancestry across the hybrid chromosomes (backcrosses only). The dashed line identifies the cut-off used to consider outlier regions (0.14; mean + 2SD). (C) Local ancestry in the genome of two wolves sampled in Sierra Morena in 1927 and 1991. Genomic blocks were colored based on the attributed local ancestry: red for wolf (both chromosomes), blue for dog (both chromosomes), and half red/blue for wolf/dog (one chromosome of each).

Figure 5.3. Genetic diversity in the historical and contemporary populations. (A) Proportion of heterozygous sites present in 52 individuals from the historical (green) and contemporary (orange) populations across time (year of the original collection). Historical samples from the contracting (N = 17) and stable/expanding (N = 35) are represented in dark and light green, respectively. Each dot denotes an individual and black dashed lines represent the average value in each population. (B) Distribution of heterozygosity across the genome of historical (red) and contemporary (orange) individuals (N = 22, in each), in non-overlapping windows of 10 SNPs (~350 kb). The red and orange lines identify the cut-off used to define high variation windows in each population. (C) Inverse cumulative average number of runs of homozygosity (ROH) per individual across 1 Mb non-overlapping windows relative to their average size, for each population. (D) Average size of ROH per individual (represented by a single bar) in the historical population (further information is available in Supplementary Table 5.5). (E) Decay of linkage disequilibrium across populations. Each dot represents the averaged pairwise r2 values between pairs of SNPs in non-overlapping 1 kb windows. The inset plot represents a zoom-in view of the first 50 kb, where each dot represents the averaged pairwise r2 values between pairs of SNPs in non-overlapping 100 bp windows.

Chapter VI.

Figure 6.1. Diagramatic overview of how reference based transcripts are removed by TVscript. Steps (i) to (viii) indicate actions taken. Reference transcripts (green circles) are showen for diagramtic purposes in order to highlight how read counts (gray circles) across replicates are treated for each transcript independently. Read counts are grouped into individual files (gray rectangles) in accordance to replicate. These files are grouped into one of two conditions (blue and brown boxes). Remaining keys are indicated at the bottom.

Figure 6.2. Over expressed transcripts pre- and post-filtering using simulated data. The number of transcripts identified by DESEQ2 as being over expressed both prior to (light gray) and post (dark gray) filtering of count datasets within each of the 100 iterations performed at each level of introduced random intra-condition count variation. Each iteration involved initially simulating ten count datasets divided into conditions A and B following which DESEQ2 was run to attempt to identify the 100 transcripts selected for over representation as described in the methods. Following this the ten simulated datasets were filtered using TVscript with a 95th percentile threshold to generate corresponding filtered datasets (divided into corresponding conditions A' and B') on which DESeq2 was re-run.

Figure 6.3. Characterization of intra-condition variation. Percentile range of intra-condition variation scores (x-axis) observed prior to filtering, across both case studies, (A) wolves (orange) and dogs (red); (B) tame (dark blue) and aggressive (light blue) foxes. PCA plots based on normalized non-filtered count data of the individual datasets comparing (C) wolf and dog, and (D) tame and aggressive fox. In the latter only individual samples that were positioned within a distant cluster are labelled with the sample ID.

Figure 6.4. Removal of transcripts above specified levels of intra-condition variation. Percentile range of combined intra-condition variation scores (x-axis) present in each case study, (A) wolves and dogs; (B) tame and aggressive foxes. The number of transcripts removed in the top five percentiles (from the 95th to the 99th) are presented in each panel.

Figure 6.5. Effects of removing transcripts above specified levels of intra-condition variation on differential expression analysis. (A) Number of differentially expressed transcripts (DETs) identified using non-filtered (NF) and filtered datasets, based on the top 10 percentiles (99th to the 90th), for both case studies. Up and down-regulated transcripts are represented by red and blue dots, respectively. Gray arrows identify the selected thresholds for which the results of subsequent corresponding differential expression analysis were used for the identification of candidate transcript associated with tameness within each case study. (B) Number of transcripts identified as differentially expressed within the non-filtered datasets that were associated the highest levels of intra-condtion variation (99th to 90th) within both case studies, wolves and dogs (orange dots), and tame and aggressive foxes (blue dots).

Figure 6.6. Distribution of dispersion estimates. Plots of final dispersion estimates for both case studies, (A) wolves and dogs; (B) tame and aggressive foxes, calculated using DESEQ2 for the non-filtered (NF; orange) and 10% filtered datasets (90^{th}; blue). Each black dot represents a single transcript, and red dots represent outliers. The number of outliers and correlation index (r^2) are displayed in the top right corner of each pannel. Both x and y-axis are transformed into a logarithm scale. The line in each graph corresponds to the regression analysis between the mean of normalized counts and dispersion estimates.

Figure 6.7. DETs overlapping between non-filtered and filtered datasets. Venn diagrams representing the number of overlapping DETs found following differential expression analysis using non-filtered datasets and filtered datasets (97^{th}, 95^{th} and 90^{th} percentiles), within each case studies. The inset table provides information about the number of DETs lost/added following each filter step in relation to the non-filtered dataset as well as the the percentage of total discordance.

Chapter VII.

Figure 7.1. Stepwise approach with methods to detect hybridization and genetic introgression between two populations, or species. This workflow was designed for genomic datasets containing a high-density panel of genome-wide SNPs. Having a well-annotated reference genome and chromosome-level information can help in some of this analysis (e.g., step 2, local ancestry methods).

List of tables

Chapter IV.

Table 4.1. ΔAF. Allele frequency differences in candidate genes overlapping DEGs between Iberian and non-Iberian wolves. Information about the direction of gene expression in Iberian wolves, chromosome and chromosome position, derived allele, frequency of the derived allele calculated for the RNA-seq and WGS data in Iberian and non-Iberian wolves, and the genomic consequence associated with each SNP is presented. Data absent from WGS is represented with NA.

Chapter V.

Table 5.1. Over-represented GO categories among the top 90 high variation windows found in the contemporary Iberian wolf population.

Chapter VI.

Table 6.1. Shared genes and gene families (Up-regulation). List of the gene families, and shared genes, that were commonly up-regulated in dogs and tame foxes. The number, and name, of the genes within each gene family are provided, with the corresponding log2fold-change values in brackets for each species. Within each family, single genes were charecterized as shared between dogs and tame foxes (bold), or as exclusively to each of the two groups. When more than one transcript for a specific gene was present, all the log2FC values are reported.

Table 6.2. Shared genes and gene families (Down-regulation). List of the gene families, and shared genes, that were commonly down-regulated in dogs and tame foxes. The number, and name, of the genes within each gene family are provided, with the corresponding log2fold-change values in brackets for each species. Within each family, single genes were characterized as shared between dogs and tame foxes, or as exclusively to each of the two groups. When more than one transcript for a specific gene was present, all the log2FC values are reported

Abbreviations

AIMs – ancestry informative markers

bp – base pairs

BQ – base quality

CI – confidence interval

DEGs – differentially expressed genes

DETs – differentially expressed transcripts

DNA – deoxyribonucleic acid

DP – sequencing depth

D_{XY} – absolute sequence divergence

EM – expectation maximization

EMBL-EBI – European Bioinformatics Institute

f_d – fraction of introgression

F_{ST} – fixation index

GLM - generalized linear model

GO – gene ontology

GPS – global positioning system

hDNA – historical DNA

HMM – hidden markov model

IBS – identity by state

ka – thousands of years ago

kb – kilobase pairs

LD – linkage disequilibrium

LGM – last glacial maximum

Ma – millions of years ago

MAF – minor allele frequency

MAP – maximum a posteriori

Mb – megabase pairs

MCMC – Markov chain Monte Carlo

MHC – major histocompatibility complex

MLE – maximum likelihood estimates

MQ – mapping quality

mtDNA – mitochondrial DNA

NCBI – National Center for Biotechnology Information

Ne – effective population size

NHC – natural history collection

NJ – neighbor-joining

PCA – principal component analysis

PCR – polymerase chain reaction

PE – paired end

Pi – nucleotide diversity

qPCR – quantitative PCR

RNA-seq – ribonucleic acid (RNA)-sequencing

ROH – runs of homozygosity

SD – standard deviation

SE – standard error

SNP – single nucleotide polymorphisms

TMRCA – time to the common ancestor

VCF – variant call format

WGS – whole-genome sequencing

μ – mutation rate

∆AF – allele frequency differences

Chapter I.

General Introduction

In 1859, Charles Darwin wrote, “It is not the strongest of the species that survive, nor the most intelligent, but the most responsive to change” (Darwin 1859). Hundreds of years later, understanding the mechanisms that allow species to adapt to change still is one of the seminal questions in evolutionary biology. Why do some species go extinct, while others manage to find rapid solutions to live and persist in new environments? Uncovering the genetic basis of adaptive traits is now an important research field that has been answering this question.

Recent advances in sequencing technologies and computational tools have allowed breakthrough discoveries in this research field within the last decade. These were initially driven by their relevance to our own lives, as humans display striking traits of local adaptation to extreme lifestyles. Classic examples are illustrated by the resistance to malaria due to mutations in beta hemoglobin in populations where malaria is endemic (Kwiatkowski 2005) and by populations living in high altitudes under chronic hypoxia conditions owing to the EPAS1 gene that regulates hemoglobin concentration (Yi et al. 2010). More recent efforts have been made to understand adaptation to extreme natural settings in animal species, including some that are extinct. The latter has been possible with the advent of ancient genomics (Der Sarkissian et al. 2015), which has uncovered the adaptive potential of several species. A remarkable example was the disclosure of why woolly mammoths (*Mammuthus primigenius*) went extinct. The speed of climate change and the inability of mammoths to adapt quickly enough when the landscapes transformed were the cause (Wang et al. 2021).

For at least 50,000 years, humans have been transforming landscapes around the world at an unprecedented speed (Sullivan et al. 2017). Thus, it is of utmost urgency to understand the evolutionary mechanisms shaping species’ adaptation to the recently human-modified landscapes, which so far have received less attention (Beckman et al. 2022).

1.1. Evolutionary adaptation in an anthropogenic world

Humans have dramatically changed the planet for our own benefit, or detriment, resulting in profound ecological, evolutionary, and environmental changes (Sullivan et al. 2017), including megafaunal extinctions (Broughton and Weitzel 2018; Smith et al. 2018). Anthropogenic changes can have devastating effects, resulting not only in environmental degradation and habitat loss (Sih et al. 2010) but also impacting the way species interact with one another (Robertson et al. 2013). In light of these changes, many species, especially large mammals, have vanished or decreased considerably their abundance and distribution (Díaz et al. 2019). Others, however, have adapted to thrive in close proximity to humans by changing their distribution, behavior, and morphology (Yeh and Price 2004; Partecke et al. 2006; Hunter 2007; Mueller et al. 2013; Harpak et al. 2021; Beckman et al. 2022). For example, a population genetic approach comparing urban and rural populations of the common blackbird (*Turdus merula*) found signs of positive selection in genes associated with circadian rhythms, harm avoidance, and migratory and exploratory behavior in response to urbanization events (Mueller et al. 2013). Brown rats (*Rattus norvegicus*) also represent an extraordinary case of successful adaptation to urbanization, displaying genomic signatures of positive selection in genes associated with metabolism, diet, neurodevelopment and locomotory behavior (Harpak et al. 2021). Another example of high tolerance to disturbance from human activities is illustrated by a brown bear subspecies (*Ursus arctos marsicanus*) living in the Apennine mountains in Italy. Benazzo et al. (2017) sequenced the complete genome of some of these individuals and found several fixed differences across the genome comparatively with non-Apennine bears in genes related to the digestion of dietary carbohydrates and with tame/aggressive behavior. A behavioral change toward a less aggressive temperament potentially reduced the risk perceived by local human communities, allowing bears to persist and live in proximity to humans (Benazzo et al. 2017). The parallels between these studies are particularly noteworthy, with similar genes or gene functions being targeted by positive selection, which may suggest a process of convergent evolution in response to human-modified environments.

Large carnivores are one of the most affected groups of species by humans (Chapron et al. 2014). A continuum history of conflict and direct persecution resulted in a considerable reduction in their abundance and distribution (Ripple et al. 2014), and a range of large carnivore species is currently sharing landscapes with humans around the world (Carter and Linnell 2016). Several studies have described striking behavioral

responses of large carnivores living in human-modified environments. For example, cougars (*Puma concolor*) preferably use riparian woodlands in California to avoid areas of high human density (Dickson et al. 2005); tigers (*Panthera tigris*) in Nepal and leopards (*Panthera pardus*) in India, increased their nocturnality in response to human disturbance (Carter et al. 2012; Gaynor et al. 2018), and gray wolves (*Canis lupus*) in Italy changed their routes in response to seasonal and circadian variation in human activity (Mancinelli et al. 2019). This evidence shows how adaptable large carnivores can be to human disturbance, however, adaptation in this context is likely associated with plasticity instead of natural selection. In fact, few studies on large carnivores have shown direct evidence of an evolutionary response toward human disturbance involving genetic adaptation (e.g., Benazzo et al. 2017).

1.1.1. Genetic diversity and signatures of neutral and selective events

A direct consequence of reduced population size is an increased susceptibility to experiencing low levels of genetic diversity, inbreeding, and genetic drift (Lande 1993). When the number of individuals is significantly reduced in a population, the chances of inbreeding increase, which will decrease the levels of genetic diversity (Keller and Waller 2002; Charlesworth and Willis 2009). Low levels of diversity can rapidly become a threat by leaving populations without the genetic capacity to face environmental change and more prone to random stochastic demographic processes (Lande and Shannon 1996; Reed and Frankham 2003). In such situations, genetic drift tends to prevail over natural selection, limiting adaptation and decreasing the frequency of beneficial mutations (Whitlock et al. 2000). All of these can decrease individual fitness and drive a species to high extinction risk (Kramer et al. 2009).

Empirical evidence relative to the genomic impact of long-term population decline has been described for iconic species, such as the mountain gorilla (*Gorilla beringei beringei*) (Xue et al. 2015) and the Iberian lynx (*Lynx pardinus*) (Abascal et al. 2016), and highlights the importance of maintaining genetic diversity within populations. However, associating genetic factors as the causative link for past extinctions remains a difficult task. Historical and ancient DNA have been revealed to be invaluable resources to address this issue, opening a window of understanding on past biodiversity. With recent advances in sequencing technologies and methodological procedures, the application of these types of resources has rapidly expanded to a population genomic level (Orlando et al. 2015). For example, whole-genome sequencing (WGS) of historical samples of lions (*Panthera leo*) from North Africa, where they are currently extinct, revealed extreme

levels of genomic erosion compatible with the population decline occurring at the time (de Manuel et al. 2020). The same sequencing technique when applied to samples of extinct and living rhinoceros (*Rhinoceros* spp.) has allowed for the reconstruction of their evolutionary histories and subsequently revealed that while rhinoceros genomes in general exhibit low levels of genome-wide diversity, the lowest variation was found in the genome of contemporary species (Liu et al. 2021). This is an example of how the continued loss of genetic diversity through evolutionary time has recently been exacerbated in response to anthropogenic change.

While the importance of maintaining genetic diversity is unquestionable, some examples of species persisting at extremely low levels of genomic diversity have been documented. This is the case of the Channel Island fox (*Urocyon littoralis*) that has persisted for thousands of years at extremely small population sizes with dramatic low levels of genome-wide variation, but with no signs of inbreeding depression (Robinson et al. 2016). Their enigmatic long-term persistence is apparently associated with the purging of deleterious recessive mutations (Robinson et al. 2018). The same mechanism was proposed for the Iberian lynx as an evolutionary strategy to mitigate the effects of inbreeding depression (Kleinman-Ruiz et al. 2022). Such studies suggest that selection can still operate in populations with extremely low sample sizes and low variation. Disentangling the signatures of selection and random genetic drift can be challenging in these cases. In small and isolated populations, genetic drift can be a key driver of changes in genetic diversity and even lead to speciation (Grueber et al. 2013; L. Wang et al. 2016). However, if selection pressures are strong, natural selection can act on genetic diversity in opposition to, or in combination with, genetic drift (Clegg et al. 2002). An important distinction, on which most of the studies aiming to find evidence of natural selection are based, is that demographic processes affect all loci in the genome in a similar way, whereas the effects of selection are restricted to specific loci (Bamshad and Wooding 2003; Nielsen 2005). A current common practice in the field to confirm signatures of selection is to perform demographic simulations with a range of parameters that best approximate population histories and estimate the critical values that support rejection of the neutral model (Wall 1999). Finally, an understanding of the predicted functional impact of selection in populations will be required to associate a genotype with a specific trait.

1.1.2. Gene flow as a source of novel genetic variation

Gene pools eroded by strong selection and/or genetic drift can be replenished with novel genetic variation by *de novo* mutations, recombination, and gene flow. The latter is a particularly relevant evolutionary response in the context of anthropogenic change. As landscapes are transformed, geographical and/or reproductive barriers are being disrupted and species that were formerly isolated are now getting into close contact, which can result in interspecific hybridization. Hybridization following human-mediated disturbance has been widely documented across several taxa (Grabenstein and Taylor 2018). An excellent example is provided by a recent study involving a population genomics approach on Alpine whitefish (*Coregonus* spp.), and using historical and contemporary samples before and after anthropogenic disturbance (Frei et al. 2022). Before human disturbance (i.e., eutrophication of the lake that led to the loss of deep-water spawning habitat) four whitefish species were found to form well-defined groups despite complete sympatry (one of these species is now extinct); however, post-eutrophication, all individuals from the three extant species showed high levels of admixture. Interspecific hybridization raises several concerns. The disruption of reproductive barriers can culminate in hybrid swarms, where high levels of admixture can obscure previously existing independent lineages (Brennan et al. 2014). Although the genetic legacy of such lineages is still segregating in the hybrid population, and therefore cannot be considered genetically extinct, the concept of single species becomes lost. This process was called speciation reversal (Seehausen 2006), and when mediated by anthropogenic disturbance was suggested to cause faster contemporary extinction rates than demographic decline alone (Seehausen et al. 2008). On the other hand, under habitat change, hybrids can have enhanced evolvability and increased fitness (Grant and Grant 2019).

Hybridization is a potential force that can dramatically alter the course of evolution (Barton 2008; Abbott et al. 2013). If hybridization produces offspring that are viable and fertile, hybrids can reproduce with members of the parental species producing backcrossed offspring. Following the reproduction of backcross hybrids with the same parental species, genetic variation from one of the species (donor) is passed and incorporated into the genome of the other (recipient). This process is known as “introgression” and can be a very powerful mechanism to allow adaptation (Aguillon et al. 2022). Introgressed variants have the advantage of being pre-tested in the donor species, hence, unlike *de novo* mutations, genetic introgression is not a strictly blind

process – gene variants that worked well in one species may also work equally well in the recipient species. Furthermore, in contrast to (point) mutations, introgression can cause a wholesale change of the genome, affecting thousands of nucleotides in a single event. In the era of high-throughput sequencing, where an increased number of genomes are being sequenced, researchers have come to the conclusion that introgressive hybridization is not as rare as initially thought, and instead is pervasively present in the evolutionary history of several species (Taylor and Larson 2019). As such, the most plausible explanation is that the introgression of specific variants provides increased adaptive potential in the recipient species, a process known as adaptive introgression. Proving adaptive introgression still is extremely challenging (Arnold and Kunte 2017), yet several studies have convincingly reported adaptive introgression following hybridization. One of the most emblematic cases within and outside the scientific community refers to adaptive introgression from Neanderthal DNA into the modern human genome (Prüfer et al. 2017), which has contributed to the adaptation to new environments and pathogens found in Eurasia following the out of Africa dispersal (Deschamps et al. 2016). Several other cases of adaptive introgression are being repeatedly reported in plants and animals, improving our understanding of the importance of this evolutionary force in different contexts. Classic examples in animals include the introgression of alleles in the Agouti gene that confer brown coat color from black-tailed jackrabbits (*Lepus californicus*) into snowshoe hares (*Lepus americanus*) in habitats with mild winters and low snow cover (Jones et al. 2018; Giska et al. 2019); the introgression of structural variants that control mimicry patterns between *Heliconius* butterfly species (Jay et al. 2018; Edelman et al. 2019), and the introgression of genes involved in pesticide resistance in mice from *Mus spretus* to *Mus musculus domesticus* (Liu et al. 2015). All these studies relied on whole-genome data to uncover and characterize signatures of adaptive introgression. Ranging from single loci of large effect through *cis*-regulatory mechanisms (Jones et al. 2018; Giska et al. 2019) to complex structural rearrangements (Jay et al. 2018; Edelman et al. 2019), all these studies were built based on easily recognizable traits with clear adaptive function, which facilitates the association between genotype and phenotype.

Overall, recent efforts have been made toward furthering the understanding of the evolutionary mechanisms that allow species to adapt, or not, to human-modified landscapes. We have witnessed a wide variety of evolutionary responses to anthropogenic change, but most of these case studies suggest a convergence of

selective pressures for similar phenotypes. Anthropogenic change can also promote gene flow between formerly (genetically) isolated species. While this can expose species to a high extinction rate due to the formation of hybrid swarms, enhanced adaptive potential is equally possible. The previously described evidence of adaptive introgression shows that this process is not as rare as previously anticipated and that it, in fact, can play a crucial role in shaping the evolution of species. Despite the challenges in establishing genotype-phenotype associations, and in linking them with adaptive introgression, the increased genomic resolution has allowed deep insight into several systems, as demonstrated by the studies in hares (Jones et al. 2018; Giska et al. 2019), butterflies (Jay et al. 2018; Edelman et al. 2019) and mice (Liu et al. 2015).

1.2. Genomics of animal domestication

In the context of species adapting to anthropogenic induced change, animal domestication can be seen as a special case of adaptation (Beckman et al. 2022). Domestic dogs (*Canis lupus familiaris*), as the first species to be domesticated, have received particular attention. Although remarkable discoveries have been made in the last few years, there is still no consensus in the field about the time and geographical origin of dog domestication from their wild ancestor, the gray wolf. The earliest dog-like remains were found in Europe in the Czech Republic (Germonpré et al. 2012) and Belgium (Germonpré et al. 2009) with ~27,000 and ~36,000 years old, respectively. Genetic analysis of these, and several other dog-like fossils, later suggested that some were misclassified and instead likely belong to an extinct wolf lineage (Crockford and Kuzmin 2012; Morey 2014). Over the years, the geographical origin of dog domestication has been placed in East Asia, based on the mitochondrial DNA (mtDNA) control region (Savolainen et al. 2002), and later using whole-genome data (Wang et al. 2015); in the Middle East, based on autosomal single nucleotide polymorphisms (SNP) (Vonholdt et al. 2010), and in Europe, using complete mitochondrial genomes (Thalmann et al. 2013). An important finding was later achieved using entire genomes of several dogs and wolves from distinct worldwide populations (Freedman et al. 2014; Fan et al. 2016), suggesting that dogs likely originated from a now-extinct wolf lineage. Skoglund et al. (2015) sequenced the first whole ancient genome of a 35,000-year-old wolf from northern Siberia (named “Taimyr wolf”) and showed that this individual belonged to a population that diverged from the common ancestor of present-day wolves and dogs and has contributed to the ancestry of northern high latitude dog breeds. The authors also

suggested that dogs must have originated from multiple regional wolf populations. The multi-origin theory was corroborated by Frantz et al. (2016), who sequenced an ancient genome of an Irish dog from the Bronze age to reveal a deep split between ancient and modern dogs in East Asia and Western Eurasia. A study published very recently performed the most comprehensive analysis by sequencing 66 new ancient wolf genomes from around the world, spanning the past 100,000 years (Bergström et al. 2022). The idea was to exhaustively map the genetic diversity of wolf populations across space and time to understand which populations were closest to dogs. Bergström et al. (2022) proposed that although dogs are overall more closely related to ancient wolves from eastern Eurasia, an independent domestication process in the West cannot be discarded. Hence, despite these major advances, the geographical origin of the wolf population that gave rise to dogs somewhere between 40 and 14 thousand years ago (ka) (Skoglund et al. 2015; Fan et al. 2016; Frantz et al. 2016; Perri, Feuerborn, et al. 2021; Bergström et al. 2022) remains to be discovered.

Following dog domestication, numerous other animal species were domesticated, between 11 to 4 ka, including sheep (*Ovis aries*), goats (*Capra hircus*), cattle (*Bos taurus*), pigs (*Sus scrofa*), chickens (*Gallus gallus*), rabbits (*Oryctolagus cuniculus*), cats (*Felis catus*) and horses (*Equus caballus*) (Frantz et al. 2020). Several genomic resources have been made available for these species, including ancient DNA, which has furthered our understanding of the evolutionary mechanisms underlying their domestication (MacHugh et al. 2017).

1.2.1. Origin of the common traits shared by domestic animals

The domestication process of animal species was underlined by independent mechanisms and purposes (Frantz et al. 2020), however, there seems to be a common suite of biological traits shared among them (e.g., increased tameness, prolongation in juvenile behavior, coat coloration, reduced cranial volume, floppy ears, more frequent and nonseasonal estrus cycles and changed concentrations of specific neurotransmitters), known as the “domestication syndrome” (Darwin 1868; Faegri 1981) (**Figure 1.1**). The theory behind it predicted that selection for less aggressive behaviors alone would result in the set of changes that characterize domestic animals (Belyaev 1979). This theory was initially supported by the “Belyaev Fox Farm experiment” (Trut 1999), in which foxes (*Vulpes vulpes*) that were selectively bred for tame behaviors over 50 generations started to display the set of traits that constitute the domestication syndrome (Trut et al. 2009). However, a recent study has suggested that most of the

traits found in tame foxes predated the breeding experiments, which undermined their association with the domestication syndrome as a proof of concept (Lord et al. 2020); nevertheless, the authors recognized that these two distinct fox populations represent an invaluable system to study the genetic basis of behavior. Extensive work on this system involving genomic resources has been produced over the last few years, which has provided important insights into the genetic heritability of the traits that characterize tame and aggressive foxes. For example, two recent studies applying genomic (Kukekova et al. 2018) and transcriptomic (Wang et al. 2018) approaches revealed high allele frequency changes in their genomes, as well as significant brain gene expression differences between aggressive and tame foxes, respectively.



Figure 1.1. Pictures of several animal domesticated species and tame foxes from the "Belyaev Fox Farm experiment", showing similar phenotypic traits, such as changes in coat coloration, floppy ears, and increased tameness (shown in different columns). This common set of biological traits is among those that constitute the "domestication syndrome". Figure adapted from Trut et al. 2009.

One of the most popular hypotheses proposed for the genetics underlying common traits in domesticates was the neural crest cell hypothesis (Wilkins et al. 2014). In it, an initial selective pressure for tameness would lead to deficits in neural crest cells during

embryonic development, which would indirectly produce the morphological changes described in the domestication syndrome (**Figure 1.1**). However, this hypothesis was recently discredited (Johnsson et al. 2021) after the robustness of the domestication syndrome, as the extent to which the fox experiment could be used to validate it, was questioned (Lord et al. 2020). Nevertheless, there seems to be a consensus that behavioral shifts toward reduced aggressiveness, as well as increased social tolerance toward humans, were the primary targets of selection during domestication (Sato et al. 2020; Andersson and Purugganan 2022). For example, in the case of dogs, the most widely accepted theory is that less aggressive and fearful wolves would have started to approach human settlements to profit from food leftovers (Larson et al. 2012). As such, because selection against aggression was natural and without direct human intervention, it was proposed that wolves domesticated themselves; in other words, to have self-domesticated (Hare et al. 2012; Herbeck et al. 2018). Self-domestication was firstly described in bonobo apes (*Pan paniscus*), which display less severe forms of aggression to conspecifics (Hare et al. 2012) and posteriorly associated with human evolution (Theofanopoulou et al. 2017). In the second stage of domestication, humans would have started to selectively breed wolves with lower levels of aggression, resulting in heritable genetic changes and marked divergence in behavior between wild and domestic forms (Freedman and Wayne 2017).

1.2.2. Genetic basis of the domestic behavioral phenotype

Up until today, no obvious candidate gene was identified to explain the common behavior of domestic animals. With the availability of several genomic resources for domestic animals (commercial array chips, high-quality reference genomes, and variant data) a common practice for finding “domestication genes” has been to compare the genome of both wild and domesticates and identify highly divergent regions. The rationale behind it is that a substantial fraction of the domestic animal’s genome has a high similarity to that of their wild ancestors, hence, genomic regions of strong differentiation between them must be linked to the observed phenotypic differentiation. Several studies followed this approach to uncover candidate genes in domestic chickens (Rubin et al. 2010), pigs (Rubin et al. 2012), dogs (Axelsson et al. 2013; Li et al. 2014), sheep (Kijas et al. 2012), rabbits (Carneiro et al. 2014) and foxes (Kukekova et al. 2018), revealing shifts in allele frequencies at many loci. An important observation made in these studies, particularly in dogs and rabbits, was that most of the outlier loci were located in noncoding regions, suggesting that alterations in gene regulation, and not in the protein-coding sequence,

might be underlying the phenotypic changes between domestic and wild animals. Saetre et al. (2004) was a pioneer in studying gene expression profiles between wild and domestic animals, using the microarray technology in brain tissues of dogs and wolves to uncover the genetic basis of behavior. The authors suggested that changes in brain gene expression may have been the key to rapid adaptation. The advent, and subsequent rapid evolution, of RNA-sequencing (RNA-seq) technology, made it possible to generate, characterize and compare complete transcriptomes (Wang et al. 2009), promising to revolutionize our understanding of the regulatory mechanisms that shaped domestication. Using this technology, Albert et al. (2012) performed the first comparative approach of transcriptomes derived from the frontal cortex of three pairs of wild-domestic animals, including dogs and wolves, pigs and wild boars, domestic and wild rabbits, in an attempt to find overlap genes to the different domestication events. However, the authors found almost no overlap and proposed that the majority of brain gene expression changes were specific to the given domestication event.

Over the last years, several studies employing genomic and transcriptomic approaches to single pairs of wild-domestic animals have identified a repeated set of genes, gene families, and pathways, involved in neurotransmission and cognition (e.g., **dog**: Wang et al. 2013; Li et al. 2014; Freedman et al. 2016; Pendleton et al. 2018; **tame fox**: Hekman et al. 2018; Kukekova et al. 2018; Wang et al. 2018; **horse**: Schubert et al. 2014; **pig**: Frantz et al. 2015; Li et al. 2015; **rabbit**: Carneiro et al. 2014; Sato et al. 2020; **cattle**: Qanbari et al. 2014; Eusebi et al. 2018; **cat**: Montague et al. 2014; **chicken**: Bélteky et al. 2016; Katajamaa and Jensen 2020; **sheep**: Naval-Sanchez et al. 2018, and **goat**: Dong et al. 2015; Alberto et al. 2018; Bertolini et al. 2018). The hypothesis of a convergent evolutionary mechanism underlying the domestic behavioral phenotype was recently addressed by Hou et al. (2020), who investigated common signatures of selection and gene expression across several domestic animals. Strong convergent patterns emerged in a gene family encoding glutamate ionotropic receptors, which corroborated the results of the independent studies showing multiple signatures of selection in glutamate genes (e.g., Carneiro et al. 2014; Li et al. 2014; Montague et al. 2014; Wang et al. 2016; Wang et al. 2018). Glutamate is the primary excitatory neurotransmitter in the vertebrate central nervous system, and it regulates neuronal excitability and synaptic plasticity (O'Rourke and Boeckx 2020). More importantly, glutamate genes have been associated with domestication-related behaviors such as fearfulness, exploration tendencies, anxiety and stress responses, and altered learning and memory capabilities in several studies using model vertebrate species (Schafe et al. 2001; Sheng and Kim 2002; Ko et al. 2005; Bannerman et al. 2014; Babenko et al. 2016;

Forrest et al. 2018). The role of glutamate receptors in domestication was recently reviewed in detail by O'Rourke and Boeckx (2020), who identified the genes GRIK3, GRM8, and ADGRB4 as those with signals of selection across more animal domestic species. These two studies that investigated convergent molecular responses in several domesticates seem to suggest a polygenic background for animal domestication, and although distinct genes tend to be selected in different species, there is a repeated targeting of gene families and gene pathways that encode neurotransmission.

Despite major advances in the last decade, genomic studies on animal domestication have shown that several questions remain to be answered, including the geographical origin and timing of the first domestication event and the genetic basis for the domestic behavioral phenotype. Multiple and repeated evidence of selection in genes associated with neurotransmission across domestic animals suggests a crucial role of behavior in the primordium of domestication. Future research in ancient genomics has the potential to reconstruct in detail the evolutionary histories of domestic animals, and with exciting recent developments in ancient transcriptomics (e.g., sequencing of RNA data from a 14,300-year-old canid; Smith et al. 2019), one can expect important findings about the regulatory mechanisms that shaped the transition from wild to domestic animals.

1.3. Hybridization between wild and domestic animals

In the face of increasingly human-modified environments, domestic animals may have enhanced adaptive potential, as they possess many traits that were directly selected by humans to allow co-existence. As such, some of these traits could potentially allow adaptation of their wild counterparts to anthropogenic change through genetic introgression (Beckman et al. 2022). Contrary to the previous belief that animal domestication required complete reproductive isolation between domestic and wild lineages, many studies have shown that pervasive events of gene flow happened following divergence in several species (reviewed by Marshall et al. 2014; Frantz et al. 2020). For example, studies on wild-domestic pigs have found extensive levels of introgression in the domestic species genome, however, the domestic phenotype seems to be maintained due to "domestication islands" in the pig genome that prevent introgression of maladaptive wild boar genomic variants (Frantz et al. 2015). On the other hand, there are also cases where hybridization between wild and domestic species has

disrupted the phenotypic differentiation between the two, resulting in highly admixed populations. This is the current situation of the wildcat (*Felis silvestris*) in Scotland, that due to extensive hybridization with domestic cats, almost all individuals show some level of admixture, resembling a hybrid swarm (Howard-McCombe et al. 2021).

Hybridization between wild and domestic species is considered to be of a different nature from that occurring naturally and is often referred to as “anthropogenic hybridization” (Allendorf et al. 2001). Anthropogenic hybridization is primarily seen as a threat to the conservation of wild species, as the introgression of domestic genomic variants into the wild species genome can promote genetic homogenization and disrupt local adaptations to the wild (Randi 2008). Domestic animals possess only a fraction of the genetic variability present in wild populations due to the initial domestication bottleneck, and due to strong subsequent artificial selection during breed formation (Frantz et al. 2020). The latter also led to an increase in the frequency of deleterious mutations in the genome of domestic animals, inflating the mutational load (Cruz et al. 2008; Charlier et al. 2016; Marsden et al. 2016; Derks et al. 2019; Fages et al. 2019; Orlando and Librado 2019). While the consequences of high mutational load may be masked in domestic animals due to human intervention (e.g., animal care, breeding strategies), its effect on small wild populations can be extremely severe and reduce individual fitness. This can also affect the efficiency of purifying selection in wild populations, and hence the capacity to purge maladaptive domestic variants. Predicting the outcome of hybridization, whether the populations will maintain distinct or become a hybrid swarm, such as that of the wild and domestic cats in Scotland, is very difficult as it will depend on many factors. To start, it will be strictly correlated with the frequency and extent to which hybridization occurs, and with the size of the wild population. Very often the size of the domestic population tends to exceed largely the size of the wild population, which increases the chances of encounters and consequently, hybridization. Understanding the dynamics of hybridization in the past might be a valuable way to predict its effects in the future (Chaturvedi et al. 2020). The use of ancient and historical samples of wild and domestic animals can play a crucial role to estimate levels of admixture in the past, but also to identify potential drivers of hybridization (Grabenstein and Taylor 2018).

In the last two decades, several cases of adaptive introgression have also been reported between wild and domestic animals, which has raised the debate about the ecological and evolutionary role of hybrids in the wild. The most emblematic case refers to the introgression of a beta-defensin gene, CBD103, from domestic dogs into North American gray wolves that confer them a black coat color (Anderson et al. 2009). Initially, it was

thought that black wolves would have increased fitness because of improved camouflage in forested habitats (Anderson et al. 2009), however, it was later proposed that the target of natural selection could be in fact immunological (Schweizer et al. 2018), as CBD103 also has antimicrobial activity against respiratory pathogens (Erles and Brownlie 2010). Adaptive introgression in the opposite direction was also described, in a study suggesting that Tibetan Mastiff dogs successfully adapted to high altitudes because of introgression of the hypoxia adaptive gene, EPAS1, from Tibetan gray wolves (Miao et al. 2017). In this regard, a recent study generated hundreds of genomes from gray wolves and dogs inhabiting high altitudes and found evidence that the EPAS1 was derived from an as-yet-unrecognized wolf-like lineage (Wang et al. 2020). Whether the gene introgressed independently into both Tibetan wolves and dogs, or sequentially, is still not clear. Similarly, the adaptation of both domestic sheep and pigs to high altitudes is likely to have been underlined by adaptive introgression from their wild counterparts in genes related to oxygen transportation (Ai et al. 2015; Hu et al. 2019). All these reported cases of adaptive introgression were observed despite complete genetic differentiation between wild and domestic species (i.e., were not hybrid populations). This implies that, regardless of interbreeding between wild and domestic species, selective mechanisms (purifying and positive selection) operate together in the genome to maintain both phenotypes distinct. It is important to note that hybridization and introgression are very distinct phenomena, and with different outcomes in the evolution of a species. Although both terms are often used interchangeably, most hybridization events do not result in genetic introgression. The time scale on which they can be detected also differentiates the two; while dynamics of hybridization tend to characterize rates of gene flow at the time the population was sampled, evidence of adaptive introgression requires many generations after the event to be traceable (Taylor and Larson 2019; Aguillon et al. 2022).

1.3.1. The case of wolf-dog hybridization

Under the scope of this dissertation, it is of particular interest to explore in more detail the role of hybridization between gray wolves and domestic dogs. The genetic integrity of wolf populations has been for decades a real concern among conservationists (Blanco et al. 1992; Boitani 1993; Donfrancesco et al. 2019; Salvatori et al. 2020), raising the need to have the ancestry of gray wolves unfolded. Although there is documented evidence of gene flow between gray wolves and dogs following domestication, little is known about the frequency, and extent, to which hybridization occurred in the past. Their

genomes, each distributed across the same number of chromosomes (38 pairs of autosomes and 2 sex chromosomes, the X and Y), are remarkably similar (0.04-0.21% of nucleotide differentiation) (Lindblad-Toh et al. 2005), and this makes the identification of backcross hybrids a difficult task. From the use of the mtDNA control region (Vilà et al. 1997) to the current sequencing of entire genomes (Smeds et al. 2021) within wolf-dog hybridization studies, our understanding of the process has significantly improved.

Episodes of wolf-dog hybridization have been reported in North America (vonHoldt et al. 2011), but have been more exhaustively studied, and reported, throughout Europe (Vilà et al. 2003; Godinho et al. 2011; Hindrikson et al. 2012; Randi et al. 2014; Galaverni et al. 2017; Pacheco et al. 2017; Kusak et al. 2018; Pilot et al. 2018; Smeds et al. 2021; reviewed by Dziech 2021). Taken together, these studies have shown that:

(i) hybridization between wolves and dogs is associated with the disruption of the social structure within wolf packs and it is most frequent in peripheral areas of the wolf range in addition to areas near human settlements where dogs are more frequent. For example, Galaverni et al. (2017) identified an increased rate of hybridization between dogs and Italian wolves after a severe bottleneck in the 1970s which decimated almost all the wolf population. In the Iberian Peninsula, most of the individuals identified with a hybrid origin were found in peripheral areas of the wolf range, where the wolf population was expanding (Godinho et al. 2011).

(ii) mtDNA and Y-chromosome analysis revealed that hybridization between wolves and dogs typically occurs between female wolves and male dogs (Vilà et al. 2003; Godinho et al. 2011; but see Hindrikson et al. 2012 and Randi et al. 2014 for exceptions). This asymmetric pattern has been explained by distinct breeding physiology of wolves and dogs. While female and male wolves have a single breeding period per year, male dogs can breed throughout the year (Asa and Valdespino 1998) and are thus potentially always available to breed with solitary wolf females on estrus (Vilà and Wayne 1999).

(iii) phenotypic traits alone cannot be used as reliable characters for identifying wolf-dog hybrids. Although some morphological features might differ between wolves and hybrids (e.g., hybrids can have higher variation in coat color; white claws instead of the black or dark gray typically observed in wolves, and dewclaw, a relict fifth finger; **Figure 1.2**; Galaverni et al. 2017; Kusak et al. 2018), several studies have shown no clear association with their genomic ancestry. For example, a study investigating the occurrence of wolf-dog hybridization in Croatia has shown that “hybrid-like” individuals with a black and white coat color were in fact wolves, after subsequent confirmation with

genetic analysis (Kusak et al. 2018). Furthermore, the authors estimated that the number of hybrids identified on the basis of phenotype was almost four times higher than that confirmed with genetic analysis.

(iv) the prevalence of wolf-dog hybrids detected within wild populations has been estimated to be close to 3-8% (Godinho et al. 2011; Galaverni et al. 2017; Kusak et al. 2018; Pilot et al. 2018). However, most individuals presented a low proportion of dog ancestry, suggesting that despite past events of gene flow, most of the dog ancestry was purged from the wolf genome after several generations of backcrossing with other wolves.

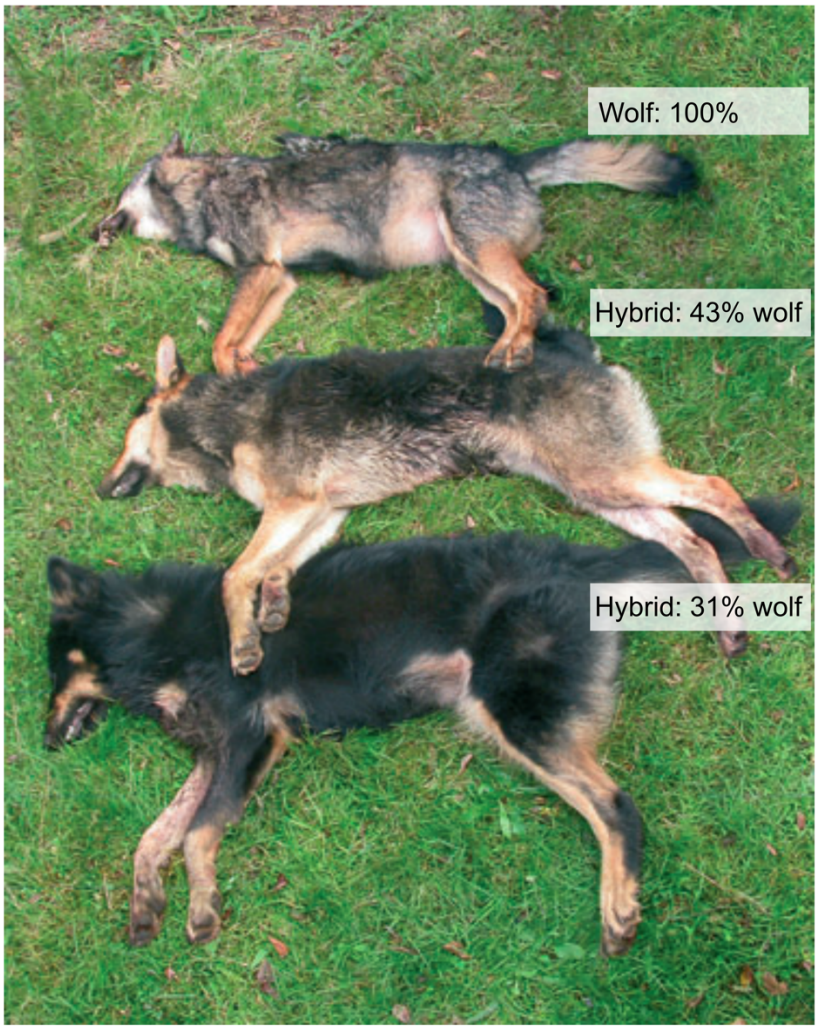


Figure 1.2. Picture of an Iberian wolf (*Canis lupus signatus*; top) and two wolf-dog hybrids (middle and bottom) found within the same pack in Sierra de Penouta, Western Asturias, Spain. Individual assignments to the wolf cluster estimated through Bayesian admixture analyses are presented (Godinho et al. 2011). Phenotypic differentiation between the three

individuals can be seen particularly in the coat color, nose elongation, and body size. Figure adapted from Godinho et al. 2011.

Given the high level of similarity between wolf and dog genomes, it is very difficult to detect hybrids resulting from older generations of backcrossing, even with a dense panel of loci across the genome. In the last years, more studies have relied on the power of single nucleotide polymorphisms (SNPs) to identify wolf-dog hybrids, showing that these can provide more detailed information about past events of hybridization than microsatellites (Galaverni et al. 2017; Pilot et al. 2018; Smeds et al. 2021). Deciding where to trace the line between a “pure” and “hybrid” individual can be extremely challenging, especially when both species are so closely related. As such, defining reliable thresholds based on the ability of the markers used for correct assignments is fundamental (Dziech 2021).

Currently, genomic approaches allow to characterize the local level of ancestry across the genome at a higher resolution, showing, for example, that the amount of dog ancestry in Eurasian wolves can range from ~2% up to 25% (Fan et al. 2016; Pilot et al. 2021). The presence of small genomic regions of dog ancestry within the wolf genome is compatible with ancient events of gene flow and suggests that introgression may be more frequent than initially thought (Gómez-Sánchez et al. 2018). The previously mentioned cases of adaptive introgression between wolves and dogs corroborate this idea. The increased resolution over the wolf and dog genomes has also revealed that different chromosomes seem to retain differential levels of introgression (Galaverni et al. 2017; Gómez-Sánchez et al. 2018; Pilot et al. 2021; Smeds et al. 2021), similar to the “domestication islands” observed in the pig genome (Frantz et al. 2015). This, however, does not seem to be correlated with differential recombination rates along the genome, nor with barriers to introgression between wolves and dogs (Pilot et al. 2021). Different selection mechanisms, together with genetic drift, may be responsible for such signatures, as previously proposed for other mammals (Geraldes et al. 2006; Good et al. 2010; Turner and Harr 2014; Frantz et al. 2015).

Years of research into the interaction between wild-domestic animals have revealed multiple pieces of evidence that imply interbreeding following divergence. Despite this, wild and domestic species seem to remain genetically and phenotypically distinct. Sequencing the entire genome of these species has shed some light on the evolutionary mechanisms that are maintaining such differentiation, revealing regions in the genome

with different permeability to, or retention of, introgressed variants. Analyzing in more detail the genomic regions that appear in wild species genomes with domestic ancestry may allow determining its adaptive potential. Introgression of domestic variants into wild species genome, and vice versa, has already proven to allow adaptation to extreme environments. Therefore, with increased anthropogenic change it is not only expectable that the dynamics of wild-domestic hybridization will change, but also that domestic variants would offer an adaptive advantage to wild species upon introgression in highly human-modified environments. This may be particularly relevant at a time when several wild species are experiencing a reduction in their distribution and abundance.

1.4. Insights into the evolutionary history of gray wolves

The gray wolf is not only the most widely distributed of all large carnivores across the world but also one of the most adaptable (Mech and Boitani 2003). Wolves can endure the extremely low temperatures in the arctic but can also inhabit deserts and high-altitude mountains (**Figure 1.3**).

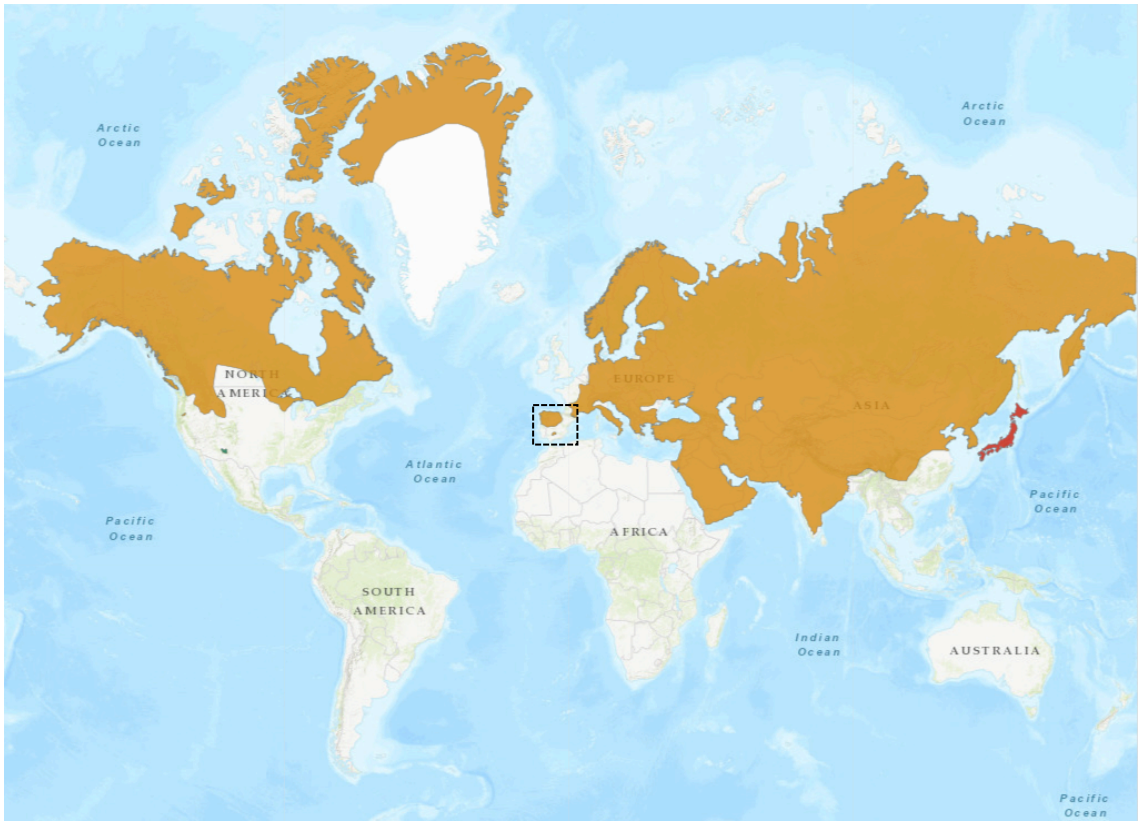


Figure 1.3. Map showing the IUCN geographical range of gray wolves in the world (2018). Orange areas represent the extant distribution, green represents areas where the species was reintroduced, and red represents areas where the gray

wolf went extinct (in this map only Japan is considered as an extinct area, but gray wolves disappeared also from Mexico, most of the USA, and Western Europe). The dashed box highlights the Iberian Peninsula, which includes Portugal and Spain. Map retrieved from the IUCN website, last accessed on 23rd of August 2022.

Gray wolves belong to the genus *Canis*, from the Canidae family (order Carnivora), composed of dogs, coyotes (*Canis latrans*), African wolves (*Canis lupaster*), golden jackals (*Canis aureus*), and the Ethiopian wolf (*Canis simensis*), that together with the dhole (*Cuon alpinus*), African wild dog (*Lycaon pictus*), side-striped jackal (*Canis adustus*) and black-backed jackal (*Canis mesolemas*) form the wolf-like canids monophyletic clade (Lindblad-Toh et al. 2005) (**Figure 1.4A**). The first *Canis* members are believed to have evolved around 5 million years ago (Ma) from the extinct Old-World genus *Eucyon*, followed by a fast expansion and diversification across Eurasia in the Early Pleistocene (Perri, Mitchell, et al. 2021). Gray wolves diverged from coyotes and African wolves at 1.1 and 1.7 Ma, respectively (vonHoldt et al. 2011; Perri, Mitchell, et al. 2021) (**Figure 1.4B**). Following divergence, extensive gene flow has been described between species of the genus *Canis*, including genetic contribution from unknown canid species (i.e., ghost introgression) into the ancestor of the gray wolf and coyote (Gopalakrishnan et al. 2018). Admixture between gray wolves and coyotes in North America has even resulted in newly recognized wolf-like species, like the red wolf (*Canis rufus*) and the Great Lakes wolf (*Canis lycaon*) (vonHoldt et al. 2016).

Gray wolves present a clear worldwide phylogeographic structure, that has formed mostly in the last 30-20 ka when New and Old-World wolves diverged (Fan et al. 2016) (**Figure 1.5A**). Several studies analyzing genomes of Pleistocene wolves have suggested that present-day gray wolves trace their origin to a wolf population that expanded from Siberia during the last glacial maximum (LGM; ~28-23 ka, Hughes and Gibbard 2015) and replaced local wolf populations (Skoglund et al. 2015; Loog et al. 2020). Moreover, there is also evidence supporting pre-LGM extinctions of multiple wolf lineages in Siberia (Ramos-Madrigal et al. 2021). Bergström et al. (2022) recently showed that wolf populations were highly connected throughout the Late Pleistocene, with very low levels of genetic differentiation. The authors also proposed that Siberia acted as a source of migration to Europe but not the other way around, and that ancestry replacement might have been incomplete as fractions of deep European ancestry persisted in present-day wolves. The genetic homogenization in European wolves could explain why Pleistocene wolves appear basal to present-day wolves, instead of the multiple extinction hypothesis (Bergström et al. 2022).

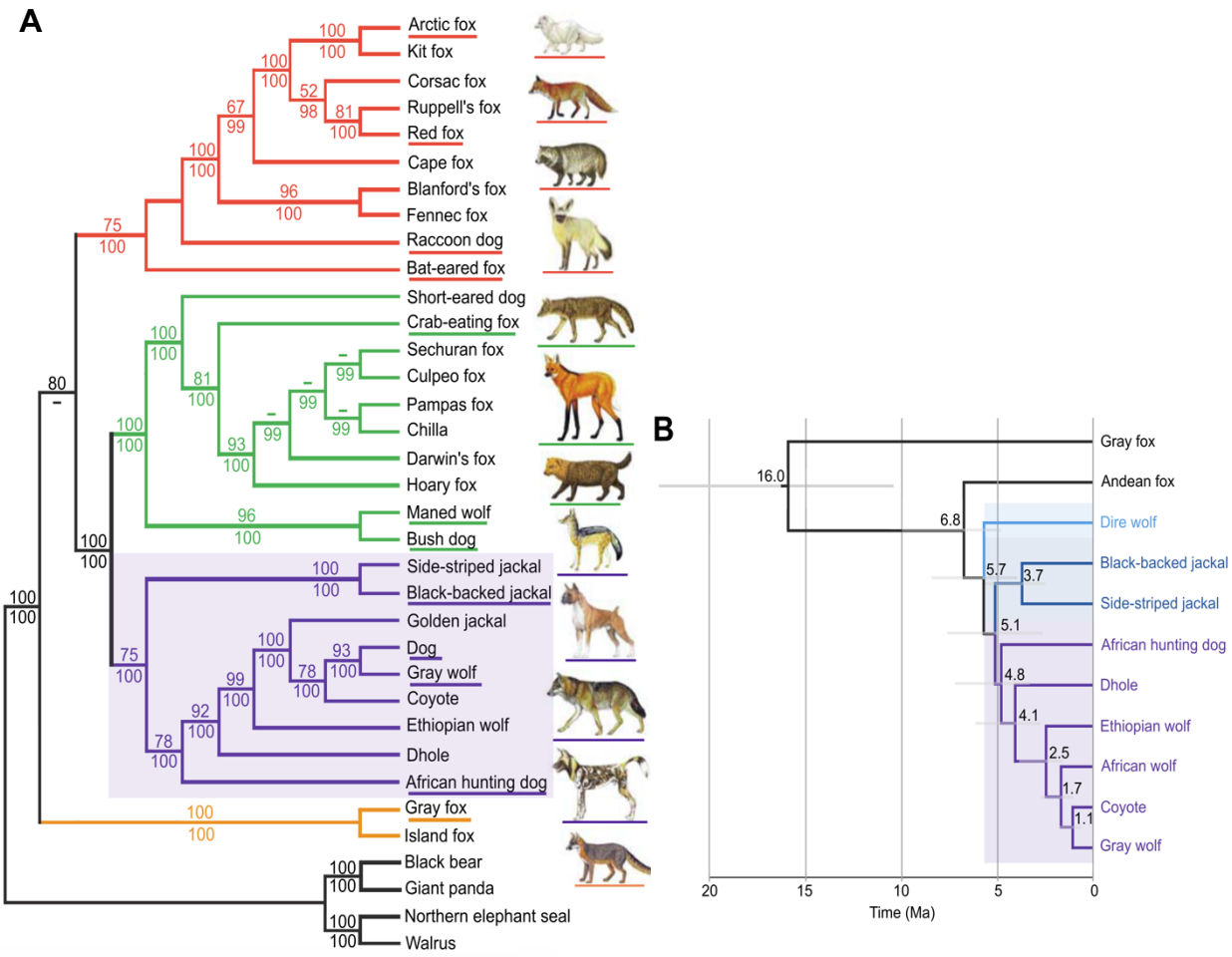


Figure 1.4. Phylogenetic relationships within the wolf-like clade. **(A)** Bayesian phylogenetic tree of the family Canidae based on ~15 kb DNA sequence from exons and introns across the genome. Different clades are shown with different colors: the red-fox-like clade in red, the South-American clade in green, the wolf-like clade in purple, and the gray and island fox clade in orange. Values above and below the internodes represent bootstrap values and Bayesian posterior probability values, respectively (dashes indicate values below 95%). Underlined species names have the corresponding illustration at the right. Figure adapted from Lindblad-Toh et al. 2005. **(B)** Increased resolution of the relationships within the wolf-like clade, including estimates of divergence times in millions of years before the present (horizontal gray bars represent 95% of highest posterior densities) using nuclear sequence alignments of 70 kb to 28 Mb. Note that in this case, both black-backed and side-striped jackals are positioned outside the wolf-like clade. Given their high divergence and distant phylogenetic position, it was recently proposed that their status should be revised and that they should be placed in a distinct genus (Álvares, 2019). An extinct species, the Dire wolf (*Canis dirus*), is also represented in the tree, found to have split from living wolf-like canids around 5.7 Ma. Figure adapted from Perri et al. 2021.

Unlike many other megafauna extinctions, gray wolves manage to persevere during the last Ice Age and have never been close to extinction. Due to its large range and stable

global trend, the gray wolf is currently listed as Least Concern by the IUCN (IUCN 2011). However, in the last few centuries due to heavy human persecution, most of the Central and Western European wolf populations went extinct (Ripple et al. 2014) (**Figure 1.5B**). During the 16th-century, wolves were completely extirpated from England, Scotland, Denmark, and the Netherlands, and later in the early 20th-century from Germany, Switzerland, France, and Scandinavia (Dufresnes et al. 2018). Wolves in Eastern Europe never faced extinction, and despite some local declines, managed to maintain large and outbred populations with genetic contribution from Russian populations, where wolf abundance is higher (Boitani 2003; Chapron et al. 2014). Additionally, small, and fragmented wolf populations persisted in Southern regions of Europe as the Italian and Iberian Peninsulas (Blanco et al. 1992; Randi 2011). Population declines were particularly severe in Italy, where the wolf population was reduced to less than 100 individuals during the 1970s (Boitani 1992). In the Iberian Peninsula, which includes Portugal and Spain (**Figure 1.3**), the population suffered a significant reduction in its distribution range but was never close to extinction (Valverde 1971; Blanco et al. 1990; Clavero et al. 2022; Nores and López-Bao 2022). Due to the small population size and complete isolation for several thousands of years, both Italian and Iberian wolf populations have been reported to have the lowest levels of genetic diversity and the highest levels of differentiation in relation to other European populations (Pilot et al. 2010; Pilot et al. 2014; Fan et al. 2016; Hindrikson et al. 2017). A recent study using whole-genomes has estimated that these two populations have diverged much earlier than the recent human-induced population declines, circa 10.3 ka with negligible gene flow ever since (Silva et al. 2020). Still, a phylogeographical analysis using mtDNA from past and present-day wolf populations has shown that the most dramatic collapse in genetic diversity in Western Europe occurred, in fact, in the 20th-century, strongly suggesting an anthropogenic role (Dufresnes et al. 2018).

In the last decades, there have been records of wolves expanding in Europe and recolonizing several areas, such as the Western Alps and Scandinavia (Chapron et al. 2014; Hindrikson et al. 2017). This recovery, which is mainly associated with the implementation of strong legislation favoring the protection of the species, allowed wolves to recolonize the continent naturally. Wolves are currently distributed across 10 populations (defined for conservation purposes), from Scandinavia, Karelian, Baltic, Central European Lowlands, Carpathian, Dinaric-Balkan, Italian Peninsula, Alpine, Northwestern Iberian Peninsula, and Sierra Morena in Spain (Chapron et al. 2014) (**Figure 1.5B**).

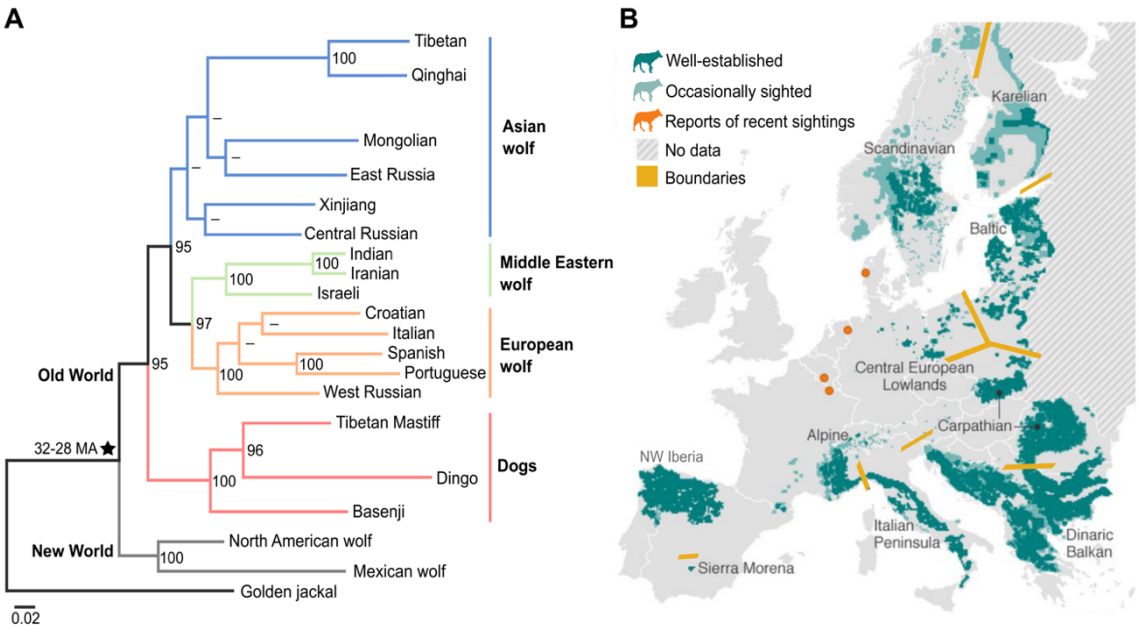


Figure 1.5. Gray wolf phylogeny and distribution across Europe. **(A)** Genome-wide phylogenetic tree based on autosomal SNPs. Estimates of divergence time between New and Old-World wolves are presented. Different colors represent different wolf populations in Eurasia and dogs. Values represent the node support from bootstrapping (dashes indicate values below 95%). Figure adapted from Fan et al. 2016. **(B)** The current distribution of wolves in Europe. Dark green patches indicate areas of permanent wolf occurrence, light green patches indicate areas of sporadic wolf occurrence, and orange dots indicate reports of recent sightings. Yellow lines denote boundaries between the 10 defined wolf populations (Chapron et al. 2014). Figure adapted from BBC/WWF.

1.4.1. The Iberian wolf as a remarkable case of persistence

In the face of fast anthropogenic induced change, and intense human persecution, some European wolf populations perish to extinction. Others, however, managed to persist across increased human-modified landscapes. The Iberian wolf subspecies (*Canis lupus signatus*, Cabrera 1907; **Figure 1.6A**), the largest population in Western Europe, is among the latter.

Similarly to other European countries, wolves in the Iberian Peninsula were heavily persecuted in the last century, mainly due to livestock depredation and competition with humans for wild prey, resulting in strong declines in the population range (Blanco et al. 1992; Ramirez et al. 2006). Until the 1930s, Iberian wolves were distributed across the entire Peninsula, while dramatic declines during the 1970s led to the almost disappearance of wolves in the South and a strong population contraction towards the

North (Valverde 1971; Garzón 1979; Blanco et al. 1990; Rico and Torrente 2000; Clavero et al. 2022; Nores and López-Bao 2022). Today, the population is mostly confined to the Northwestern region of Iberia, with a small population at the South of the Douro River in Portugal and an isolated and critically endangered population in Sierra Morena (Álvares et al. 2005; Blanco and Cortés 2007; López-Bao et al. 2018) (**Figure 1.6B**). The last census estimated that the Iberian population was composed of 2,000-2,500 wolves, organized in ~60 packs in Portugal and ~290 in Spain, over a total area of ~140,000 km^2 (Pimenta et al. 2005; MAGRAMA 2016). A large part of the population occurs in highly humanized areas (20-400 inhabitants/km^2, minimum-maximum values; INE 2016), and in some regions, the wolf density is among the highest in Europe (7 wolves/100 km^2) (Blanco and Cortés 2007; Eggermann et al. 2011). Although wolves are currently fully protected in Portugal and Spain, their annual mortality rate in some areas in Portugal was estimated to be almost 40%, which is similar to that observed in wolf populations subjected to legal harvest (Campos 2018). The causes of death were attributed to humans, through poaching, illegal snares, and collisions with vehicles. Despite the high mortality rates, the population has been expanding and already recolonized regions in Central Spain (Blanco and Cortés 2009; Chapron et al. 2014; López-Bao et al. 2018).

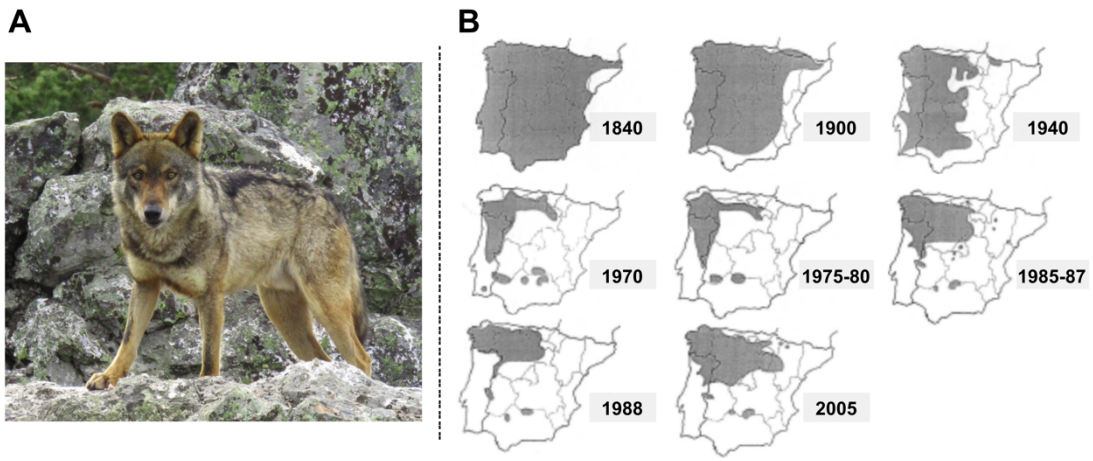


Figure 1.6. Distribution of wolves in the Iberian Peninsula. **(A)** Picture of an Iberian wolf, where is possible to see the distinctive dark marks on its front legs. This characteristic, together with dark marks on the tail (not entirely visible in this picture), were responsible for the designation *signatus*, meaning "marked" or "signed", by Ángel Cabrera in 1907, who proposed that Iberian wolves represent a distinct subspecies. Photo credits: Raquel Godinho. **(B)** Estimated wolf range in the Iberian Peninsula, including Portugal and Spain, between the 19th and 21st century. Figure adapted from Sastre 2011.

Iberian wolves display the most remarkable low dispersal rates seen among gray wolves (**Figure 1.7A**), which contributes to a strong pattern of cryptic population structure (Silva et al. 2018). Silva et al. (2018) suggested that the present-day Iberian wolf population is likely organized in four, or to a finer scale, 11 genetic clusters (**Figure 1.7B**), with moderate to high levels of differentiation between them and negligible levels of gene flow (F_{ST} = 0.09-0.19, based on microsatellites). Remarkably, the authors found perfect synchrony between the genetic clusters and the movement behavior retrieved from GPS collars, showing that only two out of the 85 collared wolves disperse outside the corresponding cluster. Similarly, Gómez-Sánchez et al. (2018) analyzed patterns of genetic diversity in four Iberian wolf genomes and found large differences in heterozygosity, which was not anticipated given their close geographical relationship. The authors pointed out that the distance between these individuals was under 250 km, which should not limit gene flow for such large mobile carnivores. For example, in Eastern European wolf populations, patterns of genetic diversity revealed the contribution of populations that were 850 km away (Hindrikson et al. 2017). Furthermore, there are no obvious physical barriers in the landscape of Iberia that could intervene in the dispersal of individuals.

Landscapes in the Iberian Peninsula have been profoundly modified by agriculture and livestock for several thousand years (Zeder 2008; Revelles et al. 2018). Despite high levels of human pressure, Iberian wolves excel in co-existing and tolerating our presence and activities. For example, wolves persist in many areas feeding almost exclusively on livestock and can exploit other human-related food resources, including carrion and garbage (Llaneza and López-Bao 2015; Newsome et al. 2016). Wolves also tend to avoid potential sources of human disturbance, like roads, trails, and windfarms, and are largely restricted to mountainous areas with accessible refuge (Llaneza et al. 2012; Rio-Maior et al. 2019), where they can rest and construct dens (Sazatornil et al. 2016). Interestingly, Iberian wolves were found to have significantly lower levels of chronic stress (reflected by the hair cortisol concentration) than wolves from the Dinaric-Balkan and Scandinavian populations (Pereira et al. 2022), which may also be a contributing factor to their increased tolerance.

Up until today, no study has tried to understand whether an evolutionary mechanism underlies the persistence of wolves in human-modified landscapes. Why did the Iberian wolf population manage to persist and adapt to anthropogenic change while most of the European populations went extinct when facing similar pressures? The co-existence with

humans for millennia has resulted in a unique set of behavioral traits (e.g., changes in dispersal behavior, trophic ecology, tolerance towards humans and their activities) that possibly have allowed the persistence of Iberian wolves in highly human-dominated landscapes. Furthermore, this unique set of behavioral phenotypes is not known among any other European wolf population. Combined, these characteristics make the Iberian wolf a particularly good model to use for understanding the evolutionary mechanisms that could have allowed wolves to persist in human-dominated landscapes.

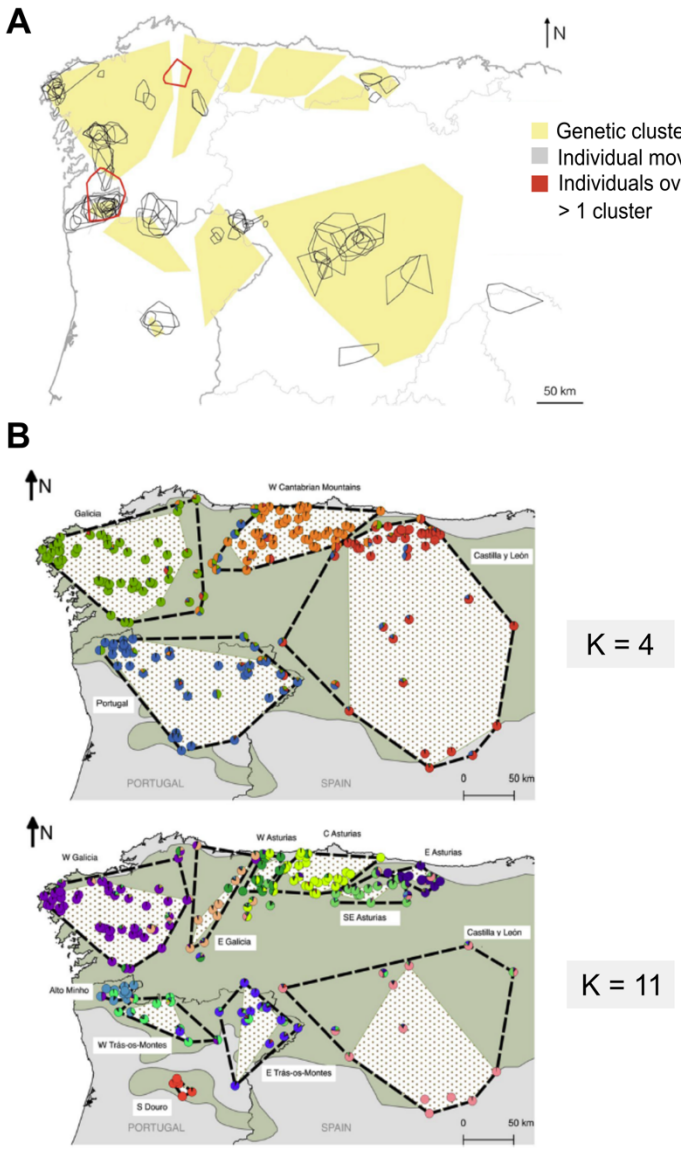


Figure 1.7. Genetic structure and low dispersal in Iberian wolves. **(A)** Spatial overlap between the movement behavior of 85 collared wolves (between 1982 and 2015) is represented by the hollow polygons, and the genetic clusters identified at K = 11 (in yellow). Red polygons represent the movement of the two wolves overlapping more than one genetic group. **(B)** Ancestry proportion of wolves to each of the genetic clusters (K = 4 and K = 11). Each circle represents a single individual and the polygons a cluster. The white dotted area identifies regions within the cluster with only non-admixed individuals. The shaded green area represents the wolf distribution in the Northwestern region of the Iberian Peninsula. Figures adapted from Silva et al. 2018.

1.5. Objectives and dissertation outline

This dissertation was designed to understand the evolutionary mechanisms underlying the behavioral traits of wolves that possibly have allowed their adaptation to human-dominated landscapes, using the Iberian wolf as a case study. In particular, the proposed objectives were:

1. Investigate the role of dog introgression in the adaptation of Iberian wolves to human-dominated landscapes. The underlying hypothesis was that "domestic" variants introgressed into the Iberian wolf genome and enhanced their adaptive potential to co-exist with humans.

2. Understand if the behavioral responses observed in Iberian wolves could conform to a process of self-domestication. The increased human pressure would have triggered natural selection towards behavioral traits favoring persistence, resembling those selected during dog domestication.

3. Characterize the genomic signatures left by the 20th-century bottleneck in the Iberian wolf genome. The idea was to understand whether population declines associated with anthropogenic disturbances depleted the Iberian wolf genome of genetic diversity and promoted an increase in the events of wolf-dog hybridization.

To this end, an integrative approach combining population genomics and transcriptomics of the Iberian wolf, several other wolf populations, and dogs, was implemented. Specifically, we generated genome-wide SNP data across a large collection of contemporary and historical Iberian wolf samples (used in objectives 1. and 3.), WGS data (objective 1.), and brain transcriptomes (objective 2.) of Iberian wolves.

This dissertation is divided into seven chapters. **Chapter I** provides a general introduction and revises the state-of-the-art literature that is most relevant in the context of this dissertation. The following five chapters present original research in the form of scientific manuscripts (two published, three in preparation to be submitted).

Chapter II comprises a methodological study where we compared the performance of different historical tissues commonly available in natural history collections of

mammalian carnivores, for genomic studies. Specifically, we collected hundreds of samples from Iberian wolf specimens across the largest museum collections on the Iberian Peninsula, as well as from private collections, and we quantify levels of endogenous DNA within each sample. We also developed a capture array to genotype SNPs contained in the commercial Illumina Canine HD BeadChip microarray system, which is commonly used to genotype SNPs in dog and wolf samples with high-quality DNA. Following this approach, we ensured compatibility between historical and contemporary datasets generated using different techniques. The historical dataset was subsequently used in Chapters III and V for population genomics. This work is published in *Frontiers in Ecology and Evolution* (doi**:** 10.3389/fevo.2022.970249; Appendix I).

In **Chapter III**, we approached objective 1. by analyzing SNP data from a large collection of contemporary and historical Iberian wolf samples, and full genome data. We started by estimating levels of dog ancestry across the Iberian wolf genome and searched for signatures of adaptive introgression. A putative introgressed region in chromosome 2 of Iberian wolves was identified in the contemporary and historical populations and further confirmed by the increased resolution of WGS data. The introgressed genomic region contained a possible candidate gene, MAST4, for the behavioral phenotype of Iberian wolves, given its association with neurological processes. The adaptive potential of the introgressed dog variants was further tested using demographic simulations. This work is currently in preparation to be submitted to a peer-reviewed journal.

In **Chapter IV,** we explored objective 2. by sequencing brain transcriptomes from the frontal cortex of four Iberian wolves living in their natural habitat, which represent the first wild wolf brain transcriptomes to be generated. We compared patterns of gene expression in Iberian wolves with those from dogs and non-Iberian wolves using transcriptomic data available in public repositories and identified potential candidate genes in Iberian wolves. Inclusively, we investigated expression profiles in a set of genes involved in neurotransmission and cognition that were previously associated with the domestic behavioral phenotype, as well as in genes associated with the introgressed dog region found in Chapter III. We also determined if differentially expressed genes in Iberian wolves overlapped with strong shifts in allele frequencies. This work is currently in preparation to be submitted to a peer-reviewed journal.

In **Chapter V,** we addressed objective 3. by performing a comparative genomic analysis between historical and contemporary populations of the Iberian wolf, using the SNP datasets formerly generated. Historical samples covered a period spanning years before,

during, and following the massive bottleneck of the 1970s, allowing us to understand how anthropogenic disturbance affected wolf population dynamics. We started by identifying the number of wolf-dog hybrids and their geographical distribution across both populations. Levels of genetic diversity were also characterized, in particular, we explored how genetic variability was distributed throughout time and space. A recent and rapid loss of genetic diversity was detected, compatible with the period of intense human persecution. Given the overall loss of variability, we investigated if there was differential retention of genetic diversity across the genome of Iberian wolves. This work is currently in preparation to be submitted to a peer-reviewed journal.

Chapter VI presents a new software tool, TVscript, that we developed to aid in the analysis of RNA-seq data. Specifically, TVscript was designed to quantify per transcript read count intra-condition variation, and to explore the effects of removing reference-based transcripts associated with high levels of such variation on subsequent differential expression analysis. We tested and revealed the usefulness of our tool in a simulated dataset and in a real case study, involving publicly available RNA-seq data from the brain of dogs, wolves, tame and aggressive foxes. For the latter, we included inter- and intra-study data to perform differential expression analysis between wolf vs. dogs and aggressive vs. tame foxes, respectively. Additionally, we used the results of these comparisons to identify genes commonly expressed in dogs and tame foxes, that could be associated with tame behavior. This work is published in *PLOS ONE* (doi**:** 10.1371/journal.pone.0274591; Appendix II).

The last chapter, **Chapter VII**, provides a general discussion of the main findings obtained in this dissertation, and how they can contribute to advancing our knowledge in the field of evolutionary biology.

1.6. References

Abascal F, Corvelo A, Cruz F, Villanueva-Cañas JL, Vlasova A, Marcet-Houben M, Martínez-Cruz B, Cheng JY, Prieto P, Quesada V, et al. 2016. Extreme genomic erosion after recurrent demographic bottlenecks in the highly endangered Iberian lynx. Genome Biol.17:251.

Abbott R, Albach D, Ansell S, Arntzen JW, Baird SJE, Bierne N, Boughman J, Brelsford A, Buerkle CA, Buggs R, et al. 2013. Hybridization and speciation. J. Evol. Biol. 26:229–246.

Aguillon SM, Dodge TO, Preising GA, Schumer M. 2022. Introgression. Curr. Biol. 32:R865–R868.

Ai H, Fang X, Yang B, Huang Z, Chen H, Mao L, Zhang F, Zhang L, Cui L, He W, et al. 2015. Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. Nat. Genet. 47:217–225.

Albert FW, Somel M, Carneiro M, Aximu-Petri A, Halbwax M, Thalmann O, Blanco-Aguiar JA, Plyusnina IZ, Trut L, Villafuerte R, et al. 2012. A Comparison of Brain Gene Expression Levels in Domesticated and Wild Animals. PLoS Genet. 8:e1002962.

Alberto FJ, Boyer F, Orozco-terWengel P, Streeter I, Servin B, de Villemereuil P, Benjelloun B, Librado P, Biscarini F, Colli L, et al. 2018. Convergent genomic signatures of domestication in sheep and goats. Nat. Commun. 9:813.

Allendorf FW, Leary RF, Spruell P, Wenburg JK. 2001. The problems with hybrids: setting conservation guidelines. Trends Ecol. Evol. 16:613–622.

Álvares F, Barroso I, Blanco JC, Correia J, Cortés Y, Costa G, Llaneza L, Moreira L, Nascimento J, Palacios V, et al. 2005. Wolf status and conservation in the Iberian Peninsula. In: Frontiers of Wolf Recovery: Southwestern US and the World. EUA: Colorado Springs. p. 76–77.

Álvares F, Bogdanowicz W, Campbell LAD, Godinho R, Hatlauf J, Jhala YV, Kitchener AC, Koepfli K-P, Krofel M, Moehlman P, et al. 2019. Old World Canis spp. with taxonomic ambiguity: Workshop conclusions and recommendations. CIBIO, Vairão, Portugal, 28th-30th May 2019. IUCN/SSC Canid Specialist Group.

Anderson TM, VonHoldt BM, Candille SI, Musiani M, Greco C, Stahler DR, Smith DW, Padhukasahasram B, Randi E, Leonard JA, et al. 2009. Molecular and Evolutionary History of Melanism in North American Gray Wolves. Science 323:1339–1343.

Andersson L, Purugganan M. 2022. Molecular genetic variation of animals and plants under domestication. Proc. Natl. Acad. Sci. 119.

Arnold ML, Kunte K. 2017. Adaptive Genetic Exchange: A Tangled History of Admixture and Evolutionary Innovation. Trends Ecol. Evol. 32:601–611.

Asa CS, Valdespino C. 1998. Canid Reproductive Biology: an Integration of Proximate Mechanisms and Ultimate Causes. Am. Zool. 38:251–259.

Axelsson E, Ratnakumar A, Arendt M-L, Maqbool K, Webster MT, Perloski M, Liberg O, Arnemo JM, Hedhammar Å, Lindblad-Toh K. 2013. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature 495:360–364.

Babenko VN, Bragin AO, Spitsina AM, Chadaeva IV, Galieva ER, Orlova GV, Medvedeva IV, Orlov YL. 2016. Analysis of differential gene expression by RNA-seq data in brain areas of

laboratory animals. J. Integr. Bioinform.13:1–15.

Bamshad M, Wooding SP. 2003. Signatures of natural selection in the human genome. Nat. Rev. Genet. 4:99–110.

Bannerman DM, Sprengel R, Sanderson DJ, McHugh SB, Rawlins JNP, Monyer H, Seeburg PH. 2014. Hippocampal synaptic plasticity, spatial memory and anxiety. Nat. Rev. Neurosci. 15:181–192.

Barton NH. 2008. The role of hybridization in evolution. Mol. Ecol. 10:551–568.

Beckman AK, Richey BMS, Rosenthal GG. 2022. Behavioral responses of wild animals to anthropogenic change: insights from domestication. Behav. Ecol. Sociobiol. 76:105.

Bélteky J, Agnvall B, Johnsson M, Wright D, Jensen P. 2016. Domestication and tameness: Brain gene expression in red junglefowl selected for less fear of humans suggests effects on reproduction and immunology. R. Soc. Open Sci. 3:160033.

Belyaev DK. 1979. Destabilizing selection as a factor in domestication. J. Hered. 70:301–308.

Benazzo A, Trucchi E, Cahill JA, Maisano Delser P, Mona S, Fumagalli M, Bunnefeld L, Cornetti L, Ghirotto S, Girardi M, et al. 2017. Survival and divergence in a small group: The extraordinary genomic history of the endangered Apennine brown bear stragglers. Proc. Natl. Acad. Sci. 114:E9589–E9597.

Bergström A, Stanton DWG, Taron UH, Frantz L, Sinding M-HS, Ersmark E, Pfrengle S, Cassatt-Johnstone M, Lebrasseur O, Girdland-Flink L, et al. 2022. Grey wolf genomic history reveals a dual ancestry of dogs. Nature 607:313-320.

Bertolini F, Servin B, Talenti A, Rochat E, Kim ES, Oget C, Palhière I, Crisà A, Catillo G, Steri R, et al. 2018. Signatures of selection and environmental adaptation across the goat genome post-domestication. Genet. Sel. Evol. 50:57.

Blanco JC, Cortés Y. 2007. Dispersal patterns, social structure and mortality of wolves living in agricultural habitats in Spain. J. Zool. 273:114–124.

Blanco JC, Cortés Y. 2009. Ecological and social constraints of wolf recovery in Spain. In: Musiani M, Boitani L, Paquet P, editors. A New Era for Wolves and People: Wolf Recovery, Human Attitudes, and Policy. Calgary: University of Calgary Press. p. 41–66.

Blanco JC, de la Cuesta L, Reig S. 1990. El lobo en España: una visión global. In: Blanco JC, de la Cuesta L, Reig S, editors. El lobo (Canis lupus) en España. Situación, problemática y apuntes sobre su ecología. Madrid: ICONA. p. 69–94.

Blanco JC, Reig S, de la Cuesta L. 1992. Distribution, status and conservation problems of the wolf Canis lupus in Spain. Biol. Conserv. 60:73–80.

Boitani L. 1992. Wolf research and conservation in Italy. Biol. Conserv. 61:125–132.

Boitani L. 1993. Wolf management actions required for conservation. In: Promberger C, Schroeder W, editors. Wolf management actions required for conservation. Munich: Wildbiologische Gesellschaft. p. 114–118.

Boitani L. 2003. Wolf conservatio and recovery. In: Mech LD, Boitani L, editors. Wolves: Behaviour, Ecology and Conservation. Chicago: University of Chicago Press. p. 317–340.

Brennan AC, Muñoz-fuentes V, Seehausen O. 2014. Hybridization due to changing species distributions: adding problems or solutions to conservation of biodiversity during global

change? Evol. Ecol. Res. 16:475–491.

Broughton JM, Weitzel EM. 2018. Population reconstructions for humans and megafauna suggest mixed causes for North American Pleistocene extinctions. Nat. Commun. 9:5441.

Campos R. 2018. Survival and Population Dynamics of the Iberian Wolf in Portugal. Dissertação de Mestrado. Universidade do Porto.

Carneiro M, Rubin CJ, Palma F Di, Albert FW, Alföldi J, Barrio AM, Pielberg G, Rafati N, Sayyab S, Turner-Maier J, et al. 2014. Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. Science 345:1074–1079.

Carter NH, Linnell JDC. 2016. Co-Adaptation Is Key to Coexisting with Large Carnivores. Trends Ecol. Evol. 31:575–578.

Carter NH, Shrestha BK, Karki JB, Pradhan NMB, Liu J. 2012. Coexistence between wildlife and humans at fine spatial scales. Proc. Natl. Acad. Sci. 109:15360–15365.

Chapron G, Kaczensky P, Linnell JDC, von Arx M, Huber D, Andrén H, López-Bao JV, Adamec M, Álvares F, Anders O, et al. 2014. Recovery of large carnivores in Europe's modern human-dominated landscapes. Science 346:1517–1519.

Charlesworth D, Willis JH. 2009. The genetics of inbreeding depression. Nat. Rev. Genet. 10:783–796.

Charlier C, Li W, Harland C, Littlejohn M, Coppieters W, Creagh F, Davis S, Druet T, Faux P, Guillaume F, et al. 2016. NGS-based reverse genetic screen for common embryonic lethal mutations compromising fertility in livestock. Genome Res. 26:1333–1341.

Chaturvedi S, Lucas LK, Buerkle CA, Fordyce JA, Forister ML, Nice CC, Gompert Z. 2020. Recent hybrids recapitulate ancient hybrid outcomes. Nat. Commun. 11:2179.

Clavero M, García-Reyes A, Fernández-Gil A, Revilla E, Fernández N. 2022. Where wolves were: setting historical baselines for wolf recovery in Spain. Anim. Conserv.

Clegg SM, Degnan SM, Moritz C, Estoup A, Kikkawa J, Owens IPF. 2002. Microevolution in island forms: the roles of drift and directional selection in morphological divergence of a passerine bird. Evolution 56:2090–2099.

Crockford SJ, Kuzmin Y V. 2012. Comments on Germonpré et al., Journal of Archaeological Science 36, 2009 "Fossil dogs and wolves from Palaeolithic sites in Belgium, the Ukraine and Russia: osteometry, ancient DNA and stable isotopes", and Germonpré, Lázkičková-Galetová, and Sablin, Jour. J. Archaeol. Sci. 39:2797–2801.

Cruz F, Vila C, Webster MT. 2008. The Legacy of Domestication: Accumulation of Deleterious Mutations in the Dog Genome. Mol. Biol. Evol. 25:2331–2336.

Darwin C. 1859. On the Origins of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life. John Murray, London, United Kingdom

Darwin C. 1868. The variation of animals and plants under domestication. John Murray, London, United Kingdom

Derks MFL, Gjuvsland AB, Bosse M, Lopes MS, van Son M, Harlizius B, Tan BF, Hamland H, Grindflek E, Groenen MAM, et al. 2019. Loss of function mutations in essential genes cause embryonic lethality in pigs. PLOS Genet. 15:e1008055.

Deschamps M, Laval G, Fagny M, Itan Y, Abel L, Casanova JL, Patin E, Quintana-Murci L. 2016. Genomic Signatures of Selective Pressures and Introgression from Archaic

Hominins at Human Innate Immunity Genes. Am. J. Hum. Genet. 98:5–21.

Díaz S, Settele J, Brondízio ES, Ngo HT, Agard J, Arneth A, Balvanera P, Brauman KA, Butchart SHM, Chan KMA, et al. 2019. Pervasive human-driven decline of life on Earth points to the need for transformative change. Science 366.

Dickson BG, Jenness J, Beier P. 2005. Influence of vegetation, topography, and roads on cougar movement in southern California.Gehrt 69:264–276.

Donfrancesco V, Ciucci P, Salvatori V, Benson D, Andersen LW, Bassi E, Blanco JC, Boitani L, Caniglia R, Canu A, et al. 2019. Unravelling the Scientific Debate on How to Address Wolf-Dog Hybridization in Europe. Front. Ecol. Evol. 7.

Dong Y, Zhang X, Xie M, Arefnezhad B, Wang Z, Wang Wenliang, Feng S, Huang G, Guan R, Shen W, et al. 2015. Reference genome of wild goat (capra aegagrus) and sequencing of goat breeds provide insight into genic basis of goat domestication. BMC Genomics 16:431.

Dufresnes C, Miquel C, Remollino N, Biollaz F, Salamin N, Taberlet P, Fumagalli L. 2018. Howling from the past: historical phylogeography and diversity losses in European grey wolves. Proc. R. Soc. B Biol. Sci. 285:20181148.

Dziech A. 2021. Identification of Wolf-Dog Hybrids in Europe – An Overview of Genetic Studies. Front. Ecol. Evol. 9.

Edelman NB, Frandsen PB, Miyagi M, Clavijo B, Davey J, Dikow RB, García-Accinelli G, Van Belleghem SM, Patterson N, Neafsey DE, et al. 2019. Genomic architecture and introgression shape a butterfly radiation. Science 366:594–599.

Eggermann J, da Costa GF, Guerra AM, Kirchner WH, Petrucci-Fonseca F. 2011. Presence of Iberian wolf (Canis lupus signatus) in relation to land cover, livestock and human influence in Portugal. Mamm. Biol. 76:217–221.

Erles K, Brownlie J. 2010. Expression of β-defensins in the canine respiratory tract and antimicrobial activity against Bordetella bronchiseptica. Vet. Immunol. Immunopathol. 135:12–19.

Eusebi PG, Cortés O, Carleos C, Dunner S, Cañon J. 2018. Detection of selection signatures for agonistic behaviour in cattle. J. Anim. Breed. Genet.135:170–177.

Faegri K. 1981. The social functions of botanical gardens in the society of the future. Bot. Jahrb. Syst. Pflanzengesch. Pflanzengeogr. 102:147–152.

Fages A, Hanghøj K, Khan N, Gaunitz C, Seguin-Orlando A, Leonardi M, McCrory Constantz C, Gamba C, Al-Rasheid KAS, Albizuri S, et al. 2019. Tracking Five Millennia of Horse Management with Extensive Ancient Genome Time Series. Cell 177:1419-1435.e31.

Fan Z, Silva P, Gronau I, Wang S, Armero AS, Schweizer RM, Ramirez O, Pollinger J, Galaverni M, Del-Vecchyo DO, et al. 2016. Worldwide patterns of genomic variation and admixture in gray wolves. Genome Res. 26:163–173.

Forrest MP, Parnell E, Penzes P. 2018. Dendritic structural plasticity and neuropsychiatric disease. Nat. Rev. Neurosci.19:215–234.

Frantz LAF, Bradley DG, Larson G, Orlando L. 2020. Animal domestication in the era of ancient genomics. Nat. Rev. Genet. 21:449–460.

Frantz LAF, Mullin VE, Pionnier-Capitan M, Lebrasseur O, Ollivier M, Perri A, Linderholm A,

Mattiangeli V, Teasdale MD, Dimopoulos EA, et al. 2016. Genomic and archaeological evidence suggest a dual origin of domestic dogs. Science 352:1228–1231.

Frantz LAF, Schraiber JG, Madsen O, Megens H-J, Cagan A, Bosse M, Paudel Y, Crooijmans RPMA, Larson G, Groenen MAM. 2015. Evidence of long-term gene flow and selection during domestication from analyses of Eurasian wild and domestic pig genomes. Nat. Genet. 47:1141–1148.

Freedman AH, Gronau I, Schweizer RM, Ortega-Del Vecchyo D, Han E, Silva PM, Galaverni M, Fan Z, Marx P, Lorente-Galdos B, et al. 2014. Genome Sequencing Highlights the Dynamic Early History of Dogs. PLoS Genet. 10:e1004016.

Freedman AH, Schweizer RM, Ortega-Del Vecchyo D, Han E, Davis BW, Gronau I, Silva PM, Galaverni M, Fan Z, Marx P, et al. 2016. Demographically-Based Evaluation of Genomic Regions under Selection in Domestic Dogs. PLoS Genet. 12:e100585.

Freedman AH, Wayne RK. 2017. Deciphering the Origin of Dogs: From Fossils to Genomes. Annu. Rev. Anim. Biosci. 5:281–307.

Frei D, De-Kayne R, Selz OM, Seehausen O, Feulner PGD. 2022. Genomic variation from an extinct species is retained in the extant radiation following speciation reversal. Nat. Ecol. Evol. 6:461–468.

Galaverni M, Caniglia R, Pagani L, Fabbri E, Boattini A, Randi E. 2017. Disentangling timing of admixture, patterns of introgression, and phenotypic indicators in a hybridizing Wolf population. Mol. Biol. Evol. 34:2324–2339.

Garzón J. 1979. La apasionada geografia del lobo. Trofeo 104:26–28.

Gaynor KM, Hojnowski CE, Carter NH, Brashares JS. 2018. The influence of human disturbance on wildlife nocturnality. Science 360:1232–1235.

Geraldes A, Ferrand N, Nachman MW. 2006. Contrasting patterns of introgression at X-linked loci across the hybrid zone between subspecies of the European rabbit (Oryctolagus cuniculus). Genetics 173:919–933.

Germonpré M, Lázničková-Galetová M, Sablin M V. 2012. Palaeolithic dog skulls at the Gravettian Předmostí site, the Czech Republic. J. Archaeol. Sci. 39:184–202.

Germonpré M, Sablin M V., Stevens RE, Hedges REM, Hofreiter M, Stiller M, Després VR. 2009. Fossil dogs and wolves from Palaeolithic sites in Belgium, the Ukraine and Russia: osteometry, ancient DNA and stable isotopes. J. Archaeol. Sci. 36:473–490.

Giska I, Farelo L, Pimenta J, Seixas FA, Ferreira MS, Marques JP, Miranda I, Letty J, Jenny H, Hackländer K, et al. 2019. Introgression drives repeated evolution of winter coat color polymorphism in hares. Proc. Natl. Acad. Sci. 116:24150–24156.

Godinho R, Llaneza L, Blanco JC, Lopes S, Álvares F, García EJ, Palacios V, Cortés Y, Javier T, Ferrand N. 2011. Genetic evidence for multiple events of hybridization between wolves and domestic dogs in the Iberian Peninsula. Mol. Ecol. 20:5154–5166.

Godinho R, Rocha RG, Layna JF. 2018. Análisis genético de la población de lobo (Canis lupus) de Guadalajara. Technical report, CIBIO, p. 26.

Gómez-Sánchez D, Olalde I, Sastre N, Enseñat C, Carrasco R, Marques-Bonet T, Lalueza-Fox C, Leonard JA, Vilà C, Ramírez O. 2018. On the path to extinction: Inbreeding and admixture in a declining grey wolf population. Mol. Ecol. 27:3599–3612.

Good JM, Giger T, Dean MD, Nachman MW. 2010. Widespread over-expression of the X

chromosome in sterile F-1 hybrid mice. PLoS Genet. 6:e1001148.

Gopalakrishnan S, Sinding MHS, Ramos-Madrigal J, Niemann J, Samaniego Castruita JA, Vieira FG, Carøe C, Montero M de M, Kuderna L, Serres A, et al. 2018. Interspecific Gene Flow Shaped the Evolution of the Genus Canis. Curr. Biol. 28:3441-3449.e5.

Grabenstein KC, Taylor SA. 2018. Breaking Barriers: Causes, Consequences, and Experimental Utility of Human-Mediated Hybridization. Trends Ecol. Evol. 33:198–212.

Grant PR, Grant BR. 2019. Hybridization increases population variation during adaptive radiation. Proc. Natl. Acad. Sci. 116:23216–23224.

Grueber CE, Wallis GP, Jamieson IG. 2013. Genetic drift outweighs natural selection at toll-like receptor (TLR) immunity loci in a re-introduced population of a threatened species. Mol. Ecol. 22:4470–4482.

Hare B, Wobber V, Wrangham R. 2012. The self-domestication hypothesis: evolution of bonobo psychology is due to selection against aggression. Anim. Behav. 83:573–585.

Harpak A, Garud N, Rosenberg NA, Petrov DA, Combs M, Pennings PS, Munshi-South J. 2021. Genetic Adaptation in New York City Rats. Genome Biol. Evol. 13:evaa247.

Hekman J, Johnson J, Edwards W, Vladimirova A, Gulevich R, Ford A, Kharlamova A, Herbeck Y, Acland G, Raetzman L, et al. 2018. Anterior Pituitary Transcriptome Suggests Differences in ACTH Release in Tame and Aggressive Foxes. G3; Genes|Genomes|Genetics 8:859–873.

Herbeck YE, Gulevich RG, Eliava M, Shepeleva DV, Trut LN, Grinevich V. 2018. Domestication: Neuroendocrine Mechanisms of Canidae -human Bonds. In: Model Animals in Neuroendocrinology. Chichester, UK: John Wiley & Sons Ltd. p. 313–334.

Hindrikson M, Männil P, Ozolins J, Krzywinski A, Saarma U. 2012. Bucking the Trend in Wolf-Dog Hybridization: First Evidence from Europe of Hybridization between Female Dogs and Male Wolves. PLoS One 7:e46465.

Hindrikson M, Remm J, Pilot M, Godinho R, Stronen AV, Baltrūnaité L, Czarnomska SD, Leonard JA, Randi E, Nowak C, et al. 2017. Wolf population genetics in Europe: a systematic review, meta-analysis and suggestions for conservation and management. Biol. Rev. 92:1601–1629.

Hou Y, Qi F, Bai X, Ren T, Shen X, Chu Q, Zhang X, Lu X. 2020. Genome-wide analysis reveals molecular convergence underlying domestication in 7 bird and mammals. BMC Genomics 21:1–20.

Howard-McCombe J, Ward D, Kitchener AC, Lawson D, Senn HV, Beaumont M. 2021. On the use of genome-wide data to model and date the time of anthropogenic hybridisation: An example from the Scottish wildcat. Mol. Ecol. 30:3688–3702.

Hu XJ, Yang J, Xie XL, Lv FH, Cao YH, Li WR, Liu MJ, Wang YT, Li JQ, Liu YG, et al. 2019. The genome landscape of tibetan sheep reveals adaptive introgression from Argali and the history of early human settlements on the Qinghai-Tibetan Plateau. Mol. Biol. Evol. 36:283–303.

Hughes PD, Gibbard PL. 2015. A stratigraphical basis for the Last Glacial Maximum (LGM). Quat. Int. 383:174–185.

Hunter P. 2007. The human impact on biological diversity. EMBO Rep. 8:316–318.

IUCN. 2011. The IUCN Red List of Threatened Species. 2011.1.

INE. 2016. Península Ibérica em números 2016. Instituto Nacional de Estatistica, Portugal. ISBN: 978-989-25-0432-2.

Jay P, Whibley A, Frézal L, Rodríguez de Cara MÁ, Nowell RW, Mallet J,

Dasmahapatra KK, Joron M. 2018. Supergene Evolution Triggered by the Introgression of a Chromosomal Inversion. Curr. Biol. 28:1839-1845.e3.

Johnsson M, Henriksen R, Wright D. 2021. The neural crest cell hypothesis: no unified explanation for domestication. Genetics 219:iyab097.

Jones MR, Mills LS, Alves PC, Callahan CM, Alves JM, Lafferty DJR, Jiggins FM, Jensen JD, Melo-Ferreira J, Good JM. 2018. Adaptive introgression underlies polymorphic seasonal camouflage in snowshoe hares. Science 360:1355–1358.

Katajamaa R, Jensen P. 2020. Selection for reduced fear in red junglefowl changes brain composition and affects fear memory. R. Soc. Open Sci. 7:200628.

Keller LF, Waller DM. 2002. Inbreeding effects in wild populations. Trends Ecol. Evol. 17:230–241.

Kijas JW, Lenstra JA, Hayes B, Boitard S, Porto Neto LR, San Cristobal M, Servin B, McCulloch R, Whan V, Gietzen K, et al. 2012. Genome-Wide Analysis of the World's Sheep Breeds Reveals High Levels of Historic Mixture and Strong Recent Selection. PLoS Biol. 10:e1001258.

Kleinman-Ruiz D, Lucena-Perez M, Villanueva B, Fernández J, Saveljev AP, Ratkiewicz M, Schmidt K, Galtier N, García-Dorado A, Godoy JA. 2022. Purging of deleterious burden in the endangered Iberian lynx. Proc. Natl. Acad. Sci. 119:e2110614119.

Ko S, Zhao M, Toyoda H, Qiu C, Zhuo M. 2005. Altered behavioral responses to noxious stimuli and fear in glutamate receptor 5 (GluR5)- or GluR6-deficient mice. J. Neurosci. 25:977–984.

Kramer AM, Dennis B, Liebhold AM, Drake JM. 2009. The evidence for Allee effects. Popul. Ecol. 51:341–354.

Kukekova A, Johnson J, Xiang X-Y, Feng S-H, Liu S, Rando H, Kharlamova A, Herbeck Y, Natalya S, Xiong Z, et al. 2018. The red fox genome assembly identifies genomic regions associated with tame and aggressive behaviors. Nat. Ecol. Evol. 2:1479–1491.

Kusak J, Fabbri E, Galov A, Gomerčić T, Arbanasić H, Caniglia R, Galaverni M, Reljić S, Huber D, Randi E. 2018. Wolf-dog hybridization in Croatia. Vet. Arh. 88:375–395.

Kwiatkowski DP. 2005. How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria. Am. J. Hum. Genet. 77:171–192.

Lande R. 1993. Risks of Population Extinction from Demographic and Environmental Stochasticity and Random Catastrophes. Am. Nat. 142:911–927.

Lande R, Shannon S. 1996. The Role of Genetic Variation in Adaptation and Population Persistence in a Changing Environment. Evolution 50:434.

Larson G, Karlsson EK, Perri A, Webster MT, Ho SYW, Peters J, Stahl PW, Piper PJ, Lingaas F, Fredholm M, et al. 2012. Rethinking dog domestication by integrating genetics, archeology, and biogeography. Proc. Natl. Acad. Sci.109:8878–8883.

Li M, Tian S, Yeung CKL, Meng X, Tang Q, Niu L, Wang X, Jin L, Ma J, Long K, et al. 2015. Whole-genome sequencing of Berkshire (European native pig) provides insights into its origin and domestication. Sci. Rep. 4:4678.

Li Y, Wang GD, Wang MS, Irwin DM, Wu DD, Zhang YP. 2014. Domestication of the Dog from the Wolf Was Promoted by Enhanced Excitatory Synaptic Plasticity: A Hypothesis. Genome Biol. Evol. 6:3115–3121.

Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ, Zody MC, et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438:803–819.

Liu KJ, Steinberg E, Yozzo A, Song Y, Kohn MH, Nakhleh L. 2015. Interspecific introgressive origin of genomic diversity in the house mouse. Proc. Natl. Acad. Sci. 112:196–201.

Liu S, Westbury MV., Dussex N, Mitchell KJ, Sinding M-HS, Heintzman PD, Duchêne DA, Kapp JD, von Seth J, Heiniger H, et al. 2021. Ancient and modern genomes unravel the evolutionary history of the rhinoceros family. Cell 184:4874-4885.e16.

Llaneza L, López-Bao JV. 2015. Indirect effects of changes in environmental and agricultural policies on the diet of wolves. Eur. J. Wildl. Res. 61:895–902.

Llaneza L, López-Bao JV., Sazatornil V. 2012. Insights into wolf presence in human-dominated landscapes: the relative role of food availability, humans and landscape attributes. Divers. Distrib. 18:459–469.

Loog L, Thalmann O, Sinding MS, Schuenemann VJ, Perri A, Germonpré M, Bocherens H, Witt KE, Samaniego Castruita JA, Velasco MS, et al. 2020. Ancient DNA suggests modern wolves trace their origin to a Late Pleistocene expansion from Beringia. Mol. Ecol. 29:1596–1610.

López-Bao JV, Fleurke F, Chapron G, Trouwborst A. 2018. Legal obligations regarding populations on the verge of extinction in Europe: Conservation, Restoration, Recolonization, Reintroduction. Biol. Conserv. 227:319–325.

Lord KA, Larson G, Coppinger RP, Karlsson EK. 2020. The History of Farm Foxes Undermines the Animal Domestication Syndrome. Trends Ecol. Evol. 35:125–136.

MacHugh DE, Larson G, Orlando L. 2017. Taming the Past: Ancient DNA and the Study of Animal Domestication. Annu. Rev. Anim. Biosci. 5:329–351.

MAGRAMA (Spanish Ministry of Agriculture, Food and Environment). 2016. Censo 2012-2014 de Lobo Ibérico (Canis lupus, Linnaeus, 1758) en España. Secretaría de Estado de Medio Ambiente.

Mancinelli S, Falco M, Boitani L, Ciucci P. 2019. Social, behavioural and temporal components of wolf (Canis lupus) responses to anthropogenic landscape features in the central Apennines, Italy. J. Zool. 309:114–124.

de Manuel M, Barnett R, Sandoval-Velasco M, Yamaguchi N, Garrett Vieira F, Zepeda Mendoza ML, Liu Shiping, Martin MD, Sinding M-HS, Mak SST, et al. 2020. The evolutionary history of extinct and living lions. Proc. Natl. Acad. Sci. 117:10927–10934.

Marsden CD, Ortega-Del Vecchyo D, O'Brien DP, Taylor JF, Ramirez O, Vilà C, Marques-Bonet T, Schnabel RD, Wayne RK, Lohmueller KE. 2016. Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. Proc. Natl. Acad. Sci. 113:152–157.

Marshall FB, Dobney K, Denham T, Capriles JM. 2014. Evaluating the roles of directed breeding and gene flow in animal domestication. Proc. Natl. Acad. Sci. 111:6153–6158.

Mech LD, Boitani L. 2003. Wolves: behavior, ecology, and conservation. (Mech LD, Boitani L, editors). University of Chicago Press

Miao B, Wang Z, Li Y. 2017. Genomic analysis reveals hypoxia adaptation in the tibetan mastiff by introgression of the gray Wolf from the tibetan plateau. Mol. Biol. Evol. 34:734–743.

Montague MJ, Li G, Gandolfi B, Khan R, Aken BL, Searle SMJ, Minx P, Hillier LW, Koboldt DC, Davis BW, et al. 2014. Comparative analysis of the domestic cat genome reveals genetic signatures underlying feline biology and domestication. Proc. Natl. Acad. Sci. 111: 17230-17235.

Morey DF. 2014. In search of Paleolithic dogs: a quest with mixed results. J. Archaeol. Sci. 52:300–307.

Mueller JC, Partecke J, Hatchwell BJ, Gaston KJ, Evans KL. 2013. Candidate gene polymorphisms for behavioural adaptations during urbanization in blackbirds. Mol. Ecol. 22:3629–3637.

Naval-Sanchez M, Nguyen Q, McWilliam S, Porto-Neto LR, Tellam R, Vuocolo T, Reverter A, Perez-Enciso M, Brauning R, Clarke S, et al. 2018. Sheep genome functional annotation reveals proximal regulatory elements contributed to the evolution of modern breeds. Nat. Commun. 9:859.

Newsome TM, Boitani L, Chapron G, Ciucci P, Dickman CR, Dellinger JA, López-Bao J V., Peterson RO, Shores CR, Wirsing AJ, et al. 2016. Food habits of the world's grey wolves. Mamm. Rev. 46:255–269.

Nielsen R. 2005. Molecular Signatures of Natural Selection. Annu. Rev. Genet. 39:197–218.

Nores C, López-Bao JV. 2022. Historical data to inform the legal status of species in Europe: An example with wolves. Biol. Conserv. 272:109639.

O'Rourke T, Boeckx C. 2020. Glutamate receptors in domestication and modern human evolution. Neurosci. Biobehav. Rev. 108:341–357.

Orlando L, Gilbert MTP, Willerslev E. 2015. Reconstructing ancient genomes and epigenomes. Nat. Rev. Genet. 16:395–408.

Orlando L, Librado P. 2019. Origin and Evolution of Deleterious Mutations in Horses. Genes 10:649.

Pacheco C, López-Bao JV, García EJ, Lema FJ, Llaneza L, Palacios V, Godinho R. 2017. Spatial assessment of wolf-dog hybridization in a single breeding period. Sci. Rep. 7:42475.

Partecke J, Schwabl I, Gwinner E. 2006. Stress and the city: Urbanization and its effects on the stress physiology in European blackbirds. Ecology 87:1945–1952.

Pendleton AL, Shen F, Taravella AM, Emery S, Veeramah KR, Boyko AR, Kidd JM. 2018. Comparison of village dog and wolf genomes highlights the role of the neural crest in dog domestication. BMC Biol. 16.

Pereira P, Fandos Esteruelas N, Nakamura M, Rio-Maior H, Krofel M, Di Blasio A, Zoppi S, Robetto S, Llaneza L, García E, et al. 2022. Hair cortisol concentration reflects the life cycle and management of grey wolves across four European populations. Sci. Rep. 12:5697.

Perri AR, Feuerborn TR, Frantz LAF, Larson G, Malhi RS, Meltzer DJ, Witt KE. 2021. Dog domestication and the dual dispersal of people and dogs into the Americas. Proc. Natl. Acad. Sci. 118:e2010083118.

Perri AR, Mitchell KJ, Mouton A, Álvarez-Carretero S, Hulme-Beaman A, Haile J, Jamieson A, Meachen J, Lin AT, Schubert BW, et al. 2021. Dire wolves were the last of an ancient New World canid lineage. Nature 591:87–91.

Pilot M, Branicki W, Jędrzejewski W, Goszczyński J, Jędrzejewska B, Dykyy I, Shkvyrya M, Tsingarska E. 2010. Phylogeographic history of grey wolves in Europe. BMC Evol. Biol. 10:104.

Pilot M, Greco C, VonHoldt BM, Jędrzejewska B, Randi E, Jędrzejewski W, Sidorovich VE, Ostrander EA, Wayne RK. 2014. Genome-wide signatures of population bottlenecks and diversifying selection in European wolves. Heredity 112:428–442.

Pilot M, Greco C, VonHoldt BM, Randi E, Jędrzejewski W, Sidorovich VE, Konopiński MK, Ostrander EA, Wayne RK. 2018. Widespread, long-term admixture between grey wolves and domestic dogs across Eurasia and its implications for the conservation status of hybrids. Evol. Appl.11:662–680.

Pilot M, Moura AE, Okhlopkov IM, Mamaev N V., Manaseryan NH, Hayrapetyan V, Kopaliani N, Tsingarska E, Alagaili AN, Mohammed OB, et al. 2021. Human-modified canids in human-modified landscapes: The evolutionary consequences of hybridization for grey wolves and free-ranging domestic dogs. Evol. Appl. 14:2433–2456.

Pimenta V, Barroso I, Álvares F, Correia J, Fernão da Costa G, Moreira L, Nascimento J, Petrucci-Fonseca F, Roque S, Santos E. 2005. Situação populacional do lobo em Portugal: resultados do censo nacional 2002/2003. Inst. da Conserv. da Natureza/Grupo Lobo, Lisboa.

Prüfer K, de Filippo C, Grote S, Mafessoni F, Korlević P, Hajdinjak M, Vernot B, Skov L, Hsieh P, Peyrégne S, et al. 2017. A high-coverage Neandertal genome from Vindija Cave in Croatia. Science 358:655–658.

Qanbari S, Pausch H, Jansen S, Somel M, Strom TM, Fries R, Nielsen R, Simianer H. 2014. Classic Selective Sweeps Revealed by Massive Sequencing in Cattle. PLoS Genet. 10:e1004148.

Ramirez O, Altet L, Enseñat C, Vilà C, Sanchez A, Ruiz A. 2006. Genetic assessment of the Iberian wolf Canis lupus signatus captive breeding program. Conserv. Genet. 7:861–878.

Ramos-Madrigal J, Sinding M-HS, Carøe C, Mak SST, Niemann J, Samaniego Castruita JA, Fedorov S, Kandyba A, Germonpré M, Bocherens H, et al. 2021. Genomes of Pleistocene Siberian Wolves Uncover Multiple Extinct Wolf Lineages. Curr. Biol. 31:198-206.e8.

Randi E. 2008. Detecting hybridization between wild species and their domesticated relatives. Mol. Ecol.17:285–293.

Randi E. 2011. Genetics and conservation of wolves Canis lupus in Europe. Mamm. Rev. 41:99–111.

Randi E, Hulva P, Fabbri E, Galaverni M, Galov A, Kusak J, Bigi D, Bolfíková BČ, Smetanová M, Caniglia R. 2014. Multilocus detection of wolf x dog hybridization in Italy, and guidelines for marker selection. PLoS One 9:e86409.

Rede Natura, ICNF. 2000. Plano Sectorial da Rede Natura 2000 para a fauna de mamíferos - Canis lupus. p. 1-9.

Reed DH, Frankham R. 2003. Correlation between Fitness and Genetic Diversity. Conserv. Biol. 7:230–237.

Revelles J, Burjachs F, Palomo A, Piqué R, Iriarte E, Pérez-Obiol R, Terradas X. 2018. Human-environment interaction during the Mesolithic- Neolithic transition in the NE Iberian Peninsula. Vegetation history, climate change and human impact during the Early-Middle Holocene in the Eastern Pre-Pyrenees. Quat. Sci. Rev. 184:183–200.

Rico M, Torrente J. 2000. Caza y rarificación del lobo en España: investigación histórica y conclusiones biológicas. Galemys 12:163–179.

Rio-Maior H, Nakamura M, Álvares F, Beja P. 2019. Designing the landscape of coexistence: Integrating risk avoidance, habitat selection and functional connectivity to inform large carnivore conservation. Biol. Conserv. 235:178–188.

Ripple WJ, Estes JA, Beschta RL, Wilmers CC, Ritchie EG, Hebblewhite M, Berger J, Elmhagen B, Letnic M, Nelson MP, et al. 2014. Status and Ecological Effects of the World's Largest Carnivores. Science 343.

Robertson BA, Rehage JS, Sih A. 2013. Ecological novelty and the emergence of evolutionary traps. Trends Ecol. Evol. 28:552–560.

Robinson JA, Brown C, Kim BY, Lohmueller KE, Wayne RK. 2018. Purging of Strongly Deleterious Mutations Explains Long-Term Persistence and Absence of Inbreeding Depression in Island Foxes. Curr. Biol. 28:3487-3494.e4.

Robinson JA, Ortega-Del Vecchyo D, Fan Z, Kim BY, VonHoldt BM, Marsden CD, Lohmueller KE, Wayne RK. 2016. Genomic Flatlining in the Endangered Island Fox. Curr. Biol. 26:1183–1189.

Rubin C-J, Megens H-J, Barrio AM, Maqbool K, Sayyab S, Schwochow D, Wang C, Carlborg Ö, Jern P, Jørgensen CB, et al. 2012. Strong signatures of selection in the domestic pig genome. Proc. Natl. Acad. Sci. 109:19529–19536.

Rubin C-J, Zody MC, Eriksson J, Meadows JRS, Sherwood E, Webster MT, Jiang L, Ingman M, Sharpe T, Ka S, et al. 2010. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature 464:587–591.

Saetre P, Lindberg J, Ellegren H, Vila C, Jazin E, Leonard JA, Olsson K, Pettersson U, Bergstrom TF. 2004. From wild wolf to domestic dog: Gene expression changes in the brain. Mol. Brain Res. 126:198–206.

Salvatori V, Donfrancesco V, Trouwborst A, Boitani L, Linnell JDC, Alvares F, Åkesson M, Balys V, Blanco JC, Chiriac S, et al. 2020. European agreements for nature conservation need to explicitly address wolf-dog hybridisation. Biol. Conserv. 248:108525.

Der Sarkissian C, Allentoft ME, Ávila-Arcos MC, Barnett R, Campos PF, Cappellini E, Ermini L, Fernández R, da Fonseca R, Ginolhac A, et al. 2015. Ancient genomics. Philos. Trans. R. Soc. B Biol. Sci. 370:20130387.

Sastre N. 2011. Genética de la conservación: el lobo gris (Canis lupus). PhD thesis, Universidad Autónoma de Barcelona: Spain

Sato DX, Rafati N, Ring H, Younis S, Feng C, Blanco-Aguiar JA, Rubin C-J, Villafuerte R, Hallböök F, Carneiro M, et al. 2020. Brain Transcriptomics of Wild and Domestic Rabbits Suggests That Changes in Dopamine Signaling and Ciliary Function Contributed to Evolution of Tameness. Genome Biol. Evol. 12:1918–1928.

Savolainen P, Zhang Y, Luo J, Lundeberg J, Leitner T. 2002. Genetic Evidence for an East

Asian Origin of Domestic Dogs. Science 298:1610–1613.

Sazatornil V, Rodríguez A, Klaczek M, Ahmadi M, Álvares F, Arthur S, Blanco JC, Borg BL, Cluff D, Cortés Y, et al. 2016. The role of human-related risk in breeding site selection by wolves. Biol. Conserv. 201:103–110.

Schafe GE, Nader K, Blair HT, LeDoux JE. 2001. Memory consolidation of Pavlovian fear conditioning: a cellular and molecular perspective. Trends Neurosci. 24:540–546.

Schubert M, Jónsson H, Chang D, Der Sarkissian C, Ermini L, Ginolhac A, Albrechtsen A, Dupanloup I, Foucal A, Petersen B, et al. 2014. Prehistoric genomes reveal the genetic foundation and cost of horse domestication. Proc. Natl. Acad. Sci. 111.

Schweizer RM, Durvasula A, Smith J, Vohr SH, Stahler DR, Galaverni M, Thalmann O, Smith DW, Randi E, Ostrander EA, et al. 2018. Natural Selection and Origin of a Melanistic Allele in North American Gray Wolves. Mol. Biol. Evol. 35:1190–1209.

Seehausen O. 2006. Conservation: Losing Biodiversity by Reverse Speciation. Curr. Biol. 16:R334–R337.

Seehausen O, Takimoto G, Roy D, Jokela J. 2008. Speciation reversal and biodiversity dynamics with hybridization in changing environments. Mol. Ecol. 17:30–44.

Sheng M, Kim MJ. 2002. Postsynaptic Signaling and Plasticity Mechanisms. Science 298:776–780.

Sih A, Stamps J, Yang LH, McElreath R, Ramenofsky M. 2010. Behavior as a Key Component of Integrative Biology in a Human-altered World. Integr. Comp. Biol. 50:934–944.

Silva P, Galaverni M, Ortega-Del Vecchyo D, Fan Z, Caniglia R, Fabbri E, Randi E, Wayne R, Godinho R. 2020. Genomic evidence for the Old divergence of Southern European wolf populations. Proc. R. Soc. B Biol. Sci. 287:20201206.

Silva P, López-Bao JV, Llaneza L, Álvares F, Lopes S, Blanco JC, Cortés Y, García E, Palacios V, Rio-Maior H, et al. 2018. Cryptic population structure reveals low dispersal in Iberian wolves. Sci. Rep. 8:14108.

Skoglund P, Ersmark E, Palkopoulou E, Dalén L. 2015. Ancient wolf genome reveals an early divergence of domestic dog ancestors and admixture into high-latitude breeds. Curr. Biol. 25:1515–1519.

Smeds L, Aspi J, Berglund J, Kojola I, Tirronen K, Ellegren H. 2021. Whole-genome analyses provide no evidence for dog introgression in Fennoscandian wolf populations. Evol. Appl. 14:721–734.

Smith FA, Elliott Smith RE, Lyons SK, Payne JL. 2018. Body size downgrading of mammals over the late Quaternary. Science 360:310–313.

Smith O, Dunshea G, Sinding M-HS, Fedorov S, Germonpre M, Bocherens H, Gilbert MTP. 2019. Ancient RNA from Late Pleistocene permafrost and historical canids shows tissue-specific transcriptome survival. PLOS Biol. 17:e3000166.

Sullivan AP, Bird DW, Perry GH. 2017. Human behaviour as a long-term ecological driver of non-human evolution. Nat. Ecol. Evol. 1:0065.

Taylor SA, Larson EL. 2019. Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. Nat. Ecol. Evol. 3:170–177.

Thalmann O, Shapiro B, Cui P, Schuenemann VJ, Sawyer SK, Greenfield DL, Germonpré MB, Sablin M V., López-Giráldez F, Domingo-Roura X, et al. 2013. Complete mitochondrial genomes of ancient canids suggest a European origin of domestic dogs. Science 342: 871-874.

Theofanopoulou C, Gastaldon S, O'Rourke T, Samuels BD, Messner A, Martins PT, Delogu F, Alamri S, Boeckx C. 2017. Self-domestication in Homo sapiens: Insights from comparative genomics. PLoS One 12:e0185306.

Trut L. 1999. Early Canid Domestication: The Farm-Fox Experiment. Am. Sci. 87:160–169.

Trut L, Oskina I, Kharlamova A. 2009. Animal evolution during domestication: The domesticated fox as a model. BioEssays 31:349–360.

Turner LM, Harr B. 2014. Genome-wide mapping in a house mouse hybrid zone reveals hybrid sterility loci and Dobzhansky-Muller interactions. Elife 3:1–25.

Valverde J. 1971. El lobo español. Montes 159:229–241.

Vilà C, Savolainen P, Maldonado JE, Amorim IR, Rice JE, Honeycutt RL, Crandall KA, Lundeberg J, Wayne RK. 1997. Multiple and Ancient Origins of the Domestic Dog. Science 276:1687–1689.

Vilà C, Walker C, Sundqvist A-K, Flagstad Ø, Andersone Z, Casulli A, Kojola I, Valdmann H, Halverson J, Ellegren H. 2003. Combined use of maternal, paternal and bi-parental genetic markers for the identification of wolf–dog hybrids. Heredity 90:17–24.

Vilà C, Wayne RK. 1999. Hybridization between Wolves and Dogs. Conserv. Biol. 13:195–198.

vonHoldt BM, Cahill JA, Fan Z, Gronau I, Robinson J, Pollinger JP, Shapiro B, Wall J, Wayne RK. 2016. Whole-genome sequence analysis shows two endemic species of North American wolf are admixtures of the coyote and gray wolf. Sci. Adv. 2:e1602250.

vonHoldt BM, Pollinger JP, Earl DA, Knowles JC, Boyko AR, Parker H, Geffen E, Pilot M, Jedrzejewski W, Jedrzejewska B, et al. 2011. A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. Genome Res. 21:1294–1305.

Vonholdt BM, Pollinger JP, Lohmueller KE, Han E, Parker HG, Quignon P, Degenhardt JD, Boyko AR, Earl DA, Auton A, et al. 2010. Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. Nature 464: 898-902.

Wall JD. 1999. Recombination and the power of statistical tests of neutrality. Genet. Res. 74:65–79.

Wang GD, Zhai W, Yang HC, Wang L, Zhong L, Liu YH, Fan RX, Yin TT, Zhu CL, Poyarkov AD, et al. 2016. Out of southern East Asia: the natural history of domestic dogs across the world. Cell Res. 26:21–33.

Wang GD, Zhai W, Yang HC, Fan RX, Cao X, Zhong L, Wang L, Liu F, Wu H, Cheng L, et al. 2013. The genomics of selection in dogs and the parallel evolution between dogs and humans. Nat. Commun. 4:1860.

Wang L, Wan ZY, Lim HS, Yue GH. 2016. Genetic variability, local selection and demographic history: genomic evidence of evolving towards allopatric speciation in Asian seabass. Mol. Ecol. 25:3605–3621.

Wang MS, Wang S, Li Y, Jhala Y, Thakur M, Otecko NO, Si J-F, Chen H-M, Shapiro B, Nielsen R, et al. 2020. Ancient Hybridization with an Unknown Population Facilitated High-Altitude Adaptation of Canids. Mol. Biol. Evol. 37:2616–2629.

Wang X, Pipes L, Trut L, Herbeck Y, Vladimirova A, Gulevich R, Kharlamova A, Johnson J, Acland G, Kukekova A, et al. 2018. Genomic responses to selection for tame/aggressive behaviors in the silver fox (Vulpes vulpes). Proc. Natl. Acad. Sci. 115:10398–10403.

Wang Y, Pedersen MW, Alsos IG, De Sanctis B, Racimo F, Prohaska A, Coissac E, Owens HL, Merkel MKF, Fernandez-Guerra A, et al. 2021. Late Quaternary dynamics of Arctic biota from ancient environmental genomics. Nature 600:86–92.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: A revolutionary tool for transcriptomics. Nat. Rev. Genet. 10:57–63.

Whitlock MC, Ingvarsson PK, Hatfield T. 2000. Local drift load and the heterosis of interconnected populations. Heredity 84:452–457.

Wilkins AS, Wrangham RW, Tecumseh Fitch W. 2014. The “domestication syndrome” in mammals: A unified explanation based on neural crest cell behavior and genetics. Genetics 198:1771.

Xue Y, Prado-Martinez J, Sudmant PH, Narasimhan V, Ayub Q, Szpak M, Frandsen P, Chen Y, Yngvadottir B, Cooper DN, et al. 2015. Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. Science 348:242–245.

Yeh PJ, Price TD. 2004. Adaptive Phenotypic Plasticity and the Successful Colonization of a Novel Environment. Am. Nat. 164:531–542.

Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS, et al. 2010. Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. Science 329:75–78.

Zeder MA. 2008. Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. Proc. Natl. Acad. Sci. 105:11597–11604.

Chapter II.

Assessing the performance of historical skins and bones for museomics using wolf specimens as a case study

Carolina Pacheco[1,2,3,†], **Diana Lobo**[1,2,3,†], Pedro Silva[1,3], Francisco Álvares[1,3], Emilio J. García[4], Diana Castro[1,3], Jorge F. Layna[5], José Vicente López-Bao[4] & Raquel Godinho[1,2,3,6,*]

[†] These authors have contributed equally to this work and share first authorship
* Corresponding author: rgodinho@cibio.up.pt

[1]CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, *InBIO* Laboratório Associado, Campus de Vairão, Universidade do Porto, Vairão, Portugal.
[2]Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Porto, Portugal.
[3]BIOPOLIS, Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, Vairão, Portugal.
[4]Research Unit of Biodiversity (UO/CSIC/PA), University of Oviedo, Mieres, Spain.
[5]Consultores en Iniciativas Ambientales, S.L., Madrid, Spain
[6]Centre for Ecological Genomics and Wildlife Conservation, Department of Zoology, University of Johannesburg, South Africa.

Published in *Frontiers in Ecology and Evolution* (September 8th, 2022).
DOI: https://doi.org/10.3389/fevo.2022.970249
Published version available in Appendix I.

Abstract

Advances in the field of museomics have promoted a high sampling demand for natural history collections (NHCs), eventually resulting in damage to invaluable resources to understand historical biodiversity. It is thus essential to achieve a consensus about which historical tissues present the best sources of DNA. In this study, we evaluated the performance of different historical tissues from Iberian wolf NHCs in genome-wide assessments. We targeted three tissues – bone (jaw and femur), maxilloturbinal bone, and skin – that have been favored by traditional taxidermy practices for mammalian carnivores. Specifically, we performed shotgun sequencing and target capture enrichment for 100,000 single nucleotide polymorphisms (SNPs) selected from the commercial Canine HD BeadChip across 103 specimens from 1912 to 2005. The performance of the different tissues was assessed using metrics based on endogenous DNA content, uniquely high-quality mapped reads after capture, and enrichment proportions. All samples succeeded as DNA sources, regardless of their collection year or sample type. Skin samples yielded significantly higher amounts of endogenous DNA compared to both bone types, which yielded equivalent amounts. There was no evidence for a direct effect of tissue type on capture efficiency; however, the number of genotyped SNPs was strictly associated with the starting amount of endogenous DNA. Evaluation of genotyping accuracy for distinct minimum read depths across tissue types showed a consistent overall low genotyping error rate (< 7%), even at low (3x) coverage. We recommend the use of skins as reliable and minimally destructive sources of endogenous DNA for whole-genome and target enrichment approaches in mammalian carnivores. In addition, we provide a new 100,000 SNP capture array validated for historical DNA compatible to the Canine HD BeadChip for high-quality DNA. The increasing demand for NHCs as DNA sources should encourage the generation of genomic datasets comparable among studies.

Keywords: endogenous DNA, historical DNA, museomics, natural history collections, SNP genotyping errors, target enrichment.

2.1. Introduction

Natural history collections (NHCs) have been gathered since the 17th century, motivated by human curiosity about our planet's biodiversity and the breakthrough in preserving perishable material (Farrington 1915). Currently, national, regional, and private collections worldwide own irreplaceable natural resources, offering wide perspectives across distinct temporal and spatial scales that inspire research in many scientific areas (Casas-Marce et al. 2012; Tsangaras and Greenwood 2012; Lopez et al. 2020; Pearson et al. 2020). Such collections provide unique overviews of historical biodiversity, from both extinct and extant species, and are also essential resources for addressing questions about species that have sampling limitations due to financial, bureaucratic, or conservation constraints (Burrell et al. 2015). Advances in molecular genetics and sequencing technology have promoted the use of the naturally fragmented DNA of historical specimens, transforming NHCs into invaluable sources of material for investigating genetics-related questions among different fields, including phylogenetics, biogeography, and conservation (Suarez and Tsutsui 2004; Holmes et al. 2016; Bi et al. 2019).

Whereas a growing body of studies using historical DNA (hDNA) illustrate the potential of NHCs in genetic research, their regular use still poses different challenges. Sampling of genetic material is often destructive, eventually compromising the integrity of specimens and jeopardizing their future use. Therefore, because sampling techniques that minimize damages are prioritized (Pálsdóttir et al. 2019), the amount of genetic material collected is often limited (Horváth et al. 2005). Additionally, traditional taxidermy practices, which commonly use hazardous chemicals, and general carelessness in protecting specimens from environmental damage, do not favor DNA preservation. Thus, historical samples often yield limited and highly degraded DNA (Raxworthy and Smith 2021) that presents major challenges during laboratory and analytical procedures (Allentoft et al. 2012; Dabney et al. 2013). Moreover, historical DNA extracts can contain a non-negligible proportion of exogenous DNA from pre- or post-mortem sources, frequently in overwhelming ratios (Weiß et al. 2016; McDonough et al. 2018; Eisenhofer et al. 2019). These factors may explain why genetic studies using hDNA have often relied on the amplification of short nuclear or mitochondrial fragments (e.g., Schwartz et al. 2007; Maebe et al. 2016; Lonsinger et al. 2019). However, the use of hDNA is nowadays facilitated by high-throughput sequencing and by recent developments in molecular methods (e.g., Rowe et al. 2011; Staats et al. 2013; Hung et al. 2014). Methods like

sequence capture of target loci, which limits the representation of the genome to specific loci, are among the most used in genome-wide studies of low-quality DNA to achieve large and cost-effective datasets (Jones and Good 2016; McCormack et al. 2016; Derkarabetian et al. 2019). Furthermore, despite requiring *a priori* availability of the target genome to design specific baits, target enrichment has been shown to be successful using bait designs based on closely related species (Vallender 2011).

Mammals have been traditionally preserved in NHCs by archiving skins, bones, teeth, or mounted specimens (Rowe et al. 2011). It is thus not surprising that most of the available genetic studies using hDNA rely on these tissues (Raxworthy and Smith 2021). However, DNA yields may vary greatly among different tissues and be dependent on curation history (Burrell et al. 2015). Hard tissues, such as teeth and bones, were at first thought to provide higher-quality DNA (Wandeler et al. 2007; Casas-Marce et al. 2010), encouraging proposals to use hard tissues assumed to minimize sampling damage, such as maxilloturbinal bone (Wisely et al. 2004). Yet, recent studies have shown conflicting results (Rowe et al. 2011; Lonsinger et al. 2019; Tsai et al. 2020), revealing soft tissues to be good sources of hDNA when preserved appropriately (Burrell et al. 2015). To date, few studies have implemented genomic resources to assess differences in the performance of distinct tissues, and those tackling this question are often based on very low sample sizes that hamper reliable statistical comparisons (e.g., Rowe et al. 2011; McDonough et al. 2018). Thus, the tissue of choice for increasing the quality of genomic data, while sampling mammal NHCs with minimal damaging, remains an open question (Raxworthy and Smith 2021).

In this work, we sought to evaluate the performance of different mammalian carnivore tissues generally available at NHCs in genome-wide assessments. Using the Iberian wolf (*Canis lupus signatus*) as a case study, we first collected bones (jaw and femur), maxilloturbinal bones, and skins from 103 historical specimens (**Figure 2.1A**). Then, we performed a shotgun sequencing of these samples to characterize endogenous DNA content. Third, we developed and tested a capture array of 100,000 regions overlapping single nucleotide polymorphisms (SNPs) contained in the commercial Illumina Canine HD BeadChip, ensuring compatibility between datasets generated with both low- and high-quality DNA. We intended to answer three main questions: Does endogenous DNA content differ across historical tissues? How is capture efficiency affected by historical tissue type? What is the effect of read depth on SNP genotyping error rates from hDNA?

2.2. Materials and Methods

2.2.1. Sampling, DNA extraction, and shotgun sequencing

We collected 103 samples from Iberian wolf specimens housed at the three largest museum collections in the Iberian Peninsula and at 15 private NHCs, consisting of 43 bones (jaw and femur), 31 maxilloturbinal bones (hereafter nasal bones), and 29 skins. The original collection years for these specimens ranged from 1912 to 1990, except for two samples from 2004 and 2005 (**Figure 2.1**; Supplementary Table 2.1). All our samples conform to the definition of historical samples by Raxworthy and Smith (2021), which restricts hDNA to that fortuitously obtained from traditional museum specimens not intended to serve as sources of DNA. Jaw and femur bones were sampled by drilling ca. 1 g of bone powder with a Dremel tool (Dremel, WI, USA; Supplementary Figure 2.1). Femurs were sampled in the patellar surface region, whereas jaws were sampled in the posterior lower region of the mandibula (see Supplementary Figure 2.1 for examples of the drilling location in each bone type). Nasal bones were collected following Wisely et al. (2004), by inserting sterilized forceps into the nasal cavity of the skull to extract the bones. Nasal bone material was posteriorly crushed into small fragments. Bones were not bleached prior to DNA extraction. Skin samples were collected from pelts or mounted specimens by extracting a patch of approximately 2 cm^2. Collecting tools, including drill bits, were cleaned with bleach, and flamed with 96% ethanol between samples to minimize cross contamination.

We prepared DNA extracts using 50 mg of sample following Dabney and Meyer (2019). For skins, we favored the inner layer for DNA extraction and discarded hairs to minimize contamination with external DNA sources. DNA concentration was measured using the Qubit fluorometer dsDNA HS Assay Kit (Thermo Fisher Scientific, MA, USA). We used 100 - 300 ng of DNA to prepare blunt-end dual-indexed DNA sequencing libraries using a full-uracil-DNA-glycosylase treatment following Meyer and Kircher's (2010) protocol with the modifications described in Kircher et al. (2012). To limit DNA contamination, DNA extractions and library preparations were conducted in dedicated rooms under sterile conditions and positive air pressure, and negative controls were used alongside the procedures. DNA libraries were diluted based on concentration measurements obtained with the Qubit fluorometer dsDNA HS Assay Kit, and library size ranges were characterized using a Bioanalyzer 2100 with High Sensitivity DNA kits (Agilent Technologies, CA, USA). To characterize the endogenous DNA content of each sample, we performed a shotgun sequencing run using one lane of an Illumina HiSeq X

instrument (Illumina, CA, USA) in PE 150 bp mode. For this, we selected libraries with concentrations > 15 ng/μL (N = 79; N_{bone} = 29; $N_{nasal\ bone}$ = 30; N_{skin} = 20) to ensure that enough library material remained for the following steps (based on the required starting amount for target capture enrichment of 14-72 ng/μl).

We also generated genomic information from two contemporary wolf muscle samples to be used as positive controls to validate the implemented approach. No animals were killed or injured for this study. We isolated DNA from these samples using the DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany). DNA quantification and library preparation were performed following the same procedures as described above for historical samples but with two modifications during library preparation: i) a shearing step using a Bioruptor Pico sonication device (Diagenode, NJ, USA) was performed to obtain fragments of ~250 bp, and ii) the USER enzyme treatment was not performed. These two samples were genotyped using two different approaches: the Illumina CanineHD BeadChip (~170,000 SNPs) and the capture array of 100,000 regions developed in this work (see next sections).

2.2.2. Bait design

We targeted a set of autosomal genome-wide SNPs whose positions were defined based on the coordinates available on the CanineHD BeadChip. With this experimental design, we ensured the compatibility of historical datasets with other datasets generated using the same SNP chip. From the ~170,000 SNPs in the CanineHD BeadChip, only uniquely mapped autosomal SNPs were considered for probe design. Based on this list of putative SNPs, we custom designed a sequence capture panel containing a final set of 100,000 SNPs distributed across the wolf genome using MYbaits Target Capture technology (Arbor Biosciences, Ann Arbor, MI, USA). The implemented methods for probe design followed the strategy described by Haak et al. (2015) and Cruz-Dávalos et al. (2017), in which four probes of 60 bp each were designed to target fragments of 120 bp per SNP: one upstream of the SNP, one downstream, and one for each possible allele, with the allele positioned at the center of the probe. RNA probes were designed and synthesized by Arbor Biosciences (product code #302016). The selection of the final set of MYbaits RNA probes was based on their capture efficiency (stringent criteria ≤ 25% repeat masked). A FASTA file providing the sequences of all 400,000 synthesized probes is available online.

2.2.3. Capture enrichment

We performed target enrichment using the MYbaits Custom Target Capture Kit following the manufacturer protocol (v.3.02). DNA libraries with similar endogenous DNA content were pooled in equimolar sets of 8 to 10 samples per capture reaction (600 ng in the final pool). Hybridization between RNA probes and the DNA library occurred at 65ºC for 40 hours. Real-time PCRs were performed to determine the number of amplification cycles required to obtain sufficient molarity for sequencing. Post-enrichment libraries were amplified using KAPA HiFi Hotstart Ready Mix (KAPA Biosystems, MA, USA), following the manufacturer recommendations, with an annealing temperature of 60ºC. The amplified enriched DNA libraries were purified in 20 µl of EB buffer using the MinElute PCR Purification kit (Qiagen) and quantified by the Qubit dsDNA HS Assay Kit. Library size ranges were characterized using a Bioanalyzer 2100 with High Sensitivity DNA kits and pooled in equimolar ratios. Pooled libraries were then sequenced using 2 lanes of an Illumina HiSeq X instrument in PE 150 bp mode.

2.2.4. Processing of sequencing reads and single nucleotide polymorphisms calling

Raw sequences from shotgun sequencing and target enrichment were processed similarly in different time periods. First, sequence reads were demultiplexed and quality assessments were done using FastQC (Andrews 2018) and MultiQC (Ewels et al. 2016). Sequence reads were then processed and aligned using the PALEOMIX v.1.2.13.2 BAM pipeline (Schubert et al. 2014). Briefly, the pipeline consisted of the following steps: i) adapter sequences were removed, low-quality and N bases were trimmed, and overlapping read pairs were collapsed using AdapterRemoval v.2.2.2 (Schubert et al. 2016) with the default parameters (on average, 96% and 95% of reads were collapsed for shotgun and capture sequencing, respectively); ii) all sequence reads were then aligned against the CanFam3.1 dog reference genome (Lindblad-Toh et al. 2005) using BWA-MEM v.0.7.17 (Heng 2013); and iii) PCR duplicates for each library were marked by the “paleomix rmdup_collapsed” tool but were not used for the following steps. The final alignment file was subjected to local realignment around indels using the GATK v.3.8 IndelRealigner tool (DePristo et al. 2011). The endogenous content was determined by the ratio of unique reads (no duplicates) that mapped to the dog reference genome with mapping quality above 20 (MQ > 20) to the total number of available reads (also referred as library complexity by some authors; Dehasque et al. 2022). For the target enrichment experiment, the number of reads mapping on target was defined as the total number of

unique high-quality (MQ > 20) reads overlapping at least 1 base of the 120 bp target region. Enrichment success was determined by the ratio of reads on target in relation to i) the total number of available reads and ii) the total number of uniquely and high-quality mapped reads. These metrics were retrieved from the PALEOMIX summary report and, additionally, with the help of SAMtools v.1.9 (Li et al. 2009). Fold enrichment was determined by the ratio of the number of on-target reads to the total number of uniquely and high-quality mapped reads, divided by the expected representation of the target regions without enrichment (i.e., the ratio of genome length, 2.4 Gb, to target length, 100,000 x 120 bp, corresponding to 0.5%).

Following target enrichment, genotypes were called using BCFtools v.1.10.2 (Li 2011) mpileup/call -m tools, with minimum Phred-scaled thresholds of 20 for base quality and read mapping quality. At the end, to evaluate the effect of sequencing depth on genotyping quality rates, we considered genotypes supported by at least three (DP ≥ 3), four (DP ≥ 4), or more reads (DP ≥ 5) using the custom python script *gtvalues2plink.py*, which was also used to convert the final VCF file to plink format. The distribution and density of SNPs in our dataset (using DP ≥ 4) was visualized using the R/Bioconductor package *karyoploteR* (Gel and Serra 2017) in R (R Development Core Team 2017).

Genotypes were ultimately validated by estimating their concordance rates with the genotypes obtained from the control samples using the Canine HD BeadChip. In this last approach, genotype calling was performed using GenomeStudio software (Illumina), following Illumina's recommendations. Sex-chromosome-related SNPs and non-uniquely mapped SNPs (SNPs with multiple positions attributed) were removed from the dataset using PLINK v.1.9 (Purcell et al. 2007), resulting in a final dataset of ~121,000 SNPs. Concordance rates were calculated for the entire set of genome-wide SNPs obtained across the three depth thresholds for each control sample, using a second custom script, *SNP_concordance.py*. We also calculated concordance rates for SNPs called exclusively with 3x and 4x coverage in each control sample and estimated the associated error rates. Additionally, to assess the genotype quality across all historical samples, we estimated the potential genotyping error rate associated with low read depth using the *ErrorCount.sh* script from the dDocent pipeline (Puritz et al. 2014). Briefly, this script reports a low range based on a 50% binomial probability of observing the second allele in a heterozygote and a high range based on a 25% probability. All genotyped SNPs were considered for this analysis without any filter for missing data.

2.2.5. Statistical analysis

To determine whether the amount of endogenous DNA and capture efficiency were influenced by sample type (bone, nasal bone, and skin), we implemented a set of generalized linear models (GLMs). We ran four different GLMs with the following dependent variables: i) proportion of endogenous DNA, i.e., number of reads that mapped uniquely with MQ > 20 in relation to the total number of available reads, following shotgun sequencing; ii) mapped reads after capture, i.e., proportion of the number of reads that mapped uniquely and with MQ > 20 to the total number of reads available after target enrichment; iii) reads on target (all), i.e., number of unique and high-quality (MQ > 20) reads that mapped on target regions in relation to the total number of reads available; and iv) reads on target (mapped), i.e., the same as above, but in relation to the total number of mapped reads. All GLMs were fitted with a binomial error distribution and a logit link. The fit of each model was further assessed using the Pearson's χ^2 residuals, which test whether any significant patterns remain in the residuals. Given the unavailability of shotgun sequencing data for all the samples, we tested levels of correlation between the proportion of endogenous DNA and the proportion of reads mapping after capture to understand if the latter could be interpreted as a proxy for endogenous content. We also tested correlation coefficients using the following variables: original collection year, DNA yield (ng) following extraction, endogenous DNA proportion, fragment length (average length of filtered reads from shotgun sequencing), and mapping length (average length of mapped and unique reads, with MQ > 20). The effect of sample type on DNA concentration was also evaluated using a Kruskal-Wallis test. All the previously mentioned tests were performed in R, and all the plots were constructed using the R package *ggpubr* (Kassambara 2020).

2.3. Results

2.3.1. Endogenous DNA content

We successfully obtained DNA extracts for all 103 historical samples, with an average DNA yield of 469.91 ± 63.18 (s.e.) ng; (range 84 - 3,812 ng). DNA yield was not correlated with the original specimen collection year ($r^2 = 0.15$) nor with the sample type (Kruskal-Wallis test, $p = 0.256$; **Figure 2.1B**). Nevertheless, even using the same quantity of starting material for DNA extraction, we cannot rule out greater effects due to histological differences between hard and soft tissues. The initial sample

characterization by shotgun sequencing resulted in an average fragment length of 97 bp (range 42 - 145 bp). Endogenous DNA content across all samples varied from 0.05 to 76.35% (mean: 14.48 ± 2.47%; **Figure 2.1C**) and was not correlated with the fragment length (r^2 = -0.42; Supplementary Figure 2.2A) nor with the original collection year (r^2 = -0.24; **Figure 2.1C**). The average mapping length across all samples was 75 bp (range 40 - 116 bp) and was not correlated with specimen original collection year (r^2 = 0.15; Supplementary Figure 2.2B). Among sample types, skin samples retrieved the highest proportion of endogenous DNA (43.49 ± 4.76%, p = 0.008; **Figure 2.1C**; Supplementary Tables 2.2 and 2.3) in relation to bones (8.14 ± 2.66%) and nasal bones (1.28 ± 0.22%).

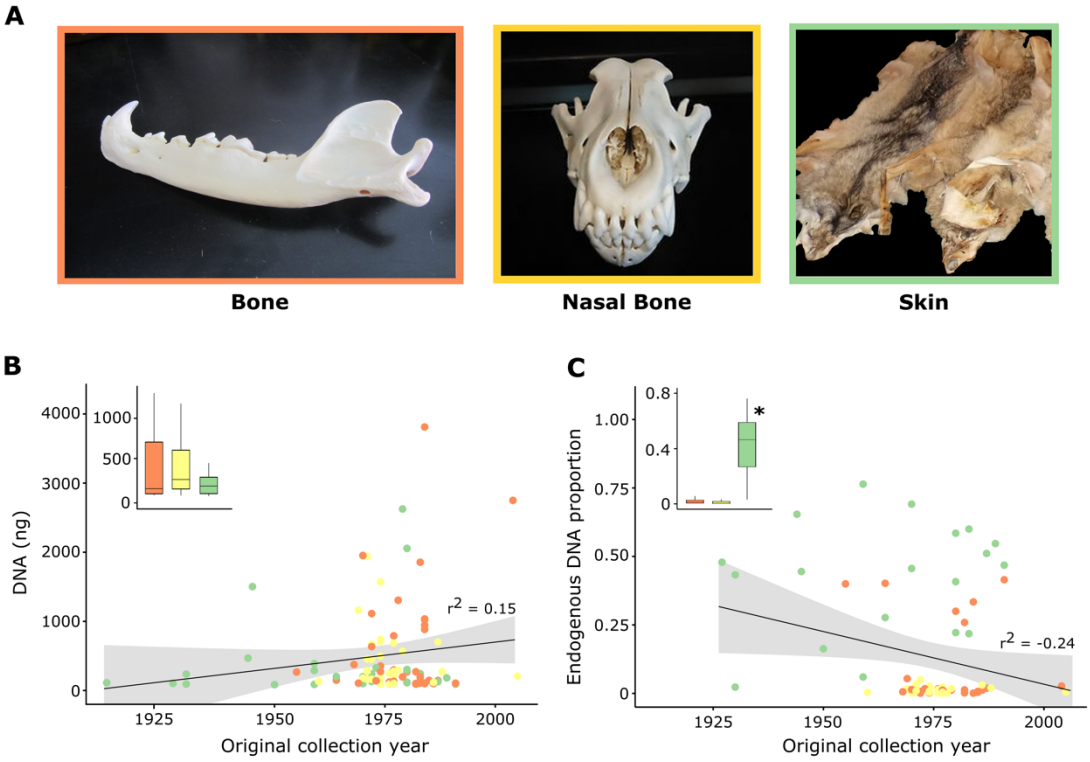


Figure 2.1. Characterization of samples and hDNA used in this work. (**A**) Photographs illustrating the specimens used for the collection of each tissue type (Image credits Raquel Godinho). (**B/C**) Correlation (r^2) between DNA yield (ng) following extraction and the proportion of endogenous DNA content with the original collection date of the specimens. Each dot represents a historical sample and inset boxplots display the median (central line) and distribution per tissue type. Significant differences (p < 0.05) between tissues are identified with an asterisk. Colors depict the three different tissue types used in this study: orange - bones (jaw and femur); yellow - maxilloturbinal (nasal) bones; and, green - skins.

After capture enrichment, the proportion of reads mapping to the reference genome ranged from 0.13 to 78.26% (mean: 19.89 ± 2.59%; Supplementary Table 2.2).

Consistently, skin samples presented a significantly higher proportion of mapped reads (49.01 ± 4.69%, p = 0.001; **Figure 2.2**; Supplementary Tables 2.2 and 2.4) relative to bones (12.57 ± 2.99%) and nasal bones (2.79 ± 0.64%). Across contemporary control samples, the average proportion of reads mapping after capture was 32.85 ± 0.56% (Supplementary Table 2.5), with 65.4% of filtered reads being duplicates. Endogenous DNA content and the proportion of reads mapping to the reference genome following capture were highly correlated (r^2 = 0.93; Supplementary Figure 2.3A; Additional Dataset 1 available online).

2.3.2. Capture efficiency across sample types

Following capture enrichment, the proportion of reads mapping on target regions ranged from 0.01 to 27.79% (mean: 4.75 ± 0.73%; Supplementary Table 2.2) in relation to all available reads and from 1 to 37% (mean: 19.82 ± 0.84%; Supplementary Table 2.2) in relation to all mapped reads. The average fold enrichment was 39.64x, with 91% of all samples presenting an enrichment > 10x. Among sample types, skin samples showed the highest proportion of reads mapping on target regardless of the metric used (12.35 ± 1.58% of all reads; 22.49 ± 1.65% of mapped reads; **Figure 2.2**). Bones (2.68 ± 0.78% of all reads; 18.13 ± 1.41% of mapped reads) and nasal bones (0.50 ± 0.11% of all reads; 19.68 ± 1.18% of mapped reads) worked less successfully; however, differences were not significant among the three sample types (**Figure 2.2**; Supplementary Tables 2.6 and 2.7). For the two contemporary control samples, the proportion of reads mapping on target regions was 17.54 ± 0.01% of all available reads and 53.42 ± 0.69% of mapped reads (Supplementary Table 2.5; Additional Dataset 1 available online).

2.3.3. Genotyping errors and coverage effect

We were able to generate genotypes for a panel of 99,982 genome-wide SNPs (Supplementary Figure 2.4) using the target enrichment approach, with distinct levels of missing data across samples. For contemporary control samples, concordance rates between genotypes obtained from the capture array developed in this study and those from the Canine HD BeadChip were above 99%. Error rates for the two control samples (i.e., rates of genotype discordance) were almost negligible but increased with decreasing coverage at the target sites (0.64 and 0.91% error rate, at coverage ≥ 5x and ≥ 3x, respectively; **Figure 2.3A**; Supplementary Table 2.8). When calculating the genotype concordance rates using SNPs called exclusively with 3x and 4x, we found a decrease to 94.1 and 96.4%, respectively (Supplementary Table 2.8). The most common

error found in genotypes called with the lowest coverage (3x) was the dropout of a second allele (miscalled homozygous in relation to the SNP chip), corresponding to 79.5% of the observed discordances. The number of loci obtained in contemporary control samples increased as the coverage decreased; for example, the average number of SNPs obtained across both samples declined from 93,948 to 88,173 at coverage ≥ 3x and ≥ 5x, respectively (**Figure 2.3B**; Supplementary Table 2.9).

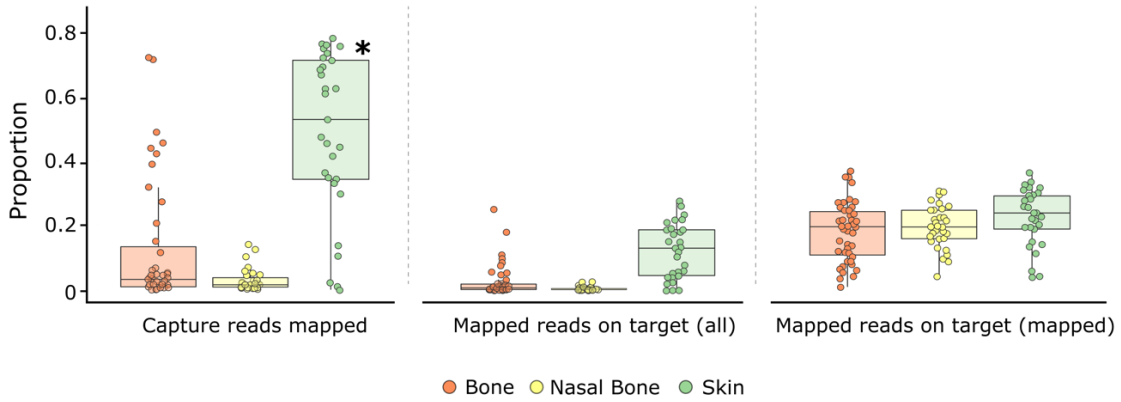


Figure 2.2. Genomic metrics assessed from capture enrichment for each tissue type: bones (orange), maxilloturbinal (nasal) bones (yellow), and skins (green). Boxplots display the proportion of reads mapping after capture enrichment and reads mapping on target region in relation to all or mapped reads for each tissue. Within boxplots, dots represent historical samples, and the central line indicates the median value. Significant differences ($p < 0.05$) between tissues for each metric are identified with an asterisk.

We found the same pattern across historical samples, with increased estimates of potential genotyping error rates associated with SNPs called with decreasing coverage (**Figure 2.3A**). Nevertheless, the highest error rate estimate for coverage ≥ 3x did not reach 7%. Concordantly, we also found an increase in the average number of SNPs across all samples for lower read depths (24,663 and 18,523 SNPs at coverage ≥ 3x and ≥ 5x; **Figure 2.3B**, Supplementary Table 2.9). Overall, the average number of SNPs obtained across historical samples was substantially lower than in contemporary control samples. Still, the highest genotyping success rates were found among three historical samples, which presented very similar rates (> 95%) to those found in the contemporary samples (**Figure 2.3B**, Supplementary Table 2.9). Skin samples presented the highest average numbers of SNPs genotyped (**Figure 2.3B**). Genotyping success rate (number of genotyped SNPs) was positively correlated with the proportion of reads mapping to the genome after capture ($r^2 = 0.85$; Supplementary Figure 2.3B).

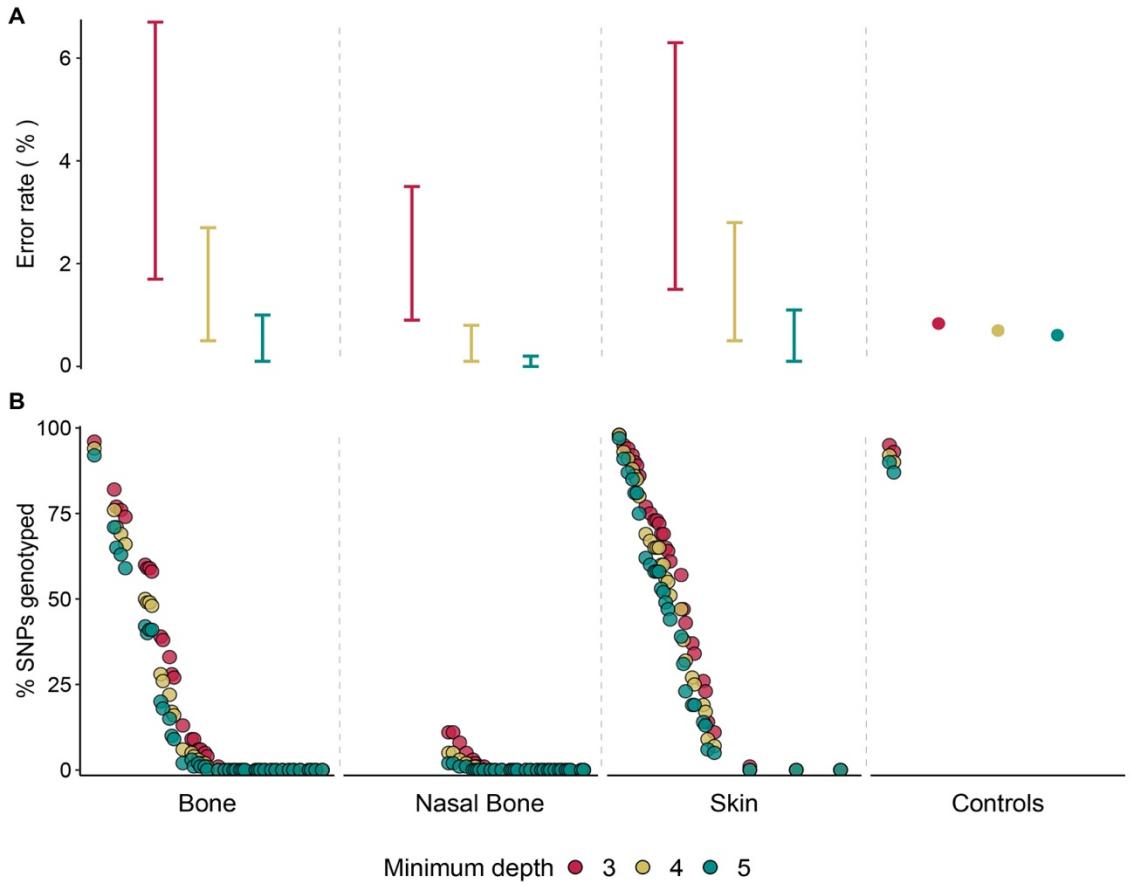


Figure 2.3. Genotyping performance of hDNA from three sample types. (**A**) Interval of potential genotyping error rate estimates (%) for the three hDNA sample types, and average empirical genotyping error rate for control samples, considering the three minimum depth levels. (**B**) Percentage of genotyped SNPs (N = 99,982) considering a minimum depth of 3x, 4x or 5x per SNP across 103 samples. Each dot represents a sample.

2.4. Discussion

In this study, we evaluated the performance of different historical tissues – bones, nasal bones, and skins – across a population sampling of Iberian wolf NHCs, in whole-genome sequencing and target enrichment. Specifically, we analyzed metrics based on endogenous DNA content, uniquely high-quality mapped reads after capture, and enrichment proportions.

Bones, nasal bones, and skins all succeeded as DNA sources but differed in endogenous DNA content. Based on equal starting amounts of DNA for shotgun sequencing, skins yielded significantly higher amounts of endogenous DNA than did the other sample types. This concords with other studies that have previously shown that

soft tissues of mammals (van der Valk et al. 2017) and birds (Tsai et al. 2020) provide higher endogenous DNA content than hard tissues, although these are based on less comprehensive sample sizes than the one used here. Although nasal bones were initially recommended for presenting higher genotyping success rates than other bones (Wisely et al., 2004), a recent study showed conflicting results (Lonsinger et al. 2019). Here, we assessed, for the first time, the proportion of endogenous DNA retrieved from nasal bones, showing that this bone type provides lower content than other commonly used bones. Despite the observed significant effect of tissue type on endogenous DNA, we cannot disregard the impact of distinct collection histories in this result (Raxworthy and Smith 2021). We did not observe significant associations between endogenous DNA content, or its mapping length, and the original collection year, suggesting no substantial DNA degradation across the period under evaluation.

There was no direct effect of tissue type on capture efficiency, with most samples presenting an enrichment > 10x. This result emphasizes the efficiency of target capture as a powerful method for building genomic datasets from hDNA and is in line with other studies using capture approaches (Carpenter et al. 2013; van der Valk et al. 2017). An *a priori* understanding of the importance of endogenous DNA content in the success of a capture experiment (Hernandez-Rodriguez et al. 2018) has driven our decision to perform shotgun sequencing prior to capture enrichment. Using this approach, we were able to demonstrate a perfect association between the number of resulting SNPs and the initial amount of endogenous DNA, emphasizing that this is a practical and cost-effective way to select samples prior to capture experiments. Furthermore, knowing the endogenous DNA content across samples also allows to minimize variation within each sequencing experiment, ensuring low rates of index hopping and reducing possible bias in downstream analysis (van der Valk et al. 2020). Our SNP genotyping results demonstrate that historical samples with high amounts of endogenous DNA can behave similarly to, or even better than, contemporary samples in capture procedures, further supporting the use of capture enrichment to genotype thousands of genome-wide SNPs in hDNA samples (Bi et al. 2013; Smith et al. 2014; Harvey et al. 2016; Lim and Braun 2016).

The final number of SNPs in a dataset can be a trade-off between coverage and genotyping error rates: relaxing the minimum read depth to increase the number of SNPs is accompanied by higher uncertainty in genotype calling. Still, we were able to generate an average of ca. 25,000 genome-wide SNPs per sample, called with ≥ 3x coverage and with low potential error rates (< 7%). Reducing coverage led to an increase in the number

of SNPs, mostly across samples with intermediate genotyping success, i.e., those where SNPs were captured but read depth was generally low. In such cases, decreasing the coverage from ≥ 4x to ≥ 3x represented an increase of ~11,000 SNPs. Regardless of the threshold used for minimum coverage, genotypes obtained for the contemporary control samples were confirmed and validated against the SNP chip genotypes. The average error rates found in the controls were within, or very close to, the potential error rate intervals estimated for the historical dataset. This overlap suggests that the use of such an analytical approach (Puritz et al. 2014) can be a reliable alternative to estimate potential error rates when no control samples are available. Observed error rates in our work for ≥ 5x coverage, which is a widely accepted threshold for SNP calling (Yi et al. 2010), were much lower than those reported by Fountain et al. (2016) at the same coverage for fresh tissue samples.

The most common error found among control samples was the dropout of a second allele after target enrichment. Given the overall low error rates, we predict minimal impact associated with allelic dropouts in downstream analysis. However, even low rates of genotyping error tend to overestimate genetic variation and can affect population genetic studies in different ways, or, to a greater extent linkage and association studies (Pompanon et al. 2005; Gautier et al. 2013). Since most studies do not have control samples available to calculate empirical concordance rates and validate genotypes (Fountain et al. 2016), predicting the effects of errors might be difficult, and in this case higher minimal depth thresholds (> 5x) may be considered. Choosing between increasing the number of loci or having high reliability in the genotypes should be considered on a case-by-case basis to ensure compatibility with downstream applications. Alternatively, historical datasets can be analyzed based on genotype likelihoods instead of genotypes, in order to take the inherent uncertainty of the genotypes into account (Nielsen et al. 2011).

2.4.1. Reducing sample damage without compromising data recovery

Skin samples were revealed to be a good source of endogenous DNA while still being minimally destructive to specimens, as small patches can be easily sampled from nonunique morphological features of hides or skin mounts. Bone sampling is generally more destructive and does not necessarily translate into higher endogenous DNA content. This applies in particular to nasal bones, which yield the lowest endogenous DNA content while relying on a quite destructive sampling process in which a large part of the structure is removed. Although we acknowledge that the consumptive sampling of

nasal bones does not compromise the utility of a specimen for morphometric or character studies (Wisely et al. 2004), our results discourage their sampling for genomic studies of mammalian carnivores.

As the demand for NHCs in molecular studies increases, conscious sampling of specimens should now become routine to not compromise their future use. The drilling procedure we used for bone sampling left a small hole in the bone tissue (**Figure 2.1A** and Supplementary Figure 2.1) but did not hamper future morphometric studies. This sampling approach is less damaging than cutting bone fragments (Supplementary Figure 2.1) and still ensures that enough material is collected for genomic analysis without further manipulation. Sampling the petrous bone, a recognized excellent source of endogenous DNA (Pinhasi et al. 2015), would only be recommendable for NHC specimens if entailing the use of damaged skulls, as its sampling is highly destructive to the skull (Charlton et al. 2019). The starting amount of endogenous DNA can also be maximized through wet lab procedures by performing several rounds of DNA extraction, creating multiple and differentially indexed DNA libraries to increase complexity levels, and/or captures per sample (Hernandez-Rodriguez et al. 2018; White et al. 2019; Fontsere et al. 2021; von Seth et al. 2021). Understanding that a reduced fragment of the appropriate tissue is enough to successfully recover molecular data without compromising the reusability of the specimen is of remarkable importance for managing NHCs.

2.4.2. Limitations and opportunities for further development

In contrast to current practices for preserving biological material in controlled environments and using sophisticated resources that prevent DNA damage, older traditional methods did not prioritize DNA integrity (Hall et al. 1997; Burrell et al. 2015; Card et al. 2021). Thus, these preservation techniques and storage conditions can greatly impact the quality and quantity of DNA. In this work, we used wolf specimens from three museums and 15 private collections distributed across a 90-year period, where most samples (~60%) were collected between 1960 and 1980. These samples likely have different collection histories, particularly those in private collections, where less-standardized preservation methods can be expected. Such heterogeneity may explain the observed levels of variability in endogenous DNA content, although our sample size hampers a statistical evaluation of these metrics across sampling origins. We acknowledge that it would be relevant to assess the effects of different preservation

techniques on DNA degradation; unfortunately, information about the method applied to each specimen used for this study was not available, as is often the case in NHCs.

One of the major challenges the museum community currently faces is improving the documentation of collection histories for individual specimens and to make it accessible to the scientific community. Global initiatives have been promoting the digitization of specimens' metadata and digital images to make them available in electronic databases that can be easily accessed (e.g., The Global Information Facility; Robertson et al. 2014). As this process evolves, it is essential that researchers working in the field of museomics synergize genomic data with NHCs metadata to enhance the scientific impact and traceability of their studies by, for example, always providing the catalogue numbers of specimens used (Card et al. 2021).

2.5. Concluding remarks

To our knowledge, this study uses the most comprehensive dataset to date – in terms of sample size and genome representation – to test the performance of three wolf tissues as sources of hDNA. Based on our findings, we recommend the use of skins for sampling mammalian carnivore specimens, as these are reliable and minimally destructive sources of endogenous DNA suitable for whole-genome and target enrichment approaches. This study should also encourage future research with the same aims but targeting different vertebrate and invertebrate groups. In addition, we provide a validated genome-wide SNP tool (i.e., probe design) that allows for direct comparison between historical and contemporary data. Although the enrichment approach presented here was based on canid genomes, its conceptual design can be implemented in any species for which SNP chips are available. We believe that the increasing demand for NHCs as DNA sources, and the requirements for minimal damage to the specimens, should encourage the generation of genomic datasets comparable among studies.

Sections containing **data availability statement**, **author contributions**, **funding**, **acknowledgments,** and **conflict of interest** that appear in the published version can be accessed in Appendix I.

2.6. References

Allentoft ME, Collins M, Harker D, Haile J, Oskam CL, Hale ML, Campos PF, Samaniego JA, Gilbert TPM, Willerslev E, et al. 2012. The half-life of DNA in bone: Measuring decay kinetics in 158 dated fossils. Proc. R. Soc. B Biol. Sci. 279:4724–4733.

Andrews S. 2018. FastQC: A quality control tool for high throughput sequence data (v. 0.11.7).

Bi K, Linderoth T, Singhal S, Vanderpool D, Patton JL, Nielsen R, Moritz C. 2019. Temporal genomic contrasts reveal rapid evolutionary responses in an alpine mammal during recent climate change. PLOS Genet. 15:e1008119.

Bi K, Linderoth T, Vanderpool D, Good JM, Nielsen R, Moritz C. 2013. Unlocking the vault: Next-generation museum population genomics. Mol. Ecol. 22:6018–6032.

Burrell AS, Disotell TR, Bergey CM. 2015. The use of museum specimens with high-throughput DNA sequencers. J. Hum. Evol. 79:35–44.

Card DC, Shapiro B, Giribet G, Moritz C, Edwards S V. 2021. Museum Genomics. Annu. Rev. Genet. 55:633–659.

Carpenter ML, Buenrostro JD, Valdiosera C, Schroeder H, Allentoft ME, Sikora M, Rasmussen M, Gravel S, Guillén S, Nekhrizov G, et al. 2013. Pulling out the 1%: Whole-Genome capture for the targeted enrichment of ancient dna sequencing libraries. Am. J. Hum. Genet. 93:852–864.

Casas-Marce M, Revilla E, Fernandes M, Rodriguez A, Delibes M, Godoy JA. 2012. The value of hidden scientific resources: Preserved animal specimens from private collections and small museums. Bioscience 62:1077–1082.

Casas-Marce M, Revilla E, Godoy JA. 2010. Searching for DNA in museum specimens: A comparison of sources in a mammal species. Mol. Ecol. Resour. 10:502–507.

Charlton S, Booth T, Barnes I. 2019. The problem with petrous? A consideration of the potential biases in the utilization of pars petrosa for ancient DNA analysis. World Archaeol. 51:574–585.

Cruz-Dávalos DI, Llamas B, Gaunitz C, Fages A, Gamba C, Soubrier J, Librado P, Seguin-Orlando A, Pruvost M, Alfarhan AH, et al. 2017. Experimental conditions improving in-solution target enrichment for ancient DNA. Mol. Ecol. Resour. 17:508–522.

Dabney J, Knapp M, Glocke I, Gansauge MT, Weihmann A, Nickel B, Valdiosera C, García N, Pääbo S, Arsuaga JL, et al. 2013. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc. Natl. Acad. Sci. 110:15758–15763.

Dabney J, Meyer M. 2019. Extraction of highly degraded DNA from ancient bones and teeth. In: Methods in Molecular Biology. Vol. 1963. Humana Press Inc. p. 25–29.

Dehasque M, Pečnerová P, Kempe Lagerholm V, Ersmark E, Danilov GK, Mortensen P, Vartanyan S, Dalén L. 2022. Development and Optimization of a Silica Column-Based Extraction Protocol for Ancient DNA. Genes 13:687.

DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, Philippakis AA, Angel G del, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43:491–498.

Derkarabetian S, Benavides LR, Giribet G. 2019. Sequence capture phylogenomics of historical ethanol-preserved museum specimens: Unlocking the rest of the vault. Mol. Ecol. Resour. 19:1531–1544.

Eisenhofer R, Minich JJ, Marotz C, Cooper A, Knight R, Weyrich LS. 2019. Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. Trends Microbiol. 27:105–117.

Ewels P, Magnusson M, Lundin S, Käller M. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32:3047–3048.

Farrington OC. 1915. The rise of Natural History Museums. Science 42:197–208.

Fontsere C, Alvarez-Estape M, Lester J, Arandjelovic M, Kuhlwilm M, Dieguez P, Agbor A, Angedakin S, Ayuk Ayimisin E, Bessone M, et al. 2021. Maximizing the acquisition of unique reads in noninvasive capture sequencing experiments. Mol. Ecol. Resour. 21:745–761.

Fountain ED, Pauli JN, Reid BN, Palsbøll PJ, Peery MZ. 2016. Finding the right coverage: the impact of coverage and sequence quality on single nucleotide polymorphism genotyping error rates. Mol. Ecol. Resour. 16:966–978.

Gautier M, Gharbi K, Cezard T, Foucaud J, Kerdelhué C, Pudlo P, Cornuet J-M, Estoup A. 2013. The effect of RAD allele dropout on the estimation of genetic variation within and between populations. Mol. Ecol. 22:3165–3178.

Gel B, Serra E. 2017. KaryoploteR: An R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics 33:3088–3090.

Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, et al. 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature 522:207–211.

Hall LM, Willcox MS, Jones DS. 1997. Association of enzyme inhibition with methods of museum skin preparation. Biotechniques 22:928–934.

Harvey MG, Smith BT, Glenn TC, Faircloth BC, Brumfield RT. 2016. Sequence Capture versus Restriction Site Associated DNA Sequencing for Shallow Systematics. Syst. Biol. 65:910–924.

Heng L. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-Bio].

Hernandez-Rodriguez J, Arandjelovic M, Lester J, de Filippo C, Weihmann A, Meyer M, Angedakin S, Casals F, Navarro A, Vigilant L, et al. 2018. The impact of endogenous content, replicates and pooling on genome capture from faecal samples. Mol. Ecol. Resour. 18:319–333.

Holmes MW, Hammond TT, Wogan GOU, Walsh RE, Labarbera K, Wommack EA, Martins FM, Crawford JC, Mack KL, Bloch LM, et al. 2016. Natural history collections as windows on evolutionary processes. Mol. Ecol. 25:864–881.

Horváth MB, Martínez-Cruz B, Negro JJ, Kalmár L, Godoy JA. 2005. An overlooked DNA source for non-invasive genetic analysis in birds. J. Avian Biol. 36:84–88.

Hung CM, Shaner PJL, Zink RM, Liu WC, Chu TC, Huang WS, Li SH. 2014. Drastic population fluctuations explain the rapid extinction of the passenger pigeon. Proc. Natl. Acad. Sci. 111:10636–10641.

Jones MR, Good JM. 2016. Targeted capture in evolutionary and ecological genomics. Mol. Ecol. 25:185–202.

Kassambara A. 2020. ggpubr R Package: ggplot2-Based Publication Ready Plots. 1–155.

Kircher M, Sawyer S, Meyer M. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res. 40.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987–2993.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup 1000 Genome Project Data Processing. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Lim HC, Braun MJ. 2016. High-throughput SNP genotyping of historical and modern samples of five bird species via sequence capture of ultraconserved elements. Mol. Ecol. Resour. 16:1204–1223.

Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ, Zody MC, et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438:803–819.

Lonsinger RC, Daniel D, Adams JR, Waits LP. 2019. Consideration of sample source for establishing reliable genetic microsatellite data from mammalian carnivore specimens held in natural history collections. J. Mammal. 100:1678–1689.

Lopez L, Turner KG, Bellis ES, Lasky JR. 2020. Genomics of natural history collections for understanding evolution in the wild. Mol. Ecol. Resour. 20:1153–1160.

Maebe K, Meeus I, Vray S, Claeys T, Dekoninck W, Boevé JL, Rasmont P, Smagghe G. 2016. A century of temporal stability of genetic diversity in wild bumblebees. Sci. Rep. 6.

McCormack JE, Tsai WLE, Faircloth BC. 2016. Sequence capture of ultraconserved elements from bird museum specimens. Mol. Ecol. Resour. 16:1189–1203.

McDonough MM, Parker LD, McInerney NR, Campana MG, Maldonado JE. 2018. Performance of commonly requested destructive museum samples for mammalian genomic studies. J. Mammal. 99:789–802.

Meyer M, Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb. Protoc. 5.

Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. 12:443–451.

Pálsdóttir AH, Bläuer A, Rannamäe E, Boessenkool S, Hallsson JH. 2019. Not a limitless resource: Ethics and guidelines for destructive sampling of archaeofaunal remains. R. Soc. Open Sci. 6.

Pearson KD, Nelson G, Aronson MFJ, Bonnet P, Brenskelle L, Davis CC, Denny EG, Ellwood

ER, Goëau H, Mason Heberling J, et al. 2020. Machine learning using digitized herbarium specimens to advance phenological research. Bioscience 70:610–620.

Pinhasi R, Fernandes D, Sirak K, Novak M, Connell S, Alpaslan-Roodenberg S, Gerritsen F, Moiseyev V, Gromov A, Raczky P, et al. 2015. Optimal ancient DNA yields from the inner ear part of the human petrous bone. PLoS One 10:e0129102.

Pompanon F, Bonin A, Bellemain E, Taberlet P. 2005. Genotyping errors: causes, consequences and solutions. Nat. Rev. Genet. 6:847–859.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. 81:559–575.

Puritz JB, Hollenbeck CM, Gold JR. 2014. dDocent: A RADseq, variant-calling pipeline designed for population genomics of non-model organisms. PeerJ 2014:e431.

R Development Core Team. 2017. R: A language and environment for statistical computing. Vienna, Austria.

Raxworthy CJ, Smith BT. 2021. Mining museums for historical DNA: advances and challenges in museomics. Trends Ecol. Evol. 36:1049–1060.

Robertson T, Döring M, Guralnick R, Bloom D, Wieczorek J, Braak K, Otegui J, Russell L, Desmet P. 2014. The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet. PLoS One 9:e102623.

Rowe KC, Singhal S, Macmanes MD, Ayroles JF, Morelli TL, Rubidge EM, Bi K, Moritz CC. 2011. Museum genomics: Low-cost and high-accuracy genetic data from historical specimens. Mol. Ecol. Resour. 11:1082–1092.

Schubert M, Ermini L, Sarkissian C Der, Jónsson H, Ginolhac A, Schaefer R, Martin MD, Fernández R, Kircher M, McCue M, et al. 2014. Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. Nat. Protoc. 9:1056–1082.

Schubert M, Lindgreen S, Orlando L. 2016. AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. BMC Res. Notes 9.

Schwartz MK, Luikart G, Waples RS. 2007. Genetic monitoring as a promising tool for conservation and management. Trends Ecol. Evol. 22:25–33.

von Seth J, Dussex N, Díez-del-Molino D, van der Valk T, Kutschera VE, Kierczak M, Steiner CC, Liu S, Gilbert MTP, Sinding M-HS, et al. 2021. Genomic insights into the conservation status of the world's last remaining Sumatran rhinoceros populations. Nat. Commun. 12:2393.

Smith BT, Harvey MG, Faircloth BC, Glenn TC, Brumfield RT. 2014. Target Capture and Massively Parallel Sequencing of Ultraconserved Elements for Comparative Studies at Shallow Evolutionary Time Scales. Syst. Biol. 63:83–95.

Staats M, Erkens RHJ, Vossenberg B van de, Wieringa JJ, Kraaijeveld K, Stielow B, Geml J, Richardson JE, Bakker FT. 2013. Genomic Treasure Troves: Complete Genome Sequencing of Herbarium and Insect Museum Specimens. PLoS One 8:e69189.

Suarez A, Tsutsui N. 2004. The Value of Museum Collections for Research and Society. Bioscience 54:66–74.

Tsai WLE, Schedl ME, Maley JM, McCormack JE. 2020. More than skin and bones: Comparing extraction methods and alternative sources of DNA from avian museum specimens. Mol.

Ecol. Resour. 20:1220–1227.

Tsangaras K, Greenwood AD. 2012. Museums and disease: Using tissue archive and museum samples to study pathogens. Ann. Anat. 194:58–73.

van der Valk T, Lona Durazo F, Dalén L, Guschanski K. 2017. Whole mitochondrial genome capture from faecal samples and museum-preserved specimens. Mol. Ecol. Resour. 17:e111–e121.

van der Valk T, Vezzi F, Ormestad M, Dalén L, Guschanski K. 2020. Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. Mol. Ecol. Resour. 20:1171–1181.

Vallender EJ. 2011. Expanding whole exome resequencing into non-human primates. Genome Biol.129 12:1–10.

Wandeler P, Hoeck PEA, Keller LF. 2007. Back to the future: museum specimens in population genetics. Trends Ecol. Evol. 22:634–642.

Weiß CL, Schuenemann VJ, Devos J, Shirsekar G, Reiter E, Gould BA, Stinchcombe JR, Krause J, Burbano HA. 2016. Temporal patterns of damage and decay kinetics of dna retrieved from plant herbarium specimens. R. Soc. Open Sci. 3.

White LC, Fontsere C, Lizano E, Hughes DA, Angedakin S, Arandjelovic M, Granjon AC, Hans JB, Lester JD, Rabanus-Wallace MT, et al. 2019. A roadmap for high-throughput sequencing studies of wild animal populations using noninvasive samples and hybridization capture. Mol. Ecol. Resour. 19:609–622.

Wisely SM, Maldonado JE, Fleischer RC. 2004. A technique for sampling ancient DNA that minimizes damage to museum specimens. Conserv. Genet. 5:105–107.

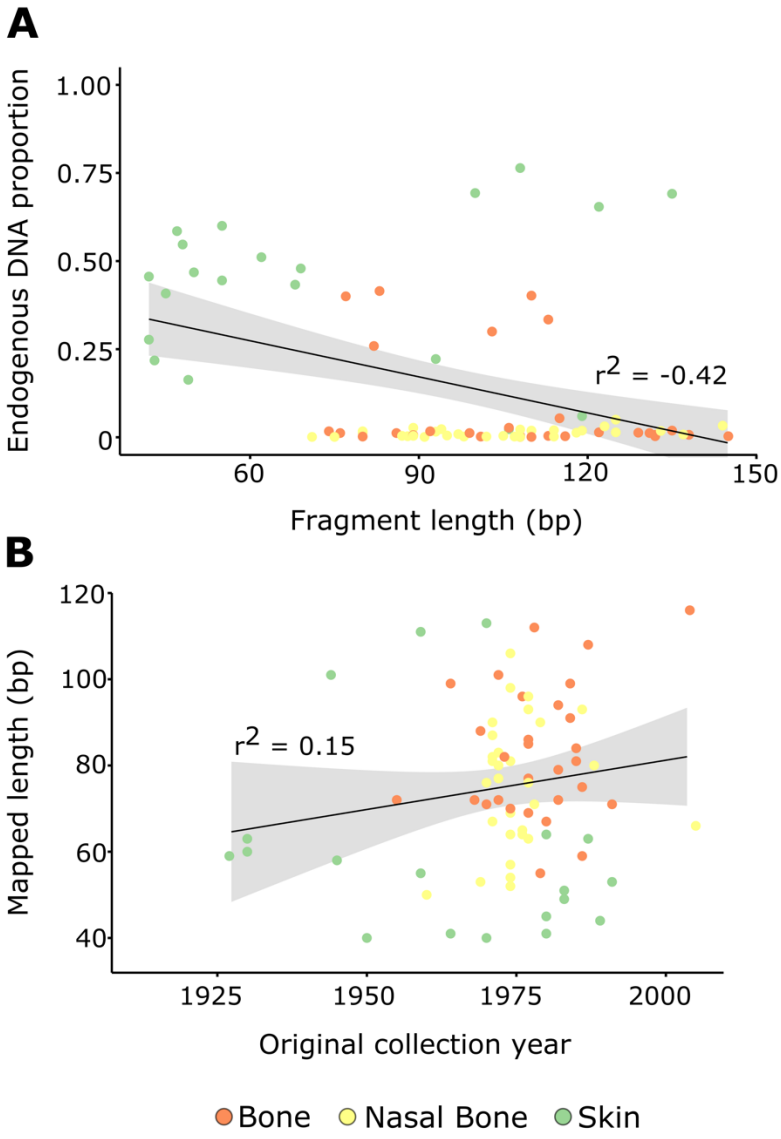
Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS, et al. 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329:75–78.

2.7. Supplementary Materials

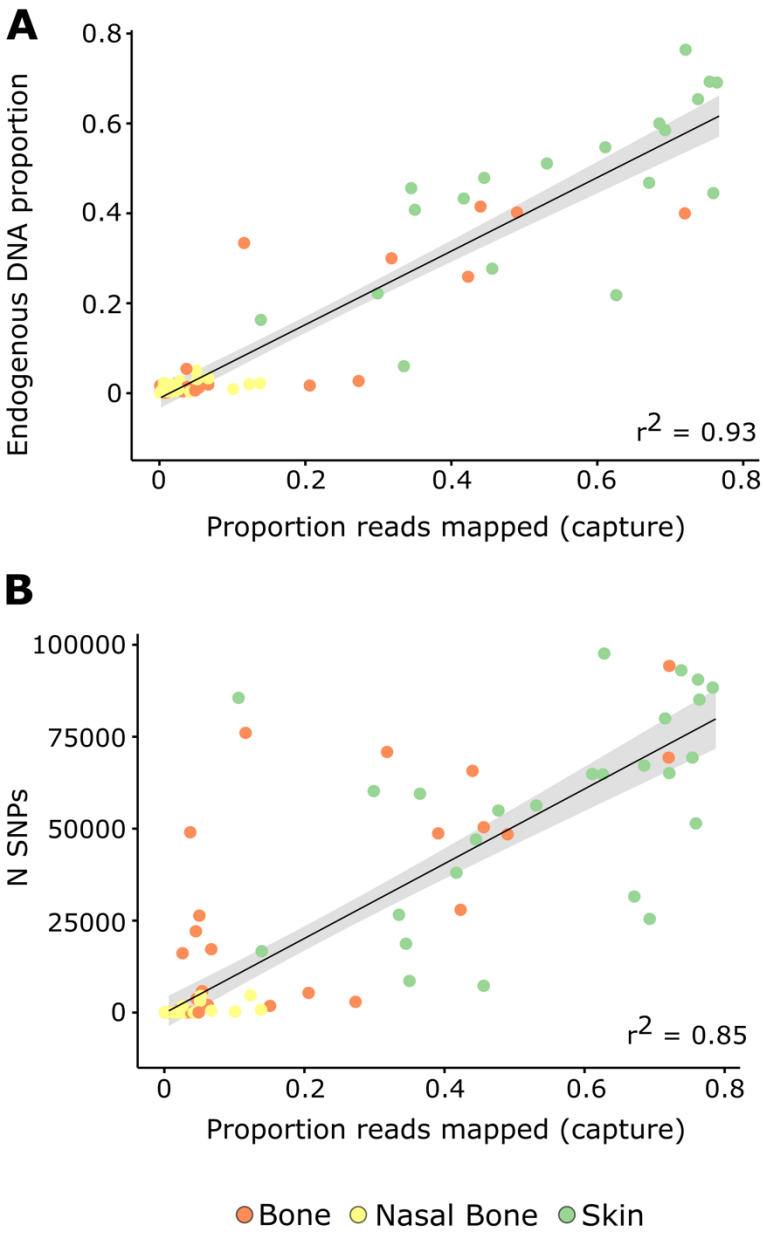
(**Additional Dataset 1** is an excel file containing detailed sampling information about the historical samples used in this study, as well as detailed statistics about the mapping of sequencing reads following shotgun resequencing and capture enrichment and is available online; see Appendix I for more information).



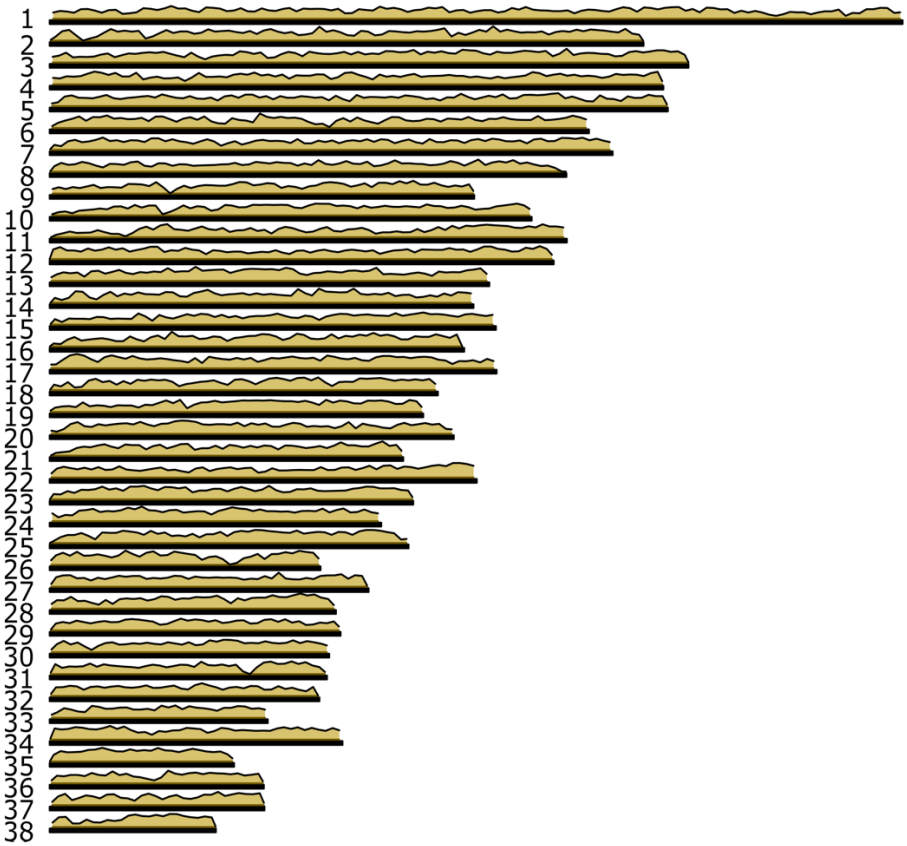
Supplementary Figure 2.1. Photographs showing the drilling location in femur (upper) and jaw (bottom) bones (delimited by a red circle). In the bottom image, marks left on the jaw from previous sampling by cutting a bone fragment are visible. Image credits Diana Lobo.



Supplementary Figure 2.2. Correlation coefficient (r^2) between (**A**) the proportion of endogenous DNA content and fragment length in shotgun sequencing (average length of reads to map). (**B**) the mapped length in shotgun sequencing (average length of reads mapped and unique MQ>20) and the specimen collection year. Each dot represents a historical sample. Colors depict the three different tissue types used in this study: orange - bones (jaw and femur); yellow - maxilloturbinal (nasal) bones; green – skins.



Supplementary Figure 2.3. Correlation coefficients (r^2) between (**A**) Endogenous DNA content assessed through shotgun sequencing and the proportion of reads mapping to the genome after capture. (**B**) Number of genotyped SNPs and the proportion of reads mapping to the genome after capture. Each dot represents a historical sample. Colors depict the three different tissue types used in this study: orange - bones (jaw and femur); yellow - maxilloturbinal (nasal) bones; green - skins.



Supplementary Figure 2.4. Genome-wide distribution and density of the 99,982 SNPs designed for the capture panel across the 38 dog autosomes.

Supplementary Table 2.1. Historical wolf samples. Description of the 103 historical wolf samples analyzed in this study using the enrichment target approach: skins, N = 29; bones, N = 43; maxilloturbinal (nasal) bones, N = 31. Information about original collection date (year), origin (museum or private collection) and collection ID (museum collections only) are provided.

Sample ID	Sample material	Year	Origin	Collection ID
CI-H001	Bones	1974	MNCN	16155
CI-H006	Bones	1970	MNCN	16218
CI-H011	Bones	1977	MNCN	16132
CI-H017	Bones	1972	MNCN	16136
CI-H018	Bones	1976	MNCN	16127
CI-H020	Bones	1974	MNCN	16141
CI-H027	Skin	1983	MNCN	16259
CI-H028	Bones	1984	MNCN	16237
CI-H032	Bones	1983	MNCN	16202
CI-H034	Bones	1984	MNCN	16212
CI-H036	Bones	1984	MNCN	16118
CI-H038	Bones	1984	MNCN	16200
CI-H039	Bones	1984	MNCN	16310
CI-H040	Skin	1983	MNCN	16245
CI-H042	Bones	1982	MNCN	16154
CI-H044	Bones	1977	MNCN	16311
CI-H046	Bones	1986	MNCN	16383
CI-H051	Bones	1983	EBD	4371
CI-H052	Skin	1979	EBD	7826
CI-H053	Bones	1982	EBD	3032
CI-H061	Nasal bone	1987	MNHNC	87.02.42
CI-H062	Bones	1987	MNHNC	87.09.41
CI-H066	Bones	1985	MNHNC	85.06.26
CI-H067	Bones	1985	MNHNC	85.10.38
CI-H068	Bones	1986	MNHNC	86.09.46
CI-H071	Bones	1977	MNHNC	77.11.01
CI-H072	Bones	1978	MNHNC	78.01.03
CI-H073	Bones	1982	MNHNC	82.00.35
CI-H076	Bones	1982	MNHNC	82.09.09
CI-H077	Bones	1983	MNHNC	83.07.06
CI-H078	Bones	1984	MNHNC	84.00.13
CI-H083	Skin	1972	P. collection #1	-
CI-H084	Skin	1972	P. collection #1	-
CI-H086	Skin	1912	P. collection #2	-
CI-H090	Skin	1980	P. collection #3	-
CI-H092	Skin	1930	P. collection #4	-
CI-H093	Skin	1930	P. collection #4	-
CI-H095	Skin	1930	P. collection #4	-
CI-H098	Skin	1959	P. collection #5	-
CI-H103	Skin	1959	P. collection #6	-
CI-H105	Skin	1959	P. collection #6	-
CI-H116	Skin	1927	P. collection #7	-
CI-H117	Skin	1970	P. collection #8	-
CI-H119	Skin	1944	P. collection #9	-
CI-H126	Skin	1973	P. collection #10	-
CI-H129	Skin	1991	MNHNC	91.00.61
CI-H131	Skin	1987	P. collection #11	-
CI-H132	Bones	1955	P. collection #12	-
CI-H135	Skin	1945	P. collection #13	-
CI-H139	Bones	1980	P. collection #14	-
CI-H140	Skin	1980	P. collection #14	-
CI-H143	Skin	1970	P. collection #14	-
CI-H145	Skin	1980	P. collection #14	-
CI-H146	Skin	1980	P. collection #14	-
CI-H148	Skin	1980	P. collection #14	-
CI-H149	Skin	1970	P. collection #15	-
CI-H159	Nasal bone	1976	EBD	2867
CI-H160	Bones	1977	EBD	2871
CI-H161	Nasal bone	1977	EBD	2872
CI-H166	Bones	1979	EBD	3144
CI-H167	Nasal bone	1979	EBD	3356
CI-H175	Nasal bone	1977	EBD	6119
CI-H177	Bones	1984	EBD	7534
CI-H179	Nasal bone	1974	EBD	7821
CI-H180	Nasal bone	1972	EBD	7822
CI-H182	Bones	1964	EBD	7831

Sample ID	Sample material	Year	Origin	Collection ID
CI-H183	Nasal bone	1971	EBD	7835
CI-H187	Nasal bone	1972	EBD	7841
CI-H189	Nasal bone	1970	EBD	7844
CI-H192	Nasal bone	1969	EBD	7848
CI-H193	Nasal bone	1971	EBD	7849
CI-H198	Nasal bone	1972	EBD	7860
CI-H200	Nasal bone	1974	EBD	7862
CI-H202	Nasal bone	1971	EBD	7866
CI-H203	Nasal bone	1974	EBD	7873
CI-H206	Bones	1972	EBD	7878
CI-H214	Nasal bone	1976	EBD	14866
CI-H216	Nasal bone	1974	EBD	14871
CI-H217	Nasal bone	1974	EBD	14872
CI-H218	Nasal bone	1974	EBD	14873
CI-H219	Nasal bone	1978	EBD	14881
CI-H221	Bones	1974	EBD	15021
CI-H222	Nasal bone	1971	EBD	15023
CI-H223	Bones	1970	EBD	15024
CI-H230	Nasal bone	1974	EBD	15039
CI-H232	Nasal bone	1977	EBD	15122
CI-H234	Bones	1969	EBD	15125
CI-H235	Bones	1968	EBD	15126
CI-H237	Bones	1972	EBD	15133
CI-H244	Bones	1973	EBD	15857
CI-H246	Bones	1977	EBD	15864
CI-H247	Nasal bone	1974	EBD	15865
CI-H249	Nasal bone	2005	EBD	15872
CI-H251	Nasal bone	1971	EBD	15880
CI-H256	Skin	1964	EBD	15932
CI-H258	Nasal bone	1986	EBD	18002
CI-H259	Nasal bone	1977	EBD	18004
CI-H260	Nasal bone	1988	EBD	18049
CI-H261	Bones	1991	EBD	19329
CI-H268	Skin	1989	EBD	23580
CI-H273	Bones	bones	EBD	27775
CI-H276	Nasal bone	1960	EBD	27942
CI-H278	Skin	1950	EBD	30410

Museum Collections:
MNCN – Museo Nacional de Ciencias Naturales (Madrid, Spain)
EBD – Estación Biológica de Doñana (Seville, Spain)
MNHNC – Museu Nacional de História Natural e da Ciência (Lisbon, Portugal)

Private Collections:
P. collection #1: Clube de caça e pesca da Covilhã (Gouveia, Portugal)
P. collection #2: Liceu Nuno Álvares (Louriçal do Campo, Castelo Branco, Portugal)
P. collection #3: Colégio/Centro Social "Padres Redentoristas" (Castelo Branco, Portugal)
P. collection #4: Herdade do Rochão (Idanha-a-Nova, Portugal)
P. collection #5: Herdade das Russianas (Barrancos, Portugal)
P. collection #6: Tapada de Mafra (Mafra, Portugal)
P. collection #7: Finca Las Juntas (Santa Elena, Jaén, Spain)
P. collection #8: Finca La Garganta (Fuencaliente, Ciudad Real, Spain)
P. collection #9: Finca El Almendral (Oliva de Plasencia, Cáceres, Spain)
P. collection #10: Private owner (Grândola, Portugal)
P. collection #11: Private owner (Melgaço, Portugal)
P. collection #12: Private owner (Vieira do Minho, Portugal)
P. collection #13: Private owner (Boticas, Portugal)
P. collection #14: Private owner (Bragança, Portugal)
P. collection #15: Private owner (Fafe, Portugal)

Supplementary Table 2.2. Mapping statistics. Percentage of endogenous DNA content, mapping reads after capture and reads mapped on target regions, in relation to all or mapped reads across all samples. All the reported mapping reads are unique and with mapping quality above 20.

Sample ID	% DNA endogenous	% Reads mapped capture	% Reads mapped target regions (mapped)	% Reads mapped target regions (all)
CI-H001	nd	2.34	14.05	0.33
CI-H006	nd	0.76	24.05	0.18
CI-H011	nd	4.64	19.66	0.91
CI-H017	nd	72.14	24.76	17.86
CI-H018	1.65	0.13	21.72	0.03
CI-H020	nd	45.59	10.30	4.70
CI-H027	21.84	62.63	19.36	12.12
CI-H028	nd	0.96	11.67	0.11
CI-H032	nd	4.54	8.90	0.40
CI-H034	nd	4.96	18.86	0.94
CI-H036	nd	39.08	25.69	10.04
CI-H038	nd	2.58	23.28	0.60
CI-H039	nd	0.22	12.00	0.03
CI-H040	60.01	68.50	27.59	18.90
CI-H042	25.86	42.33	1.00	5.58
CI-H044	0.26	0.14	4.31	0.01
CI-H046	1.23	0.55	7.22	0.04
CI-H051	nd	0.39	5.51	0.02
CI-H052	nd	62.80	30.10	18.90
CI-H053	nd	6.22	27.29	1.70
CI-H061	nd	0.22	24.84	0.05
CI-H062	1.94	6.66	20.07	1.34
CI-H066	0.37	0.77	8.56	0.07
CI-H067	0.74	3.01	28.15	0.85
CI-H068	1.24	0.64	8.06	0.05
CI-H071	0.39	3.23	25.65	0.83
CI-H072	0.29	1.62	19.76	0.32
CI-H073	1.33	1.49	12.77	0.19
CI-H076	0.17	1.13	11.53	0.13
CI-H077	nd	15.11	35.19	5.32
CI-H078	0.10	0.59	18.35	0.11
CI-H083	nd	10.58	19.30	2.04
CI-H084	nd	47.73	32.13	15.34
CI-H086	nd	71.48	18.92	13.53
CI-H090	nd	76.19	23.89	18.21
CI-H092	2.26	2.42	3.91	0.09
CI-H093	43.33	41.67	13.44	5.60
CI-H095	nd	0.17	4.34	0.01
CI-H098	nd	0.98	22.70	0.22
CI-H103	76.35	72.13	29.20	21.06
CI-H105	6.03	33.50	14.04	4.70
CI-H116	47.94	44.52	23.87	10.63
CI-H117	69.08	76.39	36.38	27.79
CI-H119	65.39	73.79	31.89	23.53
CI-H126	nd	36.47	21.97	8.01
CI-H129	46.82	67.14	26.06	17.50
CI-H131	51.12	53.14	28.38	15.08
CI-H132	40.00	72.01	34.86	25.10
CI-H135	44.48	75.90	28.97	21.99
CI-H139	29.96	31.77	33.31	10.58
CI-H140	22.21	29.93	20.34	6.09
CI-H143	45.61	34.50	14.91	5.14
CI-H145	40.80	35.02	11.37	3.98
CI-H146	nd	78.26	33.54	26.25
CI-H148	58.47	69.33	31.67	21.96
CI-H149	69.25	75.43	25.64	19.34
CI-H159	0.06	0.42	12.34	0.05
CI-H160	0.26	0.32	17.69	0.06
CI-H161	2.25	0.60	19.62	0.12
CI-H166	1.24	1.59	21.53	0.34
CI-H167	1.82	1.40	24.84	0.35
CI-H175	0.08	0.44	13.11	0.06
CI-H177	33.41	11.65	27.39	3.19
CI-H179	1.28	0.58	17.46	0.10
CI-H180	0.18	0.74	26.14	0.19

Sample ID	% DNA endogenous	% Reads mapped capture	% Reads mapped target regions (mapped)	% Reads mapped target regions (all)
CI-H182	40.17	49.03	15.26	7.48
CI-H183	1.92	1.68	22.39	0.38
CI-H187	5.00	5.09	18.24	0.93
CI-H189	0.15	1.26	30.26	0.38
CI-H192	1.56	1.04	4.18	0.04
CI-H193	3.27	6.68	10.95	0.73
CI-H198	0.15	1.00	16.05	0.16
CI-H200	1.25	0.62	27.99	0.17
CI-H202	0.23	1.59	22.53	0.36
CI-H203	1.74	4.65	29.70	1.38
CI-H206	0.73	1.31	27.16	0.35
CI-H214	1.44	1.33	19.46	0.26
CI-H216	1.85	1.80	30.46	0.55
CI-H217	2.72	2.79	21.14	0.59
CI-H218	0.10	0.58	26.88	0.16
CI-H219	0.09	0.14	21.57	0.03
CI-H221	0.21	0.99	21.20	0.21
CI-H222	2.23	13.85	17.54	2.43
CI-H223	1.40	3.78	13.77	0.52
CI-H230	0.39	0.16	15.95	0.03
CI-H232	0.25	4.18	9.74	0.41
CI-H234	5.41	3.74	23.66	0.89
CI-H235	0.57	4.90	37.00	1.81
CI-H237	0.05	0.72	6.29	0.05
CI-H244	1.71	20.60	3.60	0.74
CI-H246	1.24	5.36	21.51	1.15
CI-H247	0.84	2.56	18.53	0.47
CI-H249	0.45	2.07	25.18	0.52
CI-H251	0.92	10.08	8.86	0.89
CI-H256	27.69	45.60	5.76	2.63
CI-H258	3.11	5.17	15.14	0.78
CI-H259	0.43	1.07	21.93	0.24
CI-H260	1.99	12.34	20.23	2.50
CI-H261	41.45	44.01	19.86	8.74
CI-H268	54.72	61.12	21.64	13.22
CI-H273	2.68	27.28	6.62	1.81
CI-H276	0.36	0.44	16.79	0.07
CI-H278	16.26	13.90	31.09	4.32

Abbreviation: nd, not determined.

Supplementary Table 2.3. GLM for endogenous DNA. Output of the Generalized Linear Model (GLM) with binomial distribution for variable "Endogenous DNA Proportion" across the three sample types. Bone was listed as the intercept. The fit of this model was confirmed using Pearson's $\chi 2$ residuals and deviance analysis (both > 0.05 and over dispersion < 3). Significant p values ($p < 0.05$) are represented in bold.

Effect	Estimate	SE	z value	Pr (>|z|)
Intercept (Bone)	-2.423	0.679		
Sample type: Nasal bone	-1.923	1.760	-1.092	0.275
Sample type: Skin	2.161	0.815	2.651	**0.008**

Supplementary Table 2.4. GLM for mapped reads after capture. Output of the GLM with binomial distribution for variable "Mapped Reads after Capture" across the three sample types. Bone was listed as the intercept. The fit of this model was confirmed using Pearson's $\chi 2$ residuals and deviance analysis (both > 0.05 and over dispersion < 3). Significant p values ($p < 0.05$) are represented in bold.

Effect	Estimate	SE	z value	Pr (>|z|)
Intercept (Bone)	-1.939	0.460		
Sample type: Nasal bone	-1.612	1.184	-1.362	0.173
Sample type: Skin	1.899	0.591	3.212	**0.001**

Supplementary Table 2.5. Control samples. Description of the 2 contemporary wolf samples used in this study as controls. These samples were genotyped using two techniques: Canine HD BeadChip from Illumina (~170K genome-wide SNPs) and the enrichment target approach (~100K genome-wide SNPs). Information about the number of genotyped SNPs in the SNP chip, mapping reads after capture (%) and reads mapped on target regions, in relation to all or mapped reads (%), are described. All the reported mapping reads are unique and with mapping quality above 20.

Sample ID	N SNPs	% Reads mapped capture	% Reads mapped target regions (mapped)	% Reads mapped target regions (all)
L474	121,507	33.41	52.45	17.52
L585	121,455	32.28	54.39	17.56

Supplementary Table 2.6. GLM for mapped reads on target regions (of mapped reads). Output of the GLM with binomial distribution for variable "Mapped reads on target regions in relation to mapped reads" across the three sample types. Bone was listed as the intercept. The fit of this model was confirmed using Pearson's $\chi 2$ residuals and deviance analysis (both > 0.05 and over dispersion < 3). Significant p values ($p < 0.05$) are represented in bold.

Effect	Estimate	SE	z value	Pr(>|z|)
Intercept (Bone)	-1.508	0.396		
Sample type: Nasal bone	0.101	0.601	0.167	0.867
Sample type: Skin	0.271	0.595	0.455	0.649

Supplementary Table 2.7. GLM for mapped reads on target regions (of all reads). Output of the GLM with binomial distribution for variable "Mapped reads on target regions in relation to all reads" across the three sample types. Bone was listed as the intercept. The fit of this model was confirmed using Pearson's $\chi 2$ residuals and deviance analysis (both > 0.05 and over dispersion < 3). Significant p values ($p < 0.05$) are represented in bold.

Effect	Estimate	SE	z value	Pr(>|z|)
Intercept (Bone)	-3.591	0.944		
Sample type: Nasal bone	-1.703	2.716	-0.627	0.531
Sample type: Skin	1.630	1.100	1.483	0.138

Supplementary Table 2.8. Genotyping errors in contemporary samples. Concordance rates between genotypes called using different thresholds as minimum coverage (depth ≥ 3, ≥ 4, ≥ 5 or exclusively 3x or 4x) in the enrichment approach with those obtained in the SNP chip. Empirical values for the concordance rates were calculated using the two control contemporary samples. The number of SNPs correspond to the effective number used in the comparison.

Depth	Sample ID	N SNPs	N SNPs discordant / concordant	Discordance / concordance rate (%)
≥ 3	L474	79,609	725 / 78,884	0.91 / 99.09
	L585	81,238	616 / 80,622	0.76 / 99.24
≥ 4	L474	76,983	570 / 76,413	0.74 / 99.26
	L585	79,117	504 / 78,613	0.64 / 99.36
≥ 5	L474	74,281	473 / 73,808	0.64 / 99.36
	L585	76,724	446 / 76,278	0.58 / 99.42
3	L474	2,626	155 / 2,471	5.9 / 94.10
	L585	2,123	112 / 2,009	5.3 / 94.70
4	L474	2,702	97 / 2,605	3.6 / 96.40
	L585	2,394	58 / 2,335	2.4 / 97.60

Supplementary Table 2.9. Number of genotyped SNPs. Number of SNPs obtained following the enrichment approach across historical samples, where 100,000 SNPs were targeted, considering different thresholds as minimum coverage (depth ≥ 3, ≥ 4 or ≥ 5). Results for the two control contemporary samples are also presented.

Sample ID	N SNPs (≥ 3)	N SNPs (≥ 4)	N SNPs (≥ 5)
CL-H001	3,936	1,079	278
CL-H006	414	48	12
CL-H011	9,078	3,709	1,458
CL-H017	96,382	94,257	91,984
CL-H018	27	5	0
CL-H020	60,480	50,390	42,380
CL-H027	73,153	64,762	57,613
CL-H028	1	0	0
CL-H032	33,407	22,097	14,767
CL-H034	37,711	26,357	18,188
CL-H036	59,230	48,716	40,006
CL-H038	26,539	16,120	9,441
CL-H039	8	2	0
CL-H040	74,992	67,179	60,352
CL-H042	38,922	27,944	20,468
CL-H044	0	0	0
CL-H046	12	3	1
CL-H051	0	0	0
CL-H052	98,477	97,621	96,649
CL-H053	5,862	2,033	647
CL-H061	6	0	0
CL-H062	27,753	17,227	10,449
CL-H066	68	8	1
CL-H067	95	11	1
CL-H068	4	2	1
CL-H071	32	3	1
CL-H072	338	51	8
CL-H073	895	225	55
CL-H076	6	1	0
CL-H077	5,235	1,819	567
CL-H078	17	3	0
CL-H083	89,702	85,562	81,401
CL-H084	64,069	54,930	47,051
CL-H086	85,545	79,974	74,646
CL-H090	93,746	90,526	87,200
CL-H092	528	148	54
CL-H093	46,774	38,012	31,118
CL-H095	0	0	0
CL-H098	14	0	0
CL-H103	73,307	65,105	57,888
CL-H105	37,029	26,577	19,342
CL-H116	57,111	47,019	39,081
CL-H117	89,233	85,084	80,999
CL-H119	95,331	93,062	90,649
CL-H126	68,573	59,499	51,732
CL-H129	42,706	31,543	23,313
CL-H131	65,232	56,302	48,721
CL-H132	76,309	69,323	62,938
CL-H135	60,758	51,432	43,671
CL-H139	77,381	70,853	64,921
CL-H140	68,659	60,217	52,983
CL-H143	26,076	18,714	13,728
CL-H145	13,826	8,597	5,548
CL-H146	91,851	88,357	84,892
CL-H148	34,306	25,459	19,148
CL-H149	77,267	69,317	61,921
CL-H159	1	1	1
CL-H160	112	15	1
CL-H161	1	0	0
CL-H166	0	0	0
CL-H167	905	164	28
CL-H175	1	0	0
CL-H177	81,923	76,041	70,567
CL-H179	801	154	21
CL-H180	49	8	0
CL-H182	58,179	48,473	40,510

Sample ID	N SNPs (≥ 3)	N SNPs (≥ 4)	N SNPs (≥ 5)
CL-H183	5	0	0
CL-H187	8,114	3,136	1,144
CL-H189	11	0	0
CL-H192	0	0	0
CL-H193	1,801	514	160
CL-H198	19	1	0
CL-H200	0	0	0
CL-H202	211	24	6
CL-H203	42	3	0
CL-H206	0	0	0
CL-H214	433	67	8
CL-H216	2	0	0
CL-H217	35	5	0
CL-H218	7	0	0
CL-H219	3	0	0
CL-H221	23	1	0
CL-H222	2,537	741	188
CL-H223	441	57	12
CL-H230	1	0	0
CL-H232	48	2	0
CL-H234	58,938	49,011	40,630
CL-H235	340	37	8
CL-H237	2	1	1
CL-H244	9,495	5,349	3,365
CL-H246	12,702	5,822	2,459
CL-H247	5,364	1,728	524
CL-H249	393	54	7
CL-H251	1,113	269	73
CL-H256	11,001	7,284	5,171
CL-H258	10,647	4,509	1,849
CL-H259	16	2	0
CL-H260	11,031	4,666	1,822
CL-H261	73,823	65,720	58,567
CL-H268	72,348	64,850	58,448
CL-H273	6,020	2,901	1,511
CL-H276	7	0	0
CL-H278	22,972	16,685	12,515
L474	93,011	89,944	86,825
L585	94,884	92,346	89,521

Chapter III.

Dog introgression into the Iberian wolf genome may have facilitated adaptation to human-dominated landscapes

Lobo D. et al.

(in preparation to be submitted)

Abstract

The Iberian wolf subspecies exhibits a unique set of behavioral phenotypes not known within other wolf populations – low dispersal ability, marked consumption of human-related resources, and tolerance towards human activities. While many wolf populations have historically gone extinct across Europe, the Iberian wolf managed to cope and persist in highly humanized landscapes. We hypothesize whether the ability of wolves to persist in human-dominated landscapes is associated with adaptive introgression from domestic dogs. To identify evolutionary genomic signatures of dog introgression and selection within the Iberian wolf genome, we analyzed 146 historical and contemporary samples. Although the overall dog ancestry in the wolf genome was ~1%, confirming that the Iberian wolf genome is not threatened by admixture, we found a 90 kb dog haplotype in chromosome 2 of ~30% of the Iberian wolf population, not shared with other of the analyzed wolf populations. This haplotype introgressed 3,624 (3,572-3,676; 5-95% CI) years ago, likely through a single hybridization event. The introgressed haplotype displayed patterns consistent with selection in Iberian wolves following introgression, and forward genetic simulations rejected neutral evolution indicating that balancing selection offers the best explanation of the empirical data. The introgressed haplotype is inserted in the MAST4 gene, which interacts with glutamate receptor genes. Because glutamate neurotransmitters are associated with mammal domestication for regulating levels of anxiety and stress responses, we hypothesize that dog adaptive introgression may have altered the Iberian wolf behavior, by increasing their tolerance towards humans and human activities.

3.1. Introduction

Circa 10 thousand years ago (ka), humans started to dramatically change the planet prompted by the advent of farming following the domestication of plants and animals (Frantz et al. 2020). Over time, large carnivores have been severely affected by these rapid environmental changes, mostly because of the increasing competition with humans (e.g., persecution mediated by livestock depredation), and many have gone (locally) extinct (Chapron et al. 2014). Notably, some species managed to change their behavior and ecology to adapt to human activities in human-dominated landscapes (e.g., Sazatornil et al. 2016; Benazzo et al. 2017; Rio-Maior et al. 2019). A remarkable example of such persistence is illustrated by Eurasian gray wolves (*Canis lupus*) living in highly human-dominated landscapes, such as those found in the Iberian Peninsula (i.e., Portugal and Spain; 20-400 inhabitants/km^2 in the wolf range, min-max; INE, 2016).

Despite the high genetic similarity between Eurasian gray wolves and domestic dogs (0.04 – 0.21% nucleotide differentiation; Lindblad-Toh et al. 2005), they diverged markedly in behavior and morphology, principally as the result of domestication 40-14 ka from an as yet unidentified wolf population (Skoglund et al. 2015; Fan et al. 2016; Frantz et al. 2016; Perri et al. 2021; Bergström et al. 2022). Additionally, wolves and dogs can interbreed, and their hybrids can backcross into both the wild and the domestic populations. In the last decades, wolf-dog anthropogenic hybridization has been viewed as a major conservation threat to wolf populations, as the introgression of genes originally selected by humans may disrupt local adaptations to the wild (Rhymer and Simberloff 1996; Allendorf et al. 2001). But cases of adaptive introgression have also been reported (Anderson et al. 2009; Wang et al. 2020), suggesting it can also be a powerful force in the evolutionary response of species facing new environmental conditions (Grabenstein and Taylor 2018; Taylor and Larson 2019). Thus, it is plausible to speculate that the introgression of dog variants into the wolf genome could have facilitated adaptation to human-dominated landscapes.

The Iberian wolf subspecies (*Canis lupus signatus*; **Figure 3.1C**) may have been prone to such “introgression-fueled adaptation” for several reasons. First, hybridization between Iberian wolves and dogs is both ancient and contemporary (Godinho et al. 2011; Fan et al. 2016; Gómez-Sánchez et al. 2018). Second, the landscape in the Iberian Peninsula has been profoundly shaped by humans and livestock for several thousand years (Zeder 2008; Revelles et al. 2018), which means that the selection pressure has

been high. Indeed, Iberian wolves can persist in densely human-populated areas and tolerate high levels of human activity (Llaneza et al. 2012; Llaneza et al. 2016; Dennehy et al. 2021), showing a remarkable capability to feed on multiple human-related food sources, including livestock, carrion or garbage (Llaneza and López-Bao 2015; Newsome et al. 2016). Third, the Iberian wolf displays a very low propensity to disperse (Silva et al. 2018), and it has been genetically isolated from other Eurasian wolves for ca. 10.3 thousand years (Silva et al. 2020). This means that local adaptations are less likely to have been swamped and displaced by immigrant variants. This unique set of behavioral phenotypes is not known among other Eurasian wolf populations.

To test the hypothesis that introgression from domestic dogs may have facilitated Iberian wolf adaptation to human-dominated landscapes, we generated single nucleotide polymorphisms (SNP) for a total of 48 historical and 95 contemporary Iberian wolf samples. Additionally, we also analyzed 15 whole-genome sequences (WGS) from contemporary samples of Iberian wolf. We used these data to 1) evaluate the evidence for dog introgression in the Iberian wolf genome, 2) determine whether introgressed genes display signatures of selection, and 3) discuss the functional role of introgressed genes to understand their possible association with behavioral adaptations to human-dominated landscapes.

3.2. Results and Discussion

3.2.1. Presence of a dog block in chromosome 2 of Iberian wolves

We started by analyzing 85,000 genome-wide SNPs (hereafter 85K) across contemporary samples of unrelated 95 Iberian wolves, 55 Eurasian wolves, and 120 dogs genotyped using the Canine HD BeadChip (**Figures 3.1A**, **3.1B** and Supplementary Table 3.1; see Materials and Methods for detailed information). We observed lower levels of heterozygosity and increased linkage disequilibrium (LD) in Iberian wolves when compared to other Eurasian wolf populations (Supplementary Figures 3.1A, 3.1C). This is compatible with historical records of a major population decline in the last decades and geographical isolation (Silva et al. 2020; Nores and López-Bao 2022; Clavero et al. 2022). Almost all runs of homozygosity (ROH) in Iberian wolves were shorter than 20 Mb (~95%; Supplementary Figure 3.1B), suggesting strong effects of past demographic processes rather than recent inbreeding. A principal component analysis (PCA) and global ancestry estimation in ADMIXTURE (Alexander et

al. 2009) confirmed the documented genetic uniqueness of Iberian wolves relative to other Eurasian wolf populations and dogs (Pilot et al. 2014) (Supplementary Figures 3.2–3.4).

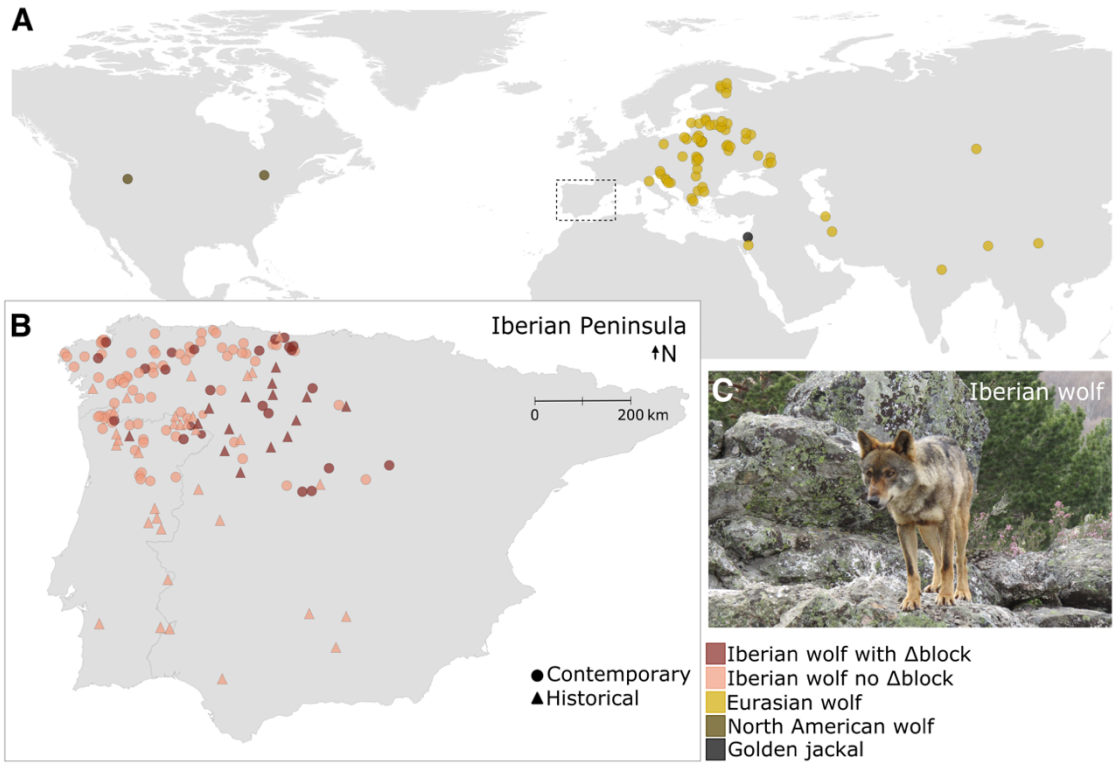


Figure 3.1. Sample location. (**A**) Geographical location of the analyzed canid samples across the world. Yellow, dark yellow and black dots denote samples from Eurasian wolves, North American wolves, and a golden jackal, respectively. (**B**) Zoomed-in view of the sampling area in the Iberian Peninsula, delimited by the dashed line in (A). Iberian wolves with and without the Δblock are represented in red and pink, respectively. Circles and triangles denote contemporary and historical samples. Further details about samples can be found in Supplementary Tables 3.1, 3.2, and 3.4. Dog samples are not represented. (**C**) Picture of an Iberian wolf (*Canis lupus signatus*); credits Raquel Godinho.

To identify signatures of dog ancestry across the Iberian wolf genome we applied two local ancestry methods, LAMP-ANC (Sankararaman et al. 2008; Pasaniuc et al. 2009) and ELAI (Guan 2014) to infer the ancestry of each SNP in the genome of Iberian wolves, considering Eurasian wolves as a reference population. We found that approximately 1% of the genome of Iberian wolves has a dog ancestry (Supplementary Figures 3.5, 3.6; Supplementary Text), which is similar to that observed in other Eurasian wolf populations (Fan et al. 2016; Gómez-Sánchez et al. 2018; Pilot et al. 2018; Pilot et al. 2021). Because blocks of dog ancestry can be stochastically spread across the wolf genome, we searched the local ancestry results for blocks commonly shared amongst

Iberian wolves. We identified one dog block on chromosome 2 spanning up to 11 Mb (**Figure 3.2B**; Supplementary Figure 3.7), detected by the genome-wide SNP data in ~30% of the population. To test whether this block had excessive dog ancestry in relation to the genome-wide average, we calculated the SNP-specific delta ancestry statistics (∆) (Tang et al. 2007) across the genome and used three standard deviations (3SD) from the mean ancestry as a threshold to identify outlier regions. This analysis confirmed an outlier region in chromosome 2 spanning ~2.6 Mb (hereafter ∆block) between positions 50.02 – 52.62 Mb (**Figure 3.2A**), comprising 78 SNPs and 22 protein-coding genes (CanFam 3.1 Ensembl annotation; Supplementary Table 3.3).

To increase the resolution over the Iberian wolf genome across the region harboring the ∆block, we complemented our dataset by sequencing the whole-genome of 12 of the Iberian wolves genotyped using the SNP chip (mean sequence coverage of 13.5x) and added publicly available WGS data from another three Iberian wolves, 10 worldwide wolves and 13 dogs (**Figure 3.1A, 3.1B**; Supplementary Table 3.2). We found the ∆block in five Iberian wolf whole-genomes (**Figure 3.3A**; Supplementary Figure 3.8), two of them carrying an intact block (chr 2: 43.5 – 55.1 Mb), whereas it was fragmented in the other three. This suggests that after a possible introgression event, recombination reassorted nucleotide variation across the genomic block into the wolf's genetic background, except for some key genomic regions that retained dog ancestry. The breakpoint positions and lengths of dog fragments were similar to those estimated using the 85K SNP dataset, except for one individual with a much smaller track only identified through WGS data (Supplementary Text).

To recreate phylogenetic relationships of Iberian wolves and dogs within the ∆block, we used the SNP chip dataset to construct a neighbor-joining (NJ) tree with the 78 SNPs spanning the ∆block. This analysis revealed that Iberian wolves with the ∆block (N = 21) cluster together with dogs and not with the other wolves where the ∆block was not observed (Supplementary Figure 3.9). Additionally, a median-joining haplotype network showed that all Iberian wolves with the ∆block shared the same haplotype in this genomic region, which is highly similar (70/78 SNPs) to a haplotype found among village dogs from Iberia (**Figure 3.2C**). To have a temporal perspective of the presence of ∆block in the population, we analyzed 48 historical samples of Iberian wolves, mostly from the 1970s (1912 to 2005; **Figure 3.1B**; Supplementary Table 3.4). For these samples, we used genotypes for a set of 18,000 genome-wide SNPs (hereafter 18K) overlapping the 85K SNP dataset, obtained through target capture enrichment (Materials and Methods; Pacheco, Lobo et al. 2022; Chapter II). By implementing local ancestry analysis using

LAMP-ANC, we found the Δblock present in 15 historical samples of Iberian wolf (Supplementary Figure 3.10, 3.11; Supplementary Text). The oldest record of the Δblock was observed in a wolf sampled in 1945. A NJ tree constructed with contemporary and historical samples of Iberian wolves and village dogs from Iberia indicated a common origin for the Δblock in wolves, grouping all individuals exhibiting the Δblock together and closer to dogs (**Figure 3.2D**).

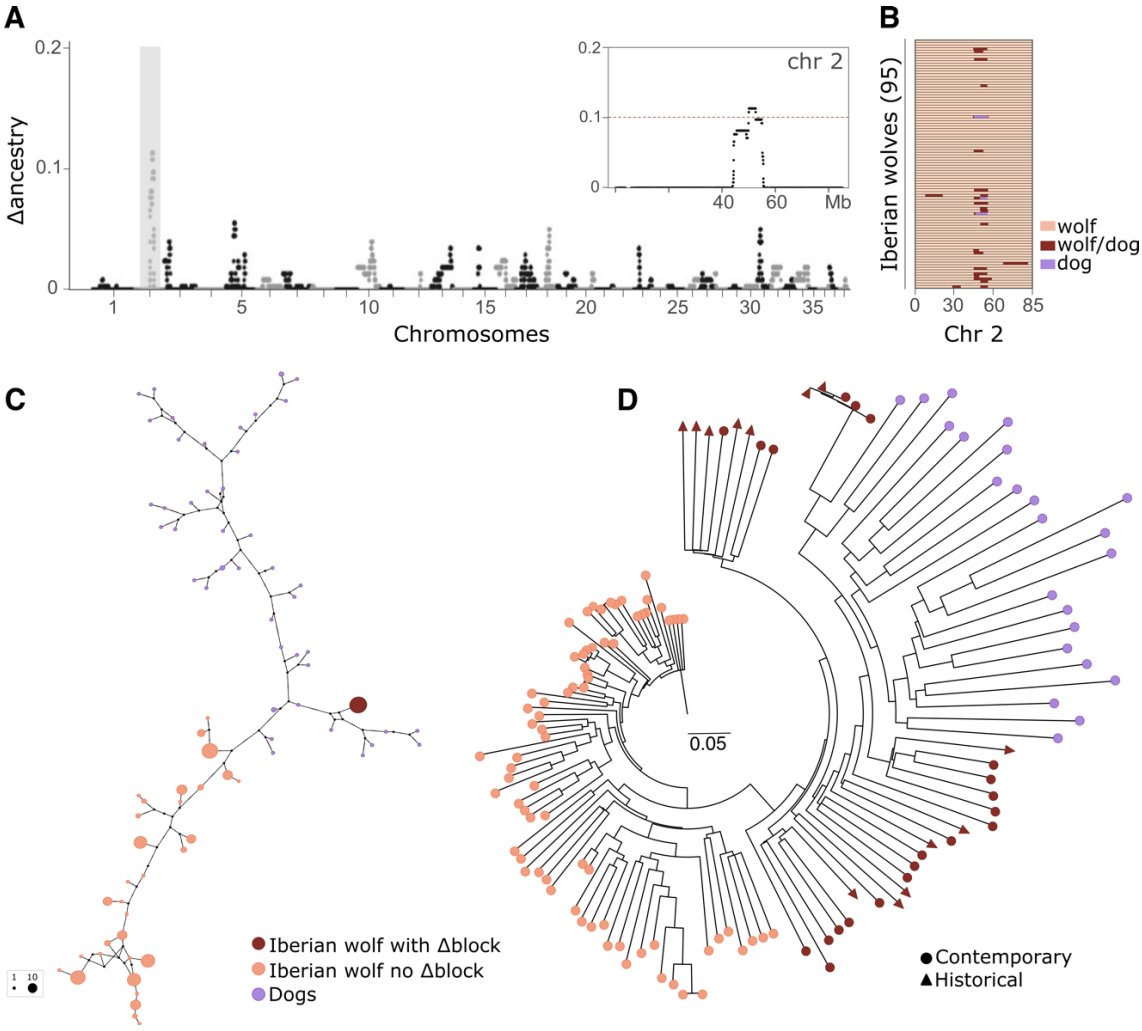


Figure 3.2. Evidence of a dog block in Iberian wolf chromosome 2. (**A**) Δancestry scores indicating the excess of dog ancestry across the Iberian wolf genome (based on ~85K SNPs across 95 contemporary samples). The Δblock on chromosome 2 is shadowed in gray. The inset displays a zoomed-in view of chromosome 2, with the dashed line indicating the cut-off used for detecting outliers. (**B**) Representation of the dog block found in chromosome 2 of Iberian wolves using ELAI local ancestry analysis (considering 10 mixture generations). Each row represents an individual (same order as in Supplementary Table 3.1) and different colors denote the attributed local ancestry: pink for homozygous wolf, red for wolf/dog (heterozygous), and purple for homozygous dog. Most individuals are heterozygous for the dog block. (**C**) Network of Δblock haplotypes (78 SNPs) for contemporary Iberian wolf samples and village dogs from Iberia. The size of the circles is scaled relative to the frequency of each haplotype. (**D**) NJ tree for the Δblock using contemporary and historical (subset > 60K SNPs; see Materials and Methods) samples of Iberian wolves, and village dogs from Iberia. In

panels C and D Iberian wolves with and without the Δblock are colored in red and pink, respectively, and dogs in purple. Contemporary and historical samples are represented by circles and triangles, respectively.

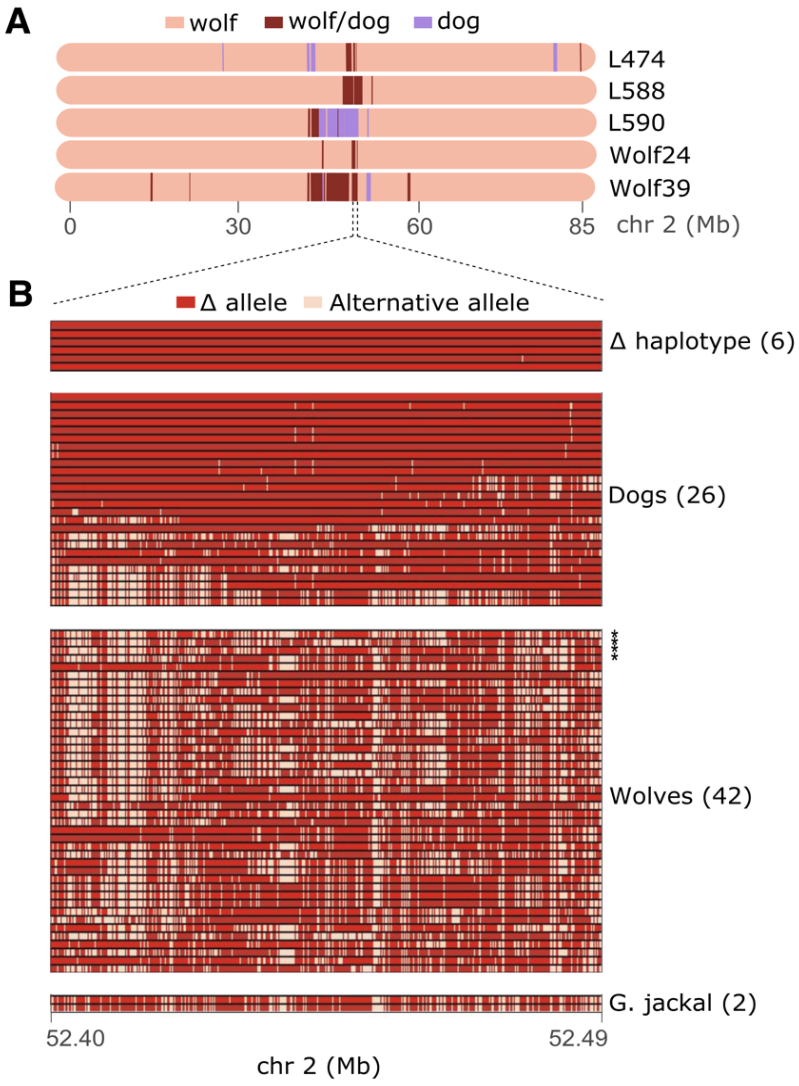


Figure 3.3. Signatures of adaptive introgression using WGS data. (**A**) Representation of the Δblock on chromosome 2 of five Iberian wolf genomes using local ancestry analysis. Different colors denote the attributed local ancestry: pink for homozygous wolf, red for wolf/dog (heterozygous), and purple for homozygous dog. (**B**) Structure of the haplotype (~90 kb; 470 SNPs) found in the Δ100 kb region in common to all Iberian wolves with the Δblock. In the top is the introgressed haplotype found among Iberian wolves; in the middle are the dog haplotypes (village dogs from Iberia appear first) and then the haplotypes found among wolves without the Δblock (Iberian, Eurasian, and North American wolves); in the bottom are the golden jackal haplotypes. Each row depicts a haplotype, and each position is colored according to whether it carries the introgressed allele (Δ in red) or the alternative allele (in pink). Haplotypes marked with an asterisk represent the alternative haplotype of Iberian wolves with the Δblock. Numbers within brackets indicate the total number of haplotypes per group.

3.2.2. Evidence that the dog block is an introgressed variant

To examine whether the ∆block represents introgressed variation from dogs into the Iberian wolf genome, we leveraged the high resolution offered by the WGS data to perform additional tests. We first used the f_d statistic (Martin et al. 2015) to calculate the fraction of introgression by measuring the excess of shared variation between Iberian wolves with the ∆block and dogs in relation to other wolves. Under a scenario of random sorting of ancestral variation, no significant excess of shared variation is expected. The 100 kb window with the highest fraction of shared variation in chromosome 2 was positioned within the ∆block (chr 2: 52.40 – 52.50 Mb, $f_d = 0.77$; hereafter ∆100 kb; **Figure 3.4A**; Supplementary Figure 3.12; Supplementary Table 3.5). Genetic differentiation (F_{ST}) in ∆100 kb between Iberian wolves with the ∆block and dogs was $F_{ST} = 0$, which is significantly lower than the genomic background differentiation (Welch two sample t-test: $p < 0.04$; **Figure 3.4B**; Supplementary Figure 3.13; Supplementary Figure 3.14 for F_{ST} estimates using the 85K SNP dataset). Next, we used HYBRIDCHECK (Ward and van Oosterhout 2016) to scan chromosome 2 haplotypes for regions of high sequence similarity, which identified ∆100 kb as an outlier region between Iberian wolves with the ∆block and dogs (average sequence similarity of 62%; Supplementary Figure 3.15).

Patterns of excess allele sharing such as those we observed can arise in the genome due to other processes rather than introgression (e.g., population structure, ancestral polymorphism, or selection; Eriksson and Manica 2012; Smith and Kronforst 2013; Zheng and Janke 2018). However, only introgression should result in extremely high levels of sequence identity between the donor and the recipient. To assess this, we compared the mean pairwise sequence divergence (d_{XY}) between Iberian wolves with and without the ∆block, Eurasian wolves, and dogs, using all sites within 6 Mb haplotypes surrounding ∆100 kb (Materials and Methods). We found no divergence between Iberian wolves with the ∆block and dogs within the ∆100 kb region ($d_{XY} = 0$), which is significantly lower than the genomic background divergence (t-test: $p < 1.47 \times 10^{-8}$; **Figure 3.4C**; Supplementary Figure 3.16). Furthermore, this divergence between Iberian wolves with the ∆block and dogs in ∆100 kb is also significantly lower compared to that observed at the same position between dogs and wolves not harboring the ∆block ($p < 2.68 \times 10^{-4}$ and $p < 7.12 \times 10^{-5}$, for Iberian wolves without the ∆block and Eurasian wolves, respectively; **Figure 3.4C**; Supplementary Figure 3.16). Concordantly, we found a single haplotype of ~90 kb within the ∆100 kb region (covering 470 SNPs using WGS data) for all Iberian wolves with the ∆block, which was highly divergent from that of other Eurasian wolves (average of 145/470 SNPs; **Figure 3.3B**; Supplementary Table 3.6). The ∆block thus

appears to coalesce to a single haplotype, which suggests it originates from one hybridization event. Importantly, a nearly identical haplotype was also found amongst dogs (**Figure 3.3B**), particularly within dogs in Iberia, from which it diverged by just a single SNP. This level of nucleotide identity is more than 150 times higher than the genome-wide average described for Eurasian wolves and dogs (~1 SNP/580 bp; Lindblad-Toh et al. 2005). Furthermore, the probability of maintaining a 90 kb haplotype without recombination in dogs and Iberian wolves since the time of divergence is extremely small ($p < 0.0003$), even considering the shortest estimate of divergence (14 ka; **Figure 3.4D**). We can thus be confident that this haplotype is of dog origin, rather than an ancestral shared polymorphism.

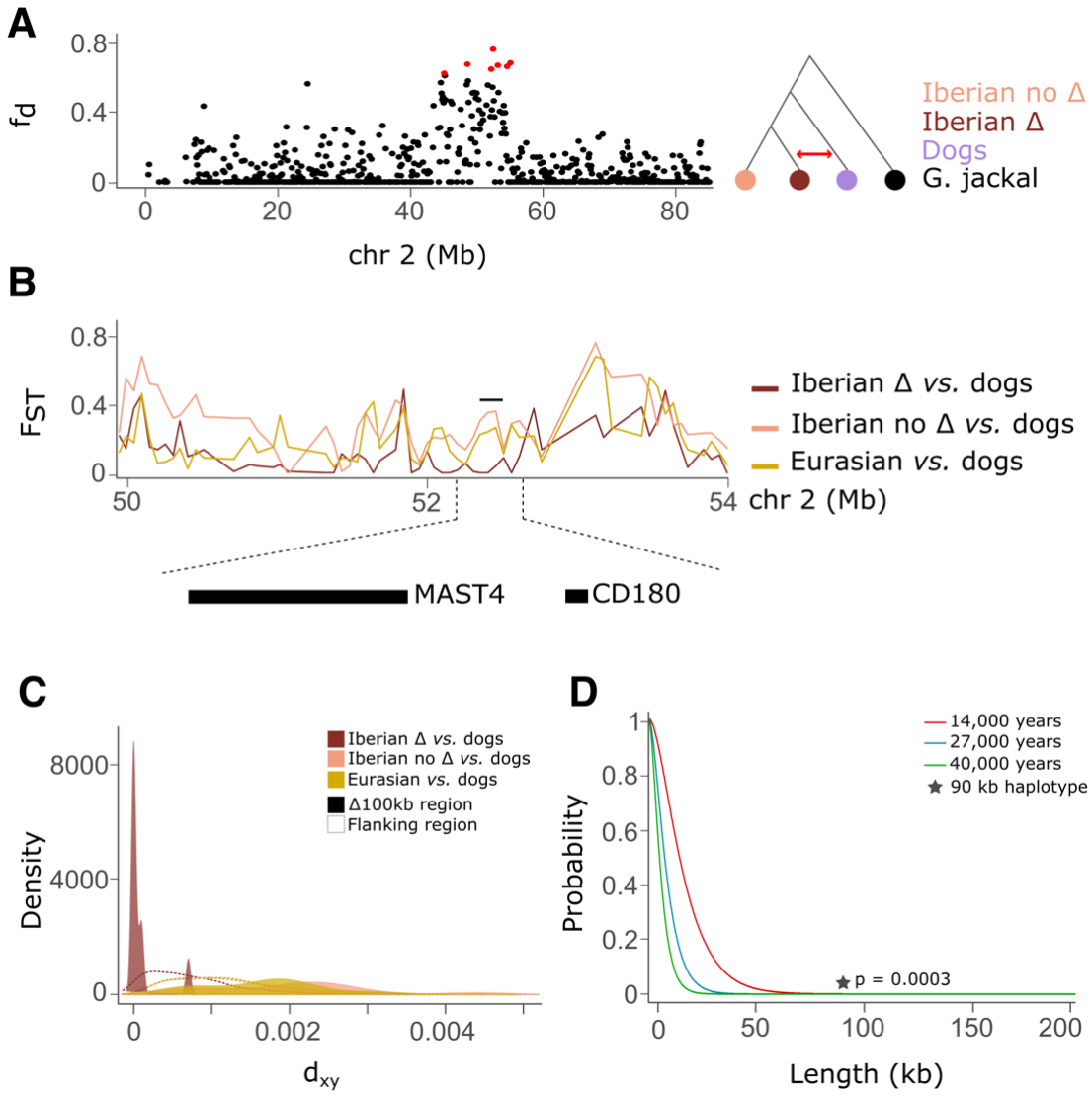


Figure 3.4. Increased allele sharing between Iberian wolves and dogs measured with WGS data. (**A**) Fraction of introgression (f_d) between Iberian wolves with the Δblock and dogs across chromosome 2 in non-overlapping 100 kb windows. Red dots indicate the highest 1% outlier windows. The test was performed following the phylogenetic

conformation illustrated on the right (P1– pink: Iberian wolves without the ∆block, N = 9; P2 – red: Iberian wolves with the ∆block, N = 5; P3 – purple: dogs, N = 13; outgroup – black: golden jackal, N = 1. (**B**) Pairwise F_{ST} between different groups (represented by distinct colors) in non-overlapping 50 kb windows within ∆block. The position of the ∆100 kb region is represented by a black bar, and the protein-coding genes located within the region, and in the vicinity, are illustrated below. (**C**) Distribution of d_{XY} between haplotypes of wolves and dogs within the ∆100 kb region and in the 6 Mb flanking region (each comparison is color coded). (**D**) The probability of maintaining a haplotype of 0-200 kb length due to ancestral shared variation, assuming different divergence times between Eurasian gray wolves and dogs. The gray star indicates the length of the introgressed haplotype (90 kb) and the corresponding p-value.

3.2.3. Simulations of selection support adaptive introgression

To critically test whether the observed pattern of admixture within the ∆100 kb region could be explained by neutral evolution, we performed individual-based forward simulations using SLiM (Haller and Messer 2019) (Materials and Methods). In neutral simulations, few events of hybridization never recovered the level of admixture observed in the ∆100 kb region (Supplementary Figure 3.17). In turn, when multiple events were simulated, we observed that the wolf genomic background rapidly become highly admixed (Supplementary Figure 3.17), which is not compatible with the empirical levels of dog ancestry in the Iberian wolf genome (< 1%). Furthermore, under neutral evolution, the expected amount of dog ancestry in the ∆100 kb region would be the same as in the genomic background (Supplementary Figure 3.18). Neutrality also did not lead to variable levels of dog admixture across genomic regions with distinct levels of recombination, suggesting that low, or suppressed, recombination rate cannot explain the empirical patterns in the ∆100 kb region (Supplementary Figure 3.19). Hence, regardless of the amount of gene flow simulated, neutral models were never a good fit for the differential levels of introgression observed in the ∆100 kb region.

Next, we examined scenarios of adaptive introgression (i.e., when genomic material from the donor increases the fitness of the recipient so that introgressed variants experience positive selection). These simulations successfully recovered the observed empirical patterns of admixture of the introgressed SNPs in the ∆100 kb region (**Figure 3.5**). Still, models of adaptive introgression with selection acting on a single SNP in the block did not lead to the observed patterns of admixture. Only models that simulated selection over multiple SNPs, specifically more than three adaptive SNPs with a combined selection of $\sum s \geq 0.1$, fit the observed level of admixture inside this region (**Figure 3.5**). Such positive selection on multiple SNPs creates linkage that would have avoided the breakdown of the introgressed block (Supplementary Figure 3.20).

Finally, we examined distinct selection models considering different times for the introgression. For a specific number of generations, there was a range of models that

could fit the empirical values of admixture (Supplementary Figure 3.21). However, we found that the overdominance model (i.e., balancing selection) provided the best fit to the empirical allelic frequencies observed within the Δ100 kb region because it avoided the fixation of the introgressed alleles (**Figure 3.5**). Therefore, we propose that balancing selection with heterozygous advantage is maintaining the introgressed dog haplotype in Iberian wolves. Unlike positive selection, the overdominance model is also consistent with the high prevalence of heterozygosity within the introgressed block (F_{IS} Δ100 kb = -0.26; F_{IS} overall chromosome = - 0.02). This pattern is identical to that observed in the dog introgressed K locus haplotype in North American black wolves (Schweizer et al. 2018), which has also been maintained through balancing selection after adaptive introgression.

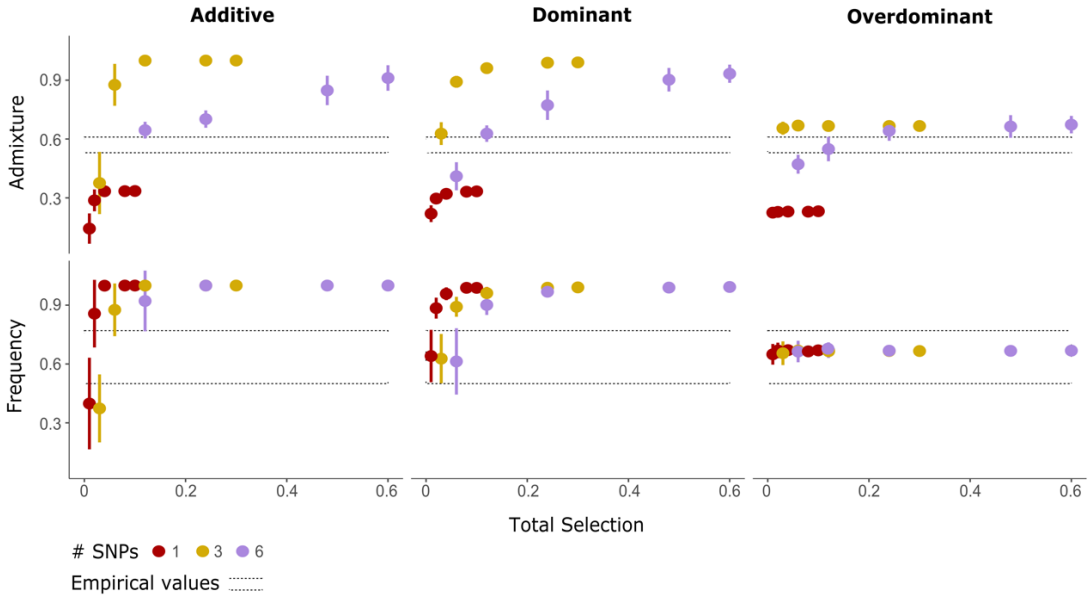


Figure 3.5. Individual-based simulations of genetic introgression. Mean and standard deviation of 100 simulated replicates in SLiM for average admixture proportion (across the entire Δ100 kb region) and mean frequency of introgressed alleles, across adaptive models with one migrant and one migration event at generation 800 (plots with results for distinct timings of introgression can be found in Supplementary Figure 3.21). Empirical estimates of admixture (0.53-0.60; 95% CI of ELAI average estimate) and mean frequency of introgressed alleles (see Supplementary Table 3.6) in Δ100 kb region, are shown in dashed lines. Simulations explored neutral evolution, as well as an additive model (h = 0.5), dominant model (h = 1) and overdominant model (h = 2), with a range of selection coefficients (s = 0-0.1) for 1, 3 or 6 introgressed SNPs (denoted with distinct colors), resulting in a total amount of selection ranging from 0 to 0.6 (x-axis).

3.2.4. Potential functional role of the introgressed haplotype

The observed evidence that a single genomic region with dog ancestry was consistently maintained at high frequency across the genome of Iberian wolves, together with the results from demographic simulations supporting selection, may suggest that the 90 kb introgressed haplotype has adaptive potential in Iberian wolves. Measurements of nucleotide diversity (pi) also supported this hypothesis; in Iberian wolves carrying the introgressed block, pi was significantly lower in the ∆100 kb region than in the genomic background (t-test: $p < 1.87x10^{-6}$; Supplementary Figure 3.22), and was significantly reduced than that estimated in haplotypes of other wolves for the ∆100 kb region ($p < 1.03x10^{-4}$ and $p < 9.43x10^{-5}$, with Iberian wolves without the ∆block and Eurasian wolves, respectively; Supplementary Figure 3.22). The recent bottleneck experienced by the Iberian wolf in the 1970s (Nores and López-Bao, 2022; Clavero et al. 2022) might be responsible for the low genome-wide levels of genetic diversity and increased LD observed, but it is unlikely to explain the reduced nucleotide variation in the introgressed haplotype. Firstly, the patterns observed for the ∆100 kb region were significantly distinct from those in the genomic background. Secondly, if the population history of the Iberian wolf was responsible for such signatures, we would expect to find the same in Iberian wolves without the ∆block, which was not the case. A similar reduction of pi in the ∆100 kb region was found within dogs from the Iberian Peninsula ($p < 4.28x10^{-6}$; Supplementary Figure 3.22), however, the fact that the introgressed haplotype carries more polymorphisms in dogs than in Iberian wolves (**Figure 3.3B**) indicates the haplotype is older in dogs.

The ∆100 kb region is inserted within a MAST4 gene intron positioned within the 5'UTR region. Phylogenetic relationships using SNPs and haplotypes spanning MAST4 (chr 2: 52.18 – 52.63 Mb; covering 2,004 SNPs using WGS data) confirmed that wolves with the ∆block unambiguously clustered with dogs and not with other wolves (Supplementary Figures 3.23B, 3.24), contradicting the expected species tree (Supplementary Figure 3.23A). MAST4 is a potential candidate gene for the behavioral responses observed in Iberian wolves, as it was previously associated with several neurocognitive disorders, such as epilepsy (Steffens et al. 2012), Alzheimer's disease (Shulman et al. 2014), and frontotemporal lobar degeneration (Martins-de-Souza et al. 2012) in several model vertebrate species. In the human brain, MAST4 was found to interact with DLG2, which encodes a protein responsible for activating NMDA glutamatergic signaling genes (Satizabal et al. 2019; GeneCards). A protein-protein interaction network that we built based on the dog reference genome, confirmed this association (Supplementary Figure

3.25). NMDA receptors are known for playing important roles in synaptic plasticity, learning, memory, and in fear conditioning. Decreased expression of these receptors activates adaptive stress responses, such as reduced anxiety and fear memory, in model vertebrate species (Babenko et al. 2016). Most importantly, glutamate genes have been found amongst the few overlapping targets of selection across several domesticates (dog, cat, fox, cattle, sheep, and rabbits; see O'Rourke and Boeckx 2020 for a review).

Although we could not directly assess the functional significance of the introgressed dog variants in MAST4, our findings provide indirect evidence of its potential role in shaping the behavior of Iberian wolves. The possible ability of MAST4 to reduce levels of anxiety and potentiate learning and memory abilities, which are behavioral traits characteristic of domestic dogs, could contribute to explaining the increased tolerance towards humans and their activities observed in Iberian wolves (e.g., feeding on anthropogenic sources of food, lack of dispersal, resting or denning in human-dominated landscapes; Newsome et al. 2016; Sazatornil et al. 2016; Silva et al. 2018). This behavioral shift may confer a fitness advantage to the population, as bolder and dispersing wolves tend to have a lower survival rate by increasing their encounters with humans (Rio-Maior et al. 2019).

3.2.5. The origin and evolution of ∆block

After finding support for adaptive introgression, we aimed to reconstruct the evolutionary history of the introgressed block, trying to assess where and when it originated. We detected a consistent pattern for the geographic prevalence of the ∆block among historical and contemporary Iberian wolf samples. Most individuals with the ∆block originated from Northcentral Iberia (**Figure 3.1B**). Here, 64% of the individuals analyzed have the ∆block, and four out of the five wolves that were homozygous in the ∆block were from this region. This suggests that the hybridization event likely occurred in Northcentral Iberia. The population genetic structure of the Iberian wolf, which may be ancient (~6.2 ka; Silva et al. 2020), could have prevented the ∆block from sweeping across the entire population range, and instead resulted in uneven spatial distribution. To estimate the time since introgression, we calculated the time to the common ancestor (TMRCA) of the ~90 kb haplotype in Iberian wolves with the ∆block using STARTMRCA (Smith et al. 2018). This method assumes the haplotype is under selection and modules LD patterns in the surrounding genomic region using a panel of haplotypes with which the selected haplotype recombined after the introgression event (Materials and Methods). Using two estimates of mutation rate for dogs and wolves, $4x10^{-9}$ (Skoglund et al. 2015) and $4.5x10^{-9}$ (Koch et al. 2019), and the recombination rate reported for

chromosome 2, $8.4x10^{-9}$ (Wong et al. 2010; Campbell et al. 2016), we estimated the TMRCA to be 816 (815-817, 5-95% CI) and 794 (793-795, 5-95% CI) generations, respectively (Supplementary Figure 3.26). Assuming the average across both estimates, and a mean generation time of 4.5 years for wolves (Mech et al. 2016), this places the time since introgression at least 3,624 years ago (3,572-3,676; 5-95% CI). This is a period concordant with the Bronze age when humans started to settle in large communities to farm the lands. The increased use of dogs to protect human and animal communities (Albizuri et al. 2021) might have boosted opportunities for hybridization between wolves and dogs at the time.

Little is known about the rate of wolf-dog hybridization in ancient and historical times. Despite the introgression of the dog ∆block in Iberian wolves, our findings provide robust evidence that the Iberian wolf genome has very low levels of admixture, as it carries an overall genomic proportion of dog ancestry < 1%. Current estimates of hybridization between Iberian wolves and dogs are about 4% (Godinho et al. 2011), but the wolf reproductive behavior of forming stable monogamous pairs for many generations should efficiently limit introgression events. Nevertheless, we showed that Iberian wolves remain clearly differentiated from dogs, ensuring the maintenance of their ecological and biological role in the Iberian Peninsula.

3.3. Conclusions

Our findings provide evidence that hybridization with dogs at least 3,500 years ago resulted in the introgression of a dog genomic block in chromosome 2 of Iberian wolves. The introgressed haplotype has increased in frequency and presents signatures of balancing selection. This selective advantage could have contributed to the persistence of wolves in human-dominated landscapes, such as that found in the Iberian Peninsula. The possible ability of the introgressed MAST4 gene to mediate fear responses and affect memory and learning abilities through its interaction with glutamate genes might hint at how Iberian wolves adapted to live in human-dominated landscapes. Although our findings point in this direction, future studies investigating the functional role of MAST4 will be needed to validate its association with the behavioral phenotypes observed in Iberian wolves, and hence, its adaptive potential. The association between the wolf genome after dog introgression and its behavioral adaptation in human-dominated landscapes could also contribute to explaining observed behaviors of other wild carnivores evolving in similar landscapes (Benazzo et al. 2017).

3.4. Materials and Methods

Population genomics approach: contemporary samples

3.4.1. Canine HD SNP BeadChip data

We generated genome-wide SNP data for 95 Iberian wolves, five Eurasian wolves and 62 dogs (Supplementary Table 3.1). Iberian wolf samples comprised muscle and blood samples and were collected across the full range distribution in Iberian Peninsula (**Figure 3.1B**), between 1996 to 2017, mainly from road kills. Our sampling approach involved a great effort to cover the full wolf range in the Iberian Peninsula to be representative of the entire population. Dog samples comprised muscle, blood and buccal swabs and were donated by local shelters and collaborators. Eurasian wolf muscle samples were donated by collaborators. No animals were killed for this study. Total genomic DNA was extracted using the QIAGEN DNeasy Blood & Tissue Kit (Qiagen) and quantified on the Qubit DNA quantification system (Thermo Fisher Scientific) using Qubit broad range assay reagents, following the manufacturer's protocol. DNA concentration across all samples was normalized to 50 ng/µl to be genotyped for approximately 170,000 genome-wide SNPs, using the Canine HD BeadChip microarray (Illumina, Inc., San Diego, CA, USA). Thereafter, the genotype calling was performed using GenomeStudio software (Illumina) following Illumina's recommendations. In addition, we added to our dataset two datasets genotyped under the same technique, namely, 58 dogs, including 30 pure breeds, from the LUPA project (Lequarré et al. 2011; Vaysse et al. 2011) and 50 European wolves from Stronen et al. 2015 (Supplementary Table 3.1). The three datasets were merged using PLINK v.1.9 (Purcell et al. 2007), after converting the SNP coordinates of the two additional datasets to the CanFam3.1 dog genome assembly using the LIFTOVER tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver). The distribution and density of SNPs along the genome were verified using the R/Bioconductor package karyoploteR (Gel and Serra 2017) in R (R Development Core Team 2017). The final dataset comprised 95 Iberian wolves, 55 Eurasian wolves, and 120 dogs.

Closely related individuals (identity-by-descent > 0.5) were excluded and only autosomal SNPs without multiple positions were maintained. Filters for high call rates per sample (> 0.95) and per SNP (> 0.98) were applied, and loci with minor allele frequency (MAF) below 0.01 were removed, resulting in a final dataset of ~85,000 SNPs. A light pruning on linkage disequilibrium (LD) was also applied for specific analyses, using $r^2 \geq 0.8$, with

a sliding-window size of 50 SNPs, shifted and recalculated every 10 SNPs (LD-pruned dataset resulted in ~79,000 SNPs). All filters were applied in PLINK.

3.4.2. Genetic diversity analysis

Proportion of heterozygous sites, LD decay curves, and runs of homozygosity (ROH) were determined across the groups (Iberian wolves, Eurasian wolves, village dogs, and purebred dogs). For heterozygosity and LD analysis, sample sizes were adjusted according to the minimum number of individuals within groups (N = 23) to avoid bias. Individuals were randomly selected. Estimates of LD decay were calculated independently for each group, based on pairwise r^2 values between all pairs of SNPs within a 1 Mb window. Values were averaged in non-overlapping 1 kb windows for plotting. ROH were considered segments longer than 300 kb spanning at least 50 SNPs and containing a maximum of 1 heterozygous SNP (Ceballos et al. 2018). Options controlling for SNP density, gap, sliding-window length, missing calls, and window threshold were kept with PLINK defaults. Inverse cumulative average counts of ROH per individual were calculated for every 1 Mb non-overlapping window until the 60 Mb for plotting. All the analyses were performed using a dataset only filtered for missing rates (~86,000 SNPs), not MAF applied, in PLINK. Plots were constructed in R.

3.4.3. Global ancestry estimation

Global ancestry proportions were firstly explored by a Principal Component Analysis (PCA) using PLINK. We then used the maximum likelihood model implemented in ADMIXTURE v.1.3 to estimate the ancestry proportions (q) in K = 2-5 different clusters. ADMIXTURE was run with the entire LD-pruned dataset in 2000 interactions and using a cross-validation procedure of 10-fold (Alexander and Lange 2011). To avoid individuals resulting from recent events of hybridization between wolves and dogs, we set a minimum assignment threshold to consider non-admixed individuals of $q \geq 0.90$ for wolves and $q \geq 0.95$ for dogs, to the wolf and dog cluster, respectively. To achieve this, we performed two extra runs of ADMIXTURE only for K = 2 between wolves and dogs, using either the Iberian or the Eurasian wolves. Because we were interested exclusively in historical signs of dog introgression in Iberian wolf genome, including recent hybrids in the following analysis could introduce bias and inflate the proportion of introgression.

3.4.4. Local ancestry analysis

We used a set of local ancestry methods to infer the identity of single chromosomal blocks within Iberian wolf genome (Liu et al. 2013; Geza et al. 2019). We tested two different methods: LAMP-ANC v.2.5 and ELAI. LAMP-ANC is a non-LD-based method that estimates within windows the most likely ancestry per SNP given the reference allele frequencies (Sankararaman et al. 2008). LAMP-ANC was run considering as admixed population the Iberian wolves (N = 95) and, as reference populations, both Eurasian wolves (N = 55) and dogs (N = 120). We assumed 10 generations since the beginning of admixture, considering a generation time of 4.5 years. This method was not designed to account for older admixture events, as more generations result in overestimation of admixture proportions, performing even poorly when the reference populations are closely related (Liu et al. 2013). Additionally, we used a recombination rate of $9.7x10^{-9}$ (Wong et al. 2010; Campbell et al. 2016), a mixture proportion of 0.99:0.01 (based on global ancestry estimates) and a LD cut-off of $r^2 > 0.1$, because LAMP-ANC assumes unlinked markers. We then ran ELAI that implements a two-layer hidden Markov model (HMM) to infer local ancestry without needing phased data to run, neither a prior definition of window size (Guan 2014). An additional feature of ELAI is the capacity to infer an unsampled reference population based on the allele frequencies of the admixed population (Seixas et al. 2018). Because we did not have a true Iberian wolf reference population, we ran ELAI considering three reference populations (Eurasian wolves, dogs and an unsampled population) by setting the number of upper-layer clusters to -C 3 and lower-layer to -c 15 (5x the number of -C, as recommended). To run ELAI, we used the filtered dataset (~85K SNPs), performed 20 expectation maximization (EM) steps, and tested several mixture generations, starting with 10 generations, followed by 100 and then in intervals of 100 up to 1000.

3.4.5. SNP-specific Δstatistics and validation of introgression

Because local ancestry methods do not attribute statistical significance to each detected block, we calculated the SNP-specific delta (Δ) statistics to find regions with a significantly higher proportion of dog introgression across individuals. Δstatistics for parental population B at marker *m* is defined as $q_B^m - q_B$, where q_B^m is the mean ancestry at the marker *m* over all introgressed individuals and q_B represents the mean genome-wide ancestry over all individuals. Positive values of Δancestry represent core hotspot regions for introgression from population B across the genome of all individuals. We calculated Δancestry for all the chromosomes using local estimates of dog ancestry

determined by ELAI. We then used three standard deviations (3SD) from the mean chromosome Δancestry, as a threshold to validate outlier regions. In addition, we only considered outlier regions those that were also detected by LAMP-ANC in the previous analysis. This analysis identified as outlier region a block on chromosome 2 of Iberian wolves (Δblock), spanning 78 SNPs.

To investigate the origin of Δblock, a NJ tree was constructed using SNPs within this region, based on pairwise genetic distances between Iberian wolves and village dogs from Iberia (most abundant dogs across the wolf territory). The distances were estimated with TASSEL 5 (Bradbury et al. 2007) as 1 – IBS (identity by state), with IBS defined as the probability that alleles drawn at random from two individuals at the same locus are the same. The NJ was built also with TASSEL 5 and drawn with FIGTREE v.1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/). We also performed a PCA with PLINK, using the same set of 78 SNPs, to investigate how individuals clustered. A median-joining haplotype network was also constructed for Δblock using POPART (Leigh and Bryant 2015) with reconstructed haplotypes of Iberian wolves and Iberian village dogs in PHASE v.2.1 (Stephens and Donnelly 2003), with the *LocusType* option set to *S* for bi-allelic SNPs.

Levels of genetic differentiation (F_{ST}) between Iberian wolves with and without the Δblock and between these two groups with dogs were determined at genome-wide level, at chromosome 2 (3091 SNPs) and within the Δblock with VCFtools v.0.1.15 (Danecek et al. 2011) according to Weir and Cockerham's formula (Weir and Cockerham 1984). F_{ST} values were displayed in Manhattan plots, using the R package *qqman* (Turner 2014).

Population genomics approach: historical samples

3.4.6. Historical samples and in-solution capture enrichment

To investigate the prevalence of the Δblock on chromosome 2 of Iberian wolves in the past, we analyzed a set of 48 historical samples collected across the former range distribution in Iberian Peninsula and covering a period spanning from 1912 to 2005 (**Figure 3.1B**; Supplementary Table 3.4). At the exception of one sample from 2005, all the samples were collected before 1990. We used the criteria defined by Raxworthy and Smith 2021, that restrict historical DNA to that fortuitously obtained from traditional museum voucher specimens not intended to serve as source for DNA, to classify samples as historical. We generated and used these samples in a parallel study (Pacheco, Lobo et al. 2022; Chapter II). Briefly, we performed an in-solution target

capture enrichment to obtain a total of 100,000 genome-wide SNPs, whose positions were defined based on the coordinates available on the Canine HD BeadChip. By following this approach, the historical dataset is compatible with any dataset genotyped using the SNP chip technology. Libraries constructed with historical DNA were pooled and sequenced on the Illumina Hiseq X instrument, using a PE 150 bp chemistry (full details about the methodology – bait design, capture experiment, processing of sequencing reads and genotyping – can be found in the original publication). Only genotypes with read depth above four were retained. The resulting genotypes were then validated using two control contemporary samples that were subjected to target enrichment and were also genotyped with the Canine HD BeadChip. In both cases, concordance rates between genotypes were > 99%. Only samples with more than 20,000 SNPs and with missing data < 10% were kept, resulting in a final dataset with ~23,000 SNPs. The historical samples were merged with the contemporary dataset of Iberian wolves (N = 95) and dogs (N = 120), and filters for call rates per sample (> 0.85) and per SNP (> 0.90) were applied, and loci with MAF < 0.01 were removed. This resulted in a combined dataset of ~18,000 SNPs. All the previous steps were performed in PLINK.

3.4.7. Local ancestry analysis in historical samples

The introgression analysis in historical samples was performed using the local ancestry method LAMP-ANC as previously described for the SNP chip dataset. Although LAMP-ANC is not as accurate as LD-based methods, we choose this software because it is the one recommended when the genome-wide SNP density is not too high (Sankararaman et al. 2008; Pasaniuc et al. 2009). We ran the analysis using the merged dataset composed by historical and contemporary samples of Iberian wolves and dogs, with ~18K SNPs. As reference populations, we used all dogs (N = 120) and the contemporary Iberian wolves without any detected signs of introgression on chromosome 2 (N = 70). As admixed population, all the samples from the historical dataset were considered. LAMP-ANC was run under the same parameters as described before. To ensure that the analysis was not biased by the limited number of loci in this dataset (chr 2 = 623 SNPs), we subsampled the historical dataset and extracted only Iberian wolf samples carrying the ∆block on chromosome 2 with at least 60,000 genome-wide SNPs (N = 13). This new dataset was merged with the contemporary samples from Iberian wolves and dogs, and all the previously described filters were applied, resulting in a combined dataset of ~48,000 SNPs. We further ran LAMP-ANC a second time, using as admixed population

these 13 Iberian wolves and maintaining all the previous parameters. For both LAMP-ANC runs outlier regions were identify using Δancestry. A NJ tree with all contemporary Iberian wolves, the subset of historical Iberian wolves (> 60K SNPs) and village dogs from Iberia, was constructed using SNPs within the Δblock (69 SNPs). Genetic distances and the NJ tree were estimated with TASSEL 5 and the tree was drawn with FIGTREE.

Whole-genome sequencing approach

3.4.8. Whole-genome sequences and SNP calling

To have the high resolution needed to detect small genomic dog blocks and further confirm the presence of the Δblock in chromosome 2 of Iberian wolves, we generated whole-genomes of 12 of the Iberian wolves used in the SNP chip dataset (**Figure 3.1B**; Supplementary Table 3.2). DNA extraction was performed using a Thermo Scientific™ KingFisher™ instrument according to the manufacturer's guidelines and resulting DNA fragments were fragmented in the Covaris LE220 plus Focused-ultrasonicator with the parameters set for getting 350 bp fragment length. Fragmented DNA was converted into BGISeq compatible double-stranded DNA sequencing libraries using the "single-tube" library building protocol BEST as initially described in Carøe et al. 2018, and modified for BGISeq technology following Mak et al. 2017. After that, libraries were pooled and sequenced on a DNBSEQ-G400 instrument, using a PE 150 bp chemistry to an average depth of 13.5x per sample at BGI-Europe, Copenhagen, Denmark. Sequence reads were processed and aligned using the BAM pipeline of PALEOMIX v.1.2.13.3 (Schubert et al. 2014). In the first step of the pipeline, low quality and N bases were trimmed from the reads, adapter sequences were removed, and overlapping read pairs were collapsed using AdapterRemoval v.2.2.0 (Schubert et al. 2016) accordingly with default parameters. Reads were mapped to the dog reference genome CanFam3.1 (Lindblad-Toh et al. 2005) using BWA v.0.7.16a (Li and Durbin 2009) with backtrack algorithm, considering a minimum base quality of 0 to ensure that all the reads were retained in the process. Reads were subsequently filtered for PCR duplicates using PICARD MarkDuplicates (https://broadinstitute.github.io/picard/). In the last step of the pipeline, local realignment around indels was performed with IndelRealigner module of the GATK v.3.8 (McKenna et al. 2010). Before SNP calling, we recalibrated the base quality scores (BQSR) using BaseRecalibrator/ApplyBQSR modules of the GATK v.4.1.8.1. Genotypes were then called for each sample using BCFtools v.1.10.2 (Li 2011) mpileup/call -m tools with minimum Phred-scaled thresholds of 20 for base quality (BQ) and read mapping quality (MQ), and all the sites were emitted in the output. At the end, the genotypes were filtered

using VCFtools to keep only bi-allelic and autosomal SNPs (indels were excluded) that have passed the previous filters (filter flag *PASS; --remove-filtered-all*), have less than 20% of missing data across the 12 samples (*--max-missing 0.8*) and were supported by at least four reads (*--min-meanDP 4*). To this dataset, we added a set of 27 publicly available whole-genome sequences from several worldwide canids (13 gray wolves, including three Iberian wolves; 13 dogs and one golden jackal) (Supplementary Table 3.2) (Auton et al. 2013; Wang et al. 2013; Freedman et al. 2014; Zhang et al. 2014; Fan et al. 2016; Marsden et al. 2016; Botigué et al. 2017; Plassais et al. 2019). The genotypes from the 27 canids were extracted from a VCF file containing 91 million variants and 722 canid genomes created by Plassais et al. 2019, available on NCBI (accession number: PRJNA448733). This VCF file was then filtered as described above and merged with the VCF file containing the genotypes of the 12 Iberian wolves, using BCFtools merge tool. In the final dataset we kept sites with less than 10% missing (*--max-missing 0.9*) and with a minor allele frequency greater or equal to 1% (*--maf 0.01*), resulting in 10.8 million SNPs ("variants only" VCF). Finally, we estimated missing and depth statistics across all individuals using VCFtools and verified that all individuals had less than 10% missing data (except the dog Boxer genome with 12%) and an average coverage of 17.6x across the 10.8 million SNPs in all samples.

3.4.9. Introgression analysis using WGS data

We first explored the genome-wide mixture proportion between wolves and dogs in K = 2-5 clusters using ADMIXTURE as formerly described for the SNP chip dataset (see results and individuals that were excluded based on dog ancestry proportion in Supplementary Text and Supplementary Figure 3.4). To validate the ∆block on chromosome 2, we ran ELAI considering three reference populations (Eurasian wolves, N = 8; dogs, N = 13; and an unsampled population) and all the Iberian wolves as admixed (N = 15), following the previously described parameters. We also calculated the fraction of introgression using f_d statistics, which is an improved and not biased derivate of the *f*-statistics. We computed f_d along chromosome 2 (~386,000 SNPs), testing non-overlapping windows of 50 kb and 100 kb, with at least 100 SNPs, using the *ABBABABAwindows.py* script from genomics_general package (available at https://github.com/simonhmartin/genomics_general). Tests were performed under the following phylogenetic conformation: case 1) P1 – Iberian wolves without the ∆block, N = 9; P2 – Iberian wolves with the ∆block, N = 5, defined by ELAI; P3 – dogs, N = 13; O – golden jackal, N = 1; and case 2) P1 – Eurasian wolves, N = 8; P2 – Iberian wolves with

the ∆block; P3 – dogs; O – golden jackal. We tested different groups as P1 to investigate if regions overlapping the ∆block would be consistently identified as outlier windows. f_d statistics will measure the excess shared variation between P3 and P2 that is not shared with P1, excluding the hypothesis of incomplete lineage sorting. We applied a conservative threshold to consider significantly introgressed regions those above the 99th percentile. F_{ST} was also estimated between the genotypes of Iberian wolves with and without the ∆block, Eurasian wolves and dogs, along chromosome 2 in non-overlapping windows of 50 kb and 100 kb, using VCFtools. Levels of differentiation between the ∆block and the genomic background (defined as the entire chromosome 2 excluding the introgressed block) were compared using a Welch two sample t-test in R.

3.4.10. Genetic sequence similarity

To identify the haplotype(s) of the ∆block and explore levels of genetic similarity between introgressed wolves and dogs, we first phased the genotypes in chromosome 2 with BEAGLE v.5.1 (Browning and Browning 2007) with the aid of dog genetic map (cM/Mb). We further used HYBRIDCHECK v.1.0 to screen the haplotypes for blocks of high sequence similarity in 1 kb windows. HYBRIDCHECK calculates pairwise sequence similarity analysis in sets of triplets, using a sliding window approach to determine the number and length of introgressed blocks. As input, we created pseudo sequences of chromosome 2 by concatenating the phased SNPs for each individual. Pairwise comparisons were then performed in sets of triplets for each of the five individuals carrying the ∆block: Iberian wolf with the ∆block – Iberian wolf without the ∆block – Portuguese village dog. We choose L409 as an individual without the ∆block based on ELAI results and used it across all comparisons. Blocks with significantly high sequence similarity between Iberian wolves with the ∆block and dogs were reported.

3.4.11. Common introgressed block and shared haplotype

Positions that concordantly displayed the highest levels of dog allele dosage across all five Iberian wolves with the ∆block were recorded. We perform this analysis looking to ELAI results, f_d and F_{ST}1% outlier regions, and HYBRIDCHECK outlier blocks. A genomic window of 100 kb (chr 2: 52.40 – 52.50 Mb, hereafter ∆100 kb) stand as a common outlier in the five individuals. Using TASSEL, we manually scanned the SNPs within this region to find signs of allelic shared variation between dogs and Iberian wolves with the ∆block that were not observed with other wolves (e.g., major allele in dogs was only present in Iberian wolves with the ∆block). To distinguish between ancestral shared

polymorphism and introgression, we estimate the mean pairwise sequence divergence (d_{XY}) between haplotypes of Iberian wolves with and without the ∆block, Eurasian wolves, and dogs from Iberian Peninsula. The latter were chosen to increase specificity in the analysis, as these dogs should represent the closest source for the donor of introgression. Because d_{XY} analysis requires invariant sites to be present in the dataset, we considered that all the genotyped invariant sites that were not present in the combined VCF ("variants only") were homozygous for the reference. To do this, we used the *--missing-to-ref* option from BCFtools. Then, we extracted all the sites in a 6 Mb region surrounding the mid position of ∆100 kb (chr 2: 49.44-55.44 Mb), which has resulted in 5.4 million sites. Flanking regions were treated as the genomic background. Haplotypes within this 6 Mb region were reconstructed using BEAGLE v.5.1 and d_{XY} was calculated using the *popgenWindows.py* script from genomics_general package in 10 kb non-overlapping windows with at least 100 sites, and with the -f flag set to *haplo* to deal with haploid data. We used Welch's t-test to compare sequence divergence between the genomic background and the ∆100 kb region, and calculated density distributions using the R package *sm* (Bowman and Azzalini, 2021), both in R.

According to the CanFam3.1 Ensembl annotation, the ∆100 kb region is inserted within the MAST4 gene (gene position: chr 2: 52.18 – 52.63 Mb). To explore the phylogenetic relationships between all individuals for MAST4, we constructed a maximum likelihood tree using SNPs spanning the gene (2,004 SNPs) with DNAML in the PHYLIP package (Felsenstein 1989). SNPs were concatenated to generate pseudo fasta sequences, which were posteriorly converted to phylip format with *convert_fasta_to_phylip.py* script from SNPhylo package (Lee et al. 2014). In addition, we also constructed population trees at chromosome 2 level and within MAST4 in TREEMIX v.1.13 (Pickrell and Pritchard 2012), based on the allele frequencies of six groups: Iberian wolves without the ∆block, N = 9; Iberian wolves with the ∆block, N = 5; Eurasian wolves, N = 8; North American wolves, N = 2; dogs, N = 13; and one golden jackal set as the root. The allelic frequencies were estimated with PLINK, and the output file was converted into the TREEMIX input with *plink2treemix.py* script, provided in TREEMIX. Maximum likelihood trees were constructed running TREEMIX in blocks of 100 SNPs with 200 bootstrap replicates and considering no migration events. TREEMIX was run 100 times and the trees were summarized with SumTrees of the DendroPy library (Sukumaran and Holder 2010) and visualized with FIGTREE. We also built a median-joining network with the haplotypes of MAST4, using POPART. Additionally, to further explore the functional relationships around MAST4, we performed a protein-protein interaction analysis in STRING v. 11

(Szklarczyk et al. 2019) (https://string-db.org/), considering a minimum required interaction score of 0.4.

3.4.12. Probability of maintaining the 90 kb haplotype from shared ancestral variation

After identifying a common haplotype of ~90 kb to all Iberian wolves with the ∆block and dogs, we calculated the probability of this haplotype being maintained due to ancestral polymorphism. We applied the method used by Huerta-Sánchez et al. 2014 and Miao et al. 2017, which follows a Gamma distribution:

$$p = 1 - pgamma\,(m, shape = 2, rate = 1/L)$$

where L is the expected length of a shared ancestral haplotype, and is defined by $L = 1/(r \times t)$, where r is the recombination rate per generation per bp and t is the time since divergence between wolves and dogs. We considered a divergence range of 40 to 14 ka (Skoglund et al. 2015; Fan et al. 2016; Silva et al. 2020; Perri et al. 2021; Bergström et al. 2022), and the recombination rate estimated for chromosome 2 of 8.4×10^{-9}. The p-value was calculated using the shortest estimate of divergence (14,000 years) to be conservative.

3.4.13. Testing for adaptive introgression and SLiM simulations

To test for the adaptive potential of the introgression, we implemented individual-based forward simulations in SLIM 3 to investigate if the introgression of the ∆block could have occurred under neutrality, and to estimate the amount of selection required to recover the level of admixture observed in the empirical data. We simulated the entire chromosome 2 of the dog reference genome CanFam3.1 and its underlaying recombination map to capture realistic dynamics and the genomic background. Our model simulates two divergent lineages (dog and wolf) that experience migration events with a range number of migrants. We simulated the wolf lineage with an effective population size of Ne = 3,000 according to Silva et al. 2020. The dog lineage possessed unique SNPs every 10 kb across the entire chromosome, to track their ancestry when introgressed into the wolf genome. We first simulated a neutral model across a range of possibilities for the timing and amount of introgression: number of migration events (single, one every 100 generations or one every generation) and number of migrants (one, 1%, 2% or 4% of the wolf Ne). For each combination of parameters simulations ran for 1,200 generations, starting migration at generation 10, and 100 replicates were

performed. Next, we simulated different adaptive scenarios to investigate whether non-neutral models were a better fit to the empirical data. Given that the neutral scenarios of higher and more frequent migration consistently led to a signal of higher chromosome-wide admixture that the one observed in the empirical data, we performed the adaptive scenario only with the parameters that recovered realistic dynamics in the neutral model (i.e., one migration event with a single migrant). We simulated adaptive SNPs inside ∆100 kb in the dog genome with positively selected alleles (hereafter "introgressed alleles"). To test the effect of linkage and the genomic architecture of the introgressed alleles, we simulated a range (1-6) of SNPs inside ∆100 kb. To investigate the magnitude of selection required for introgression to take place, we varied the selection (*s*) and the number of adaptive SNPs and calculate the combined selection by summing selection coefficients across adaptive SNPs ($\sum s$). To investigate the mode of selection, we varied the dominance coefficients (additive h = 0.5, dominant h = 1 or overdominance h = 2). We calculated the true admixture proportion of ancestry tracking SNPs and the frequency of introgressed SNPs inside ∆100 kb, 800 generations after migration (the estimated average time for introgression).

In addition, we compared levels of nucleotide diversity (pi) between the ∆100 kb region and the genomic background, using the haplotypes of the 6 Mb region surrounding ∆100 kb as input for *popgenWindows.py* script. The analysis was performed as previously described for the d_{XY} analysis. Significant differences were assessed through a Welch's t-test.

3.4.14. Estimating time of dog introgression

We dated time since introgression of the ∆haplotype in Iberian wolves using a Markov chain Monte Carlo (MCMC) approach implemented in STARTMRCA. This method leverages the decay of LD between the selected allele and nearby sites, as well as the appearance of new mutations in ∆100 kb. This method calculates the time to the common ancestor (TMRCA) assuming the haplotype has been subjected to positive selection. It requires a panel of reference haplotypes without the selected allele with which the selected haplotype has recombined following introgression. We used the haplotypes of Iberian wolves without the introgression to represent the reference haplotype. We centered the analysis on chr 2: 52,449,886 (center position of ∆100 kb) and included 1 Mb of upstream and downstream sequence, as recommended. This resulted in a genomic region comprising 1.8 M sites. We assumed a constant recombination rate of 8.4×10^{-9} for chromosome 2 and tested two of the most widely used mutation rates for

dogs and wolves, $4x10^{-9}$ and $4.5x10^{-9}$. We ran 10 independent MCMC chains, each with 50,000 iterations with a standard deviation of 10 for the proposal distribution. The final 10,000 iterations of each replicate were used to generate posterior TMRCA distributions. Estimates of TMRCA were converted to time in years using a generation time of 4.5 years per generation.

3.5. References

Albizuri S, Grandal-d'Anglade A, Maroto J, Oliva M, Rodríguez A, Terrats N, Palomo A, López-Cachero FJ. 2021. Dogs that Ate Plants: Changes in the Canine Diet During the Late Bronze Age and the First Iron Age in the Northeast Iberian Peninsula. J. World Prehistory 34:75–119.

Alexander DH, Lange K. 2011. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics 12:246.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19:1655–1664.

Allendorf FW, Leary RF, Spruell P, Wenburg JK. 2001. The problems with hybrids: setting conservation guidelines. Trends Ecol. Evol. 16:613–622.

Anderson TM, VonHoldt BM, Candille SI, Musiani M, Greco C, Stahler DR, Smith DW, Padhukasahasram B, Randi E, Leonard JA, et al. 2009. Molecular and Evolutionary History of Melanism in North American Gray Wolves. Science 323:1339–1343.

Auton A, Rui Li Y, Kidd J, Oliveira K, Nadel J, Holloway JK, Hayward JJ, Cohen PE, Greally JM, Wang J, et al. 2013. Genetic Recombination Is Targeted towards Gene Promoter Regions in Dogs. PLoS Genet. 9:e1003984.

Babenko VN, Bragin AO, Spitsina AM, Chadaeva IV., Galieva ER, Orlova GV., Medvedeva IV., Orlov YL. 2016. Analysis of differential gene expression by RNA-seq data in brain areas of laboratory animals. J. Integr. Bioinform. 13:1–15.

Benazzo A, Trucchi E, Cahill JA, Maisano Delser P, Mona S, Fumagalli M, Bunnefeld L, Cornetti L, Ghirotto S, Girardi M, et al. 2017. Survival and divergence in a small group: The extraordinary genomic history of the endangered Apennine brown bear stragglers. Proc. Natl. Acad. Sci. 114:E9589–E9597.

Bergström A, Stanton DWG, Taron UH, Frantz L, Sinding M-HS, Ersmark E, Pfrengle S, Cassatt-Johnstone M, Lebrasseur O, Girdland-Flink L, et al. 2022. Grey wolf genomic history reveals a dual ancestry of dogs. Nature 607:313-320.

Botigué LR, Song S, Scheu A, Gopalan S, Pendleton AL, Oetjens M, Taravella AM, Seregély T, Zeeb-Lanz A, Arbogast R-M, et al. 2017. Ancient European dog genomes reveal continuity since the Early Neolithic. Nat. Commun. 8:16082.

Bowman A, Azzalini A. 2021. sm R package: Smothing Methods for Nonparametric Regression and Density Estimation.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. 2007. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635.

Browning SR, Browning BL. 2007. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. Am. J. Hum. Genet. 81:1084–1097.

Campbell CL, Bhérer C, Morrow BE, Boyko AR, Auton A. 2016. A Pedigree-Based Map of Recombination in the Domestic Dog Genome. G3 Genes|Genomes|Genetics 6:3517–3524.

Cao YH, Xu SS, Shen M, Chen ZH, Gao L, Lv FH, Xie XL, Wang XH, Yang H, Liu CB, et al. 2021. Historical Introgression from Wild Relatives Enhanced Climatic Adaptation and Resistance to Pneumonia in Sheep. Mol. Biol. Evol. 38:838–855.

Carøe C, Gopalakrishnan S, Vinner L, Mak SST, Sinding MHS, Samaniego JA, Wales N, Sicheritz-Pontén T, Gilbert MTP. 2018. Single-tube library preparation for degraded DNA. Methods Ecol. Evol. 9:410–419.

Ceballos FC, Hazelhurst S, Ramsay M. 2018. Assessing runs of Homozygosity: a comparison of SNP Array and whole genome sequence low coverage data. BMC Genomics 19:106.

Chapron G, Kaczensky P, Linnell JDC, von Arx M, Huber D, Andrén H, López-Bao JV, Adamec M, Álvares F, Anders O, et al. 2014. Recovery of large carnivores in Europe's modern human-dominated landscapes. Science 346:1517–1519.

Chen N, Cai Y, Chen Q, Li R, Wang K, Huang Y, Hu S, Huang S, Zhang H, Zheng Z, et al. 2018. Whole-genome resequencing reveals world-wide ancestry and adaptive introgression events of domesticated cattle in East Asia. Nat. Commun. 9:2337.

Clavero M, García-Reyes A, Fernández-Gil A, Revilla E, Fernández N. 2022. Where wolves were: setting historical baselines for wolf recovery in Spain. Anim. Conserv.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

Dennehy E, Llaneza L, López-Bao JV. 2021. Contrasting wolf responses to different paved roads and traffic volume levels. Biodivers. Conserv. 30:3133-3150.

Eriksson A, Manica A. 2012. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. Proc. Natl. Acad. Sci. 109:13956–13960.

Fan Z, Silva P, Gronau I, Wang S, Armero AS, Schweizer RM, Ramirez O, Pollinger J, Galaverni M, Ortega Del-Vecchyo D, et al. 2016. Worldwide patterns of genomic variation and admixture in gray wolves. Genome Res. 26:163–173.

Felsenstein J. 1989. PHYLIP: Phylogeny Inference Package (Version 3.2). Cladistics 5:164–166.

Frantz LAF, Bradley DG, Larson G, Orlando L. 2020. Animal domestication in the era of ancient genomics. Nat. Rev. Genet. 21:449–460.

Frantz LAF, Mullin VE, Pionnier-Capitan M, Lebrasseur O, Ollivier M, Perri A, Linderholm A, Mattiangeli V, Teasdale MD, Dimopoulos EA, et al. 2016. Genomic and archaeological evidence suggest a dual origin of domestic dogs. Science 352:1228–1231.

Freedman AH, Gronau I, Schweizer RM, Ortega-Del Vecchyo D, Han E, Silva PM, Galaverni M, Fan Z, Marx P, Lorente-Galdos B, et al. 2014. Genome Sequencing Highlights the Dynamic Early History of Dogs. PLoS Genet. 10:e1004016.

Garland P, Quraishe S, French P, O'Connor V. 2008. Expression of the MAST family of serine/threonine kinases. Brain Res. 1195:12–19.

Gautier M, Klassmann A, Vitalis R. 2017. rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. Mol. Ecol. Resour. 17:78–90.

Gautier M, Vitalis R. 2012. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. Bioinformatics 28:1176–1177.

Gel B, Serra E. 2017. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics 33:3088–3090.

Geza E, Mugo J, Mulder NJ, Wonkam A, Chimusa ER, Mazandu GK. 2019. A comprehensive survey of models for dissecting local ancestry deconvolution in human genome. Brief. Bioinform. 20:1709–1724.

Godinho R, Llaneza L, Blanco JC, Lopes S, Álvares F, García EJ, Palacios V, Cortés Y, Javier T, Ferrand N. 2011. Genetic evidence for multiple events of hybridization between wolves and domestic dogs in the Iberian Peninsula. Mol. Ecol. 20:5154–5166.

Gómez-Sánchez D, Olalde I, Sastre N, Enseñat C, Carrasco R, Marques-Bonet T, Lalueza-Fox C, Leonard JA, Vilà C, Ramírez O. 2018. On the path to extinction: Inbreeding and admixture in a declining grey wolf population. Mol. Ecol. 27:3599–3612.

Grabenstein KC, Taylor SA. 2018. Breaking Barriers: Causes, Consequences, and Experimental Utility of Human-Mediated Hybridization. Trends Ecol. Evol. 33:198–212.

Guan Y. 2014. Detecting Structure of Haplotypes and Local Ancestry. Genetics 196:625–642.

Haller BC, Messer PW. 2019. SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. Mol. Biol. Evol. 36:632–637.

Hare B, Wobber V, Wrangham R. 2012. The self-domestication hypothesis: evolution of bonobo psychology is due to selection against aggression. Anim. Behav. 83:573–585.

Huerta-Sánchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, Liang Y, Yi X, He M, Somel M, et al. 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature 512:194–197.

Koch EM, Schweizer RM, Schweizer TM, Stahler DR, Smith DW, Wayne RK, Novembre J. 2019. De novo mutation rate estimation in wolves of known pedigree. Mol. Biol. Evol. 36:2536–2547.

INE (2016). Península Ibérica em números 2016. Instituto Nacional de Estatistica, Portugal. ISBN: 978-989-25-0432-2.

Lee TH, Guo H, Wang X, Kim C, Paterson AH. 2014. SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. BMC Genomics 15:162.

Leigh JW, Bryant D. 2015. popart: full-feature software for haplotype network construction. Methods Ecol. Evol. 6:1110–1116.

Lequarré A-S, Andersson L, André C, Fredholm M, Hitte C, Leeb T, Lohi H, Lindblad-Toh K, Georges M. 2011. LUPA: A European initiative taking advantage of the canine genome architecture for unravelling complex disorders in both human and dogs. Vet. J. 189:155–159.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987–2993.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.

Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ, Zody MC, et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438:803–819.

Liu Y, Nyunoya T, Leng S, Belinsky SA, Tesfaigzi Y, Bruse S. 2013. Softwares and methods for estimating genetic ancestry in human populations. Hum. Genomics 7:1.

Llaneza L, García EJ, Palacios V, Sazatornil V, López-Bao JV. 2016. Resting in risky environments: the importance of cover for wolves to cope with exposure risk in human-dominated landscape. Biodivers. Conserv. 25:1515-1528.

Llaneza L, López-Bao JV. 2015. Indirect effects of changes in environmental and agricultural policies on the diet of wolves. Eur. J. Wildl. Res. 61:895-902.

Llaneza L, López-Bao JV, Sazatornil V. 2012. Insights into wolf presence in human-dominated landscapes: the relative role of food availability, humans and landscape attributes. Divers. Distrib. 18:459–469.

Mak SST, Gopalakrishnan S, Carøe C, Geng C, Liu S, Sinding M-HS, Kuderna LFK, Zhang W, Fu S, Vieira FG, et al. 2017. Comparative performance of the BGISEQ-500 vs Illumina HiSeq2500 sequencing platforms for palaeogenomic sequencing. Gigascience 6:gix049.

Marsden CD, Ortega-Del Vecchyo D, O'Brien DP, Taylor JF, Ramirez O, Vilà C, Marques-Bonet T, Schnabel RD, Wayne RK, Lohmueller KE. 2016. Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. Proc. Natl. Acad. Sci. 113:152–157.

Martin SH, Davey JW, Jiggins CD. 2015. Evaluating the Use of ABBA–BABA Statistics to Locate Introgressed Loci. Mol. Biol. Evol. 32:244–257.

Martins-de-Souza D, Guest PC, Mann DM, Roeber S, Rahmoune H, Bauder C, Kretzschmar H, Volk B, Baborie A, Bahn S. 2012. Proteomic Analysis Identifies Dysfunction in Cellular Transport, Energy, and Protein Metabolism in Different Brain Regions of Atypical Frontotemporal Lobar Degeneration. J. Proteome Res. 11:2533–2543.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303.

Mech LD, Barber-Meyer SM, Erb J. 2016. Wolf (Canis lupus) Generation Time and Proportion of Current Breeding Females by Age. PLoS One 11:e0156682.

Miao B, Wang Z, Li Y. 2017. Genomic analysis reveals hypoxia adaptation in the tibetan mastiff by introgression of the gray Wolf from the tibetan plateau. Mol. Biol. Evol. 34:734–743.

Newsome TM, Boitani L, Chapron G, Ciucci P, Dickman CR, Dellinger JA, López-Bao J V., Peterson RO, Shores CR, Wirsing AJ, et al. 2016. Food habits of the world's grey wolves. Mamm. Rev. 46:255–269.

Nores C, López-Bao JV. 2022. Historial data to inform the legal status of species in Europe: An example with wolves. Biol. Conserv. 272:109639.

O'Rourke T, Boeckx C. 2020. Glutamate receptors in domestication and modern human evolution. Neurosci. Biobehav. Rev. 108:341–357.

Pacheco C, Lobo D, Silva P, Álvares F, García EJ, Castro D, Layna JF, López-Bao JV, Godinho R. 2022. Assessing the performance of historical skins and bones for museomics using wolf specimens as a case study. Front. Ecol. Evol. 10:970249.

Pasaniuc B, Sankararaman S, Kimmel G, Halperin E. 2009. Inference of locus-specific ancestry in closely related populations. Bioinformatics 25:i213–i221.

Perri AR, Feuerborn TR, Frantz LAF, Larson G, Malhi RS, Meltzer DJ, Witt KE. 2021. Dog

domestication and the dual dispersal of people and dogs into the Americas. Proc. Natl. Acad. Sci. 118:e2010083118.

Pickrell JK, Pritchard JK. 2012. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. PLoS Genet. 8:e1002967.

Pilot M, Greco C, VonHoldt BM, Jędrzejewska B, Randi E, Jędrzejewski W, Sidorovich VE, Ostrander EA, Wayne RK. 2014. Genome-wide signatures of population bottlenecks and diversifying selection in European wolves. Heredity 112:428–442.

Pilot M, Greco C, VonHoldt BM, Randi E, Jędrzejewski W, Sidorovich VE, Konopiński MK, Ostrander EA, Wayne RK. 2018. Widespread, long-term admixture between grey wolves and domestic dogs across Eurasia and its implications for the conservation status of hybrids. Evol. Appl.11:662–680.

Pilot M, Moura AE, Okhlopkov IM, Mamaev NsabetiV., Manaseryan NH, Hayrapetyan V, Kopaliani N, Tsingarska E, Alagaili AN, Mohammed OB, et al. 2021. Human-modified canids in human-modified landscapes: The evolutionary consequences of hybridization for grey wolves and free-ranging domestic dogs. Evol. Appl. 14:2433–2456.

Plassais J, Kim J, Davis BW, Karyadi DM, Hogan AN, Harris AC, Decker B, Parker HG, Ostrander EA. 2019. Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. Nat. Commun. 10:1489.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. 81:559–575.

R Development Core Team. 2017. R: A language and environment for statistical computing. Vienna, Austria.

Raxworthy CJ, Smith BT. 2021. Mining museums for historical DNA: advances and challenges in museomics. Trends Ecol. Evol. 36:1049–1060.

Revelles J, Burjachs F, Palomo A, Piqué R, Iriarte E, Pérez-Obiol R, Terradas X. 2018. Human-environment interaction during the Mesolithic- Neolithic transition in the NE Iberian Peninsula. Vegetation history, climate change and human impact during the Early-Middle Holocene in the Eastern Pre-Pyrenees. Quat. Sci. Rev. 184:183–200.

Rhymer JM, Simberloff D. 1996. Extinction by hybridization and introgression. Annu. Rev. Ecol. Syst. 27:83–109.

Rio-Maior H, Nakamura M, Álvares F, Beja P. 2019. Designing the landscape of coexistence: Integrating risk avoidance, habitat selection and functional connectivity to inform large carnivore conservation. Biol. Conserv. 235:178–188.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCaroll SA, Gaudet R, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. Nature 449:913-918.

Sankararaman S, Sridhar S, Kimmel G, Halperin E. 2008. Estimating Local Ancestry in Admixed Populations. Am. J. Hum. Genet. 82:290–303.

Satizabal CL, Adams HHH, Hibar DP, White CC, Knol MJ, Stein JL, Scholz M, Sargurupremraj M, Jahanshad N, Roshchupkin GV, et al. 2019. Genetic architecture of subcortical brain structures in 38,851 individuals. Nat. Genet. 51:1624–1636.

Sazatornil V, Rodríguez A, Klaczek M, Ahmadi M, Álvares F, Arthur S, Blanco JC, Borg BL,

Cluff D, Cortés Y, et al. 2016. The role of human-related risk in breeding site selection by wolves. Biol. Conserv. 201:103–110.

Schubert M, Ermini L, Sarkissian C Der, Jónsson H, Ginolhac A, Schaefer R, Martin MD, Fernández R, Kircher M, McCue M, et al. 2014. Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. Nat. Protoc. 9:1056–1082.

Schubert M, Lindgreen S, Orlando L. 2016. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res. Notes 9:88.

Schweizer RM, Durvasula A, Smith J, Vohr SH, Stahler DR, Galaverni M, Thalmann O, Smith DW, Randi E, Ostrander EA, et al. 2018. Natural Selection and Origin of a Melanistic Allele in North American Gray Wolves. Mol. Biol. Evol. 35:1190–1209.

Seixas FA, Boursot P, Melo-Ferreira J. 2018. The genomic impact of historical hybridization with massive mitochondrial DNA introgression. Genome Biol. 19:91.

Shulman JM, Imboywa S, Giagtzoglou N, Powers MP, Hu Y, Devenport D, Chipendo P, Chibnik LB, Diamond A, Perrimon N, et al. 2014. Functional screening in Drosophila identifies Alzheimer's disease susceptibility genes and implicates Tau-mediated mechanisms. Hum. Mol. Genet. 23:870–877.

Silva P, Galaverni M, Ortega-Del Vecchyo D, Fan Z, Caniglia R, Fabbri E, Randi E, Wayne R, Godinho R. 2020. Genomic evidence for the Old divergence of Southern European wolf populations. Proc. R. Soc. B Biol. Sci. 287:20201206.

Silva P, López-Bao JV, Llaneza L, Álvares F, Lopes S, Blanco JC, Cortés Y, García E, Palacios V, Rio-Maior H, et al. 2018. Cryptic population structure reveals low dispersal in Iberian wolves. Sci. Rep. 8:14108.

Skoglund P, Ersmark E, Palkopoulou E, Dalén L. 2015. Ancient wolf genome reveals an early divergence of domestic dog ancestors and admixture into high-latitude breeds. Curr. Biol. 25:1515–1519.

Smith J, Coop G, Stephens M, Novembre J. 2018. Estimating Time to the Common Ancestor for a Beneficial Allele. Mol. Biol. Evol. 35:1003–1017.

Smith J, Kronforst MR. 2013. Do Heliconius butterfly species exchange mimicry alleles? Biol. Lett. 9:20130503.

Steffens M, Leu C, Ruppert A-K, Zara F, Striano P, Robbiano A, Capovilla G, Tinuper P, Gambardella A, Bianchi A, et al. 2012. Genome-wide association analysis of genetic generalized epilepsies implicates susceptibility loci at 1q43, 2p16.1, 2q22.3 and 17q21.32. Hum. Mol. Genet. 21:5359–5372.

Stephens M, Donnelly P. 2003. A Comparison of Bayesian Methods for Haplotype Reconstruction from Population Genotype Data. Am. J. Hum. Genet. 73:1162–1169.

Stronen AV, Jędrzejewska B, Pertoldi C, Demontis D, Randi E, Niedziałkowska M, Borowik T, Sidorovich VE, Kusak J, Kojola I, et al. 2015. Genome-wide analyses suggest parallel selection for universal traits may eclipse local environmental selection in a highly mobile carnivore. Ecol. Evol. 5:4410–4425.

Sukumaran J, Holder MT. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569–1571.

Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, et al. 2019. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental

datasets. Nucleic Acids Res. 47:D607–D613.

Tang H, Choudhry S, Mei R, Morgan M, Rodriguez-Cintron W, Burchard EG, Risch NJ. 2007. Recent Genetic Selection in the Ancestral Admixture of Puerto Ricans. Am. J. Hum. Genet. 81:626–633.

Taylor SA, Larson EL. 2019. Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. Nat. Ecol. Evol. 3:170–177.

Turner DS. 2014. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots.

Vaysse A, Ratnakumar A, Derrien T, Axelsson E, Rosengren Pielberg G, Sigurdsson S, Fall T, Seppälä EH, Hansen MST, Lawley CT, et al. 2011. Identification of Genomic Regions Associated with Phenotypic Variation between Dog Breeds using Selection Mapping. PLoS Genet. 7:e1002316.

Wang G, Zhai W, Yang H, Fan R, Cao X, Zhong L, Wang L, Liu F, Wu H, Cheng L, et al. 2013. The genomics of selection in dogs and the parallel evolution between dogs and humans. Nat. Commun. 4:1860.

Wang MS, Wang S, Li Y, Jhala Y, Thakur M, Otecko NO, Si JF, Chen HM, Shapiro B, Nielsen R, et al. 2020. Ancient Hybridization with an Unknown Population Facilitated High-Altitude Adaptation of Canids. Mol. Biol. Evol. 37:2616–2629.

Ward BJ, van Oosterhout C. 2016. hybridcheck: software for the rapid detection, visualization and dating of recombinant regions in genome sequence data. Mol. Ecol. Resour. 16:534–539.

Weir BS, Cockerham CC. 1984. Estimating F-Statistics for the Analysis of Population Structure. Evolution 38:1358.

Wong AK, Ruhe AL, Dumont BL, Robertson KR, Guerrero G, Shull SM, Ziegle JS, Millon L V., Broman KW, Payseur BA, et al. 2010. A Comprehensive Linkage Map of the Dog Genome. Genetics 184:595–605.

Zeder MA. 2008. Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. Proc. Natl. Acad. Sci. 105:11597–11604.

Zhang W, Fan Z, Han E, Hou R, Zhang L, Galaverni M, Huang J, Liu H, Silva P, Li P, et al. 2014. Hypoxia Adaptations in the Grey Wolf (Canis lupus chanco) from Qinghai-Tibet Plateau. PLoS Genet.10:e1004466.

Zheng Y, Janke A. 2018. Gene flow analysis method, the D-statistic, is robust in a wide parameter space. BMC Bioinformatics 19:10.

Zhu Y, Li W, Yang B, Zhang Z, Ai H, Ren J, Huang L. 2017. Signatures of Selection and Interspecies Introgression in the Genome of Chinese Domestic Pigs. Genome Biol. Evol. 9:2592–2603.

3.6. Supplementary Materials

Supplementary Text

Global and local ancestry results

ADMIXTURE results based on the 85,000 genome-wide SNP dataset (hereafter 85K) showed that the average assignment of dogs, Iberian, and Eurasian wolves to the corresponding cluster at K = 3 was 0.98, 0.99, and 0.97, respectively (Supplementary Figure 3.2). By increasing the number of clusters up to K = 5, we found additional differentiation among dog breeds and finer population structure within Iberian wolves (Supplementary Figure 3.2). The second analysis in ADMIXTURE, for K = 2, to determine whether there were recently admixed individuals in our dataset (i.e., wolf-dog hybrids of few generations), revealed that all the individuals have met the criterion of having less than 0.1 of ancestry assigned to the dog cluster. Likewise, all dogs had less than 0.05 of wolf ancestry in their genome (Supplementary Figure 3.3).

ADMIXTURE results based on WGS data, revealed that only one wolf from Spain (Wolf27; publicly available data), displayed signs of recent admixture with dogs (q = 0.11 to dog cluster; Supplementary Figure 3.4). This is consistent with previous studies identifying Wolf27 as a hybrid of few generations (Gómez-Sánchez et al. 2018).

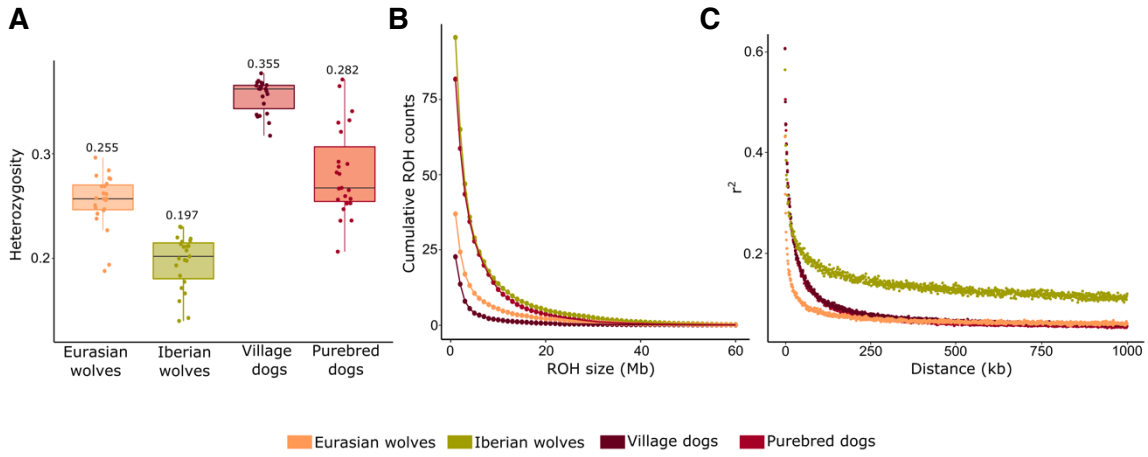
Estimates of local ancestry determined by ELAI on the 85K SNP dataset, revealed that most of the Iberian wolf ancestry (~0.99) was attributed to an unsampled reference population and only 0.2% was attributed to the Eurasian wolf population (Supplementary Figure 3.5). This suggests that relying on such divergent populations as a reference in this type of analysis can lead to overestimation of introgression in the opposite direction and should be avoided. Hence, when a close reference population is not available, we highly recommend letting ELAI infer it and use it along with the reference panel available. Both LAMP-ANC and ELAI retrieved very similar results across chromosomes using the 85K SNPs, attributing a global proportion of dog ancestry in the genome of Iberian wolves of 0.01 (Supplementary Figures 3.5 and 3.6). There were several small dog chromosomal blocks introgressed along the genome, but at low frequency among the entire population. Both methods detected a 11 Mb block on chromosome 2 (44 – 55 Mb) across the same 22 Iberian wolves.

Among the 15 Iberian wolves analyzed with WGS data, the dog block had been previously detected in two of them using the 85K SNPs (L590 and L588). Using WGS in ELAI, we confirmed the dog block in these two individuals and in another three Iberian wolves (L474 also genotyped with the 85K SNPs; Wolf 24 and Wolf39, from publicly available data). Both Wolf39 and L590 carried an intact full introgressed block (chr 2: 43.5 – 55.1 Mb), whereas L474, L588 and Wolf24 only kept parts of it. We found concordant positions, and lengths, with those estimated using the 85K dataset for individuals carrying longer introgressed tracks (L588, L590), but not for those with smaller regions. The latter being the case of L474, which was previously detected as introgressed only when we considered older events (> 100 generations in ELAI), exhibiting only a smaller fraction of the block comparative to WGS based analysis. These findings reinforce the possibility that both the frequency and extent of introgression can be underestimated, when solely based on the results of the SNP chip data. Local ancestry estimates retrieved from ELAI confirmed that Wolf27 had a substantial amount of dog ancestry in chromosome 2 (42%; Supplementary Figure 3.8), and for that reason was excluded from the following analysis.

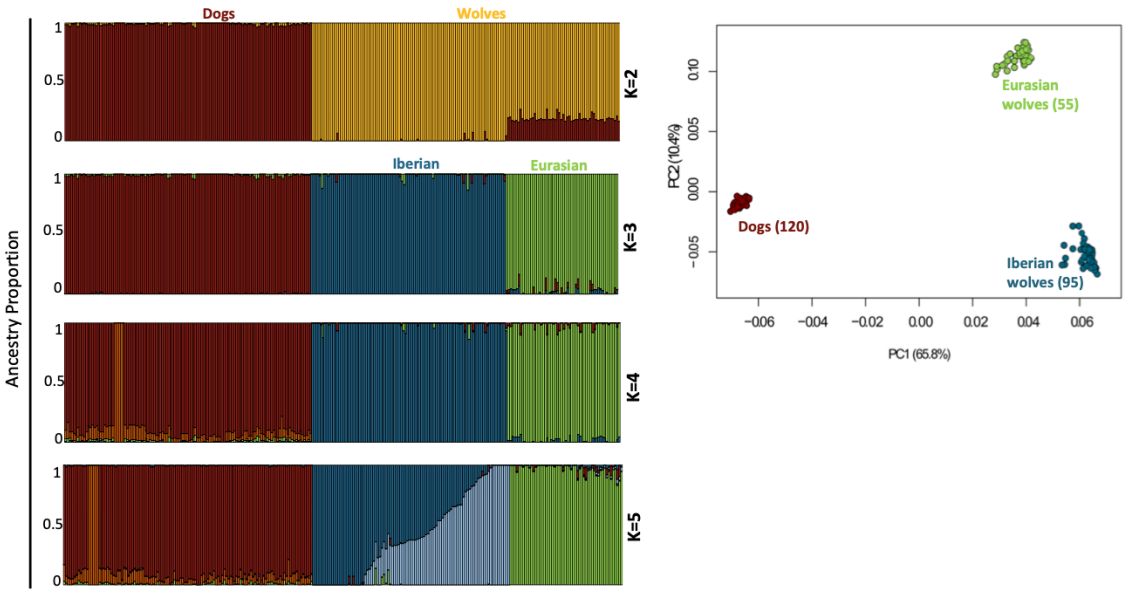
Local ancestry results across historical Iberian wolf samples

LAMP-ANC results confirmed the presence of the same dog introgressed block on chromosome 2 of 13 historical samples of Iberian wolves (Supplementary Figure 3.10). The SNP-specific Δancestry validated the dog block on chromosome 2 as an outlier region, positioned between 50.02 – 53.35 Mb spanning 24 SNPs. LAMP-ANC revealed a global proportion of dog ancestry in the genome of historical Iberian wolf samples bellow 0.01, without any signs of recent admixture with dogs. A NJ tree based on pairwise genetic distances of the SNPs within the Δblock showed two additional Iberian wolves clustering together with dogs that were not detected by LAMP-ANC. To ensure that the analysis was not biased by the limited number of loci in this dataset (chr 2 = 623 SNPs), we subsampled the historical dataset and extracted only Iberian wolf samples carrying the Δblock on chromosome 2 with at least 60,000 genome-wide SNPs. Among the 15 historical Iberian wolf samples with the Δblock, 13 had more than 60K genome-wide SNPs genotyped (including the two where the Δblock was not detected in the first LAMP-ANC run). We subsequently ran LAMP-ANC with this dataset and obtained the same results, but this time the dog block was detected in the two individuals not presenting any signs of the Δblock in the first run (Supplementary Figure 3.11). This approach has allowed us to discard possible false positives due to low-density markers, as the number

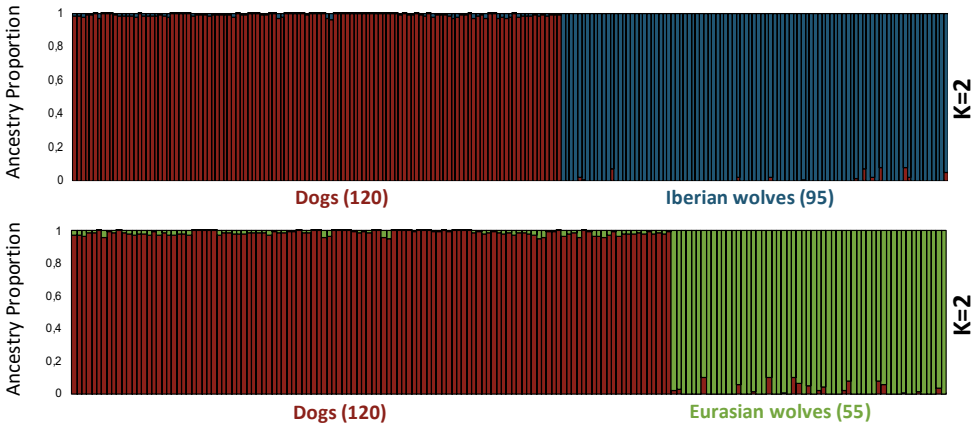
of SNPs in chromosome 2 and within the Δblock increased from 623 to 1712 and from 24 to 69 SNPs, respectively.



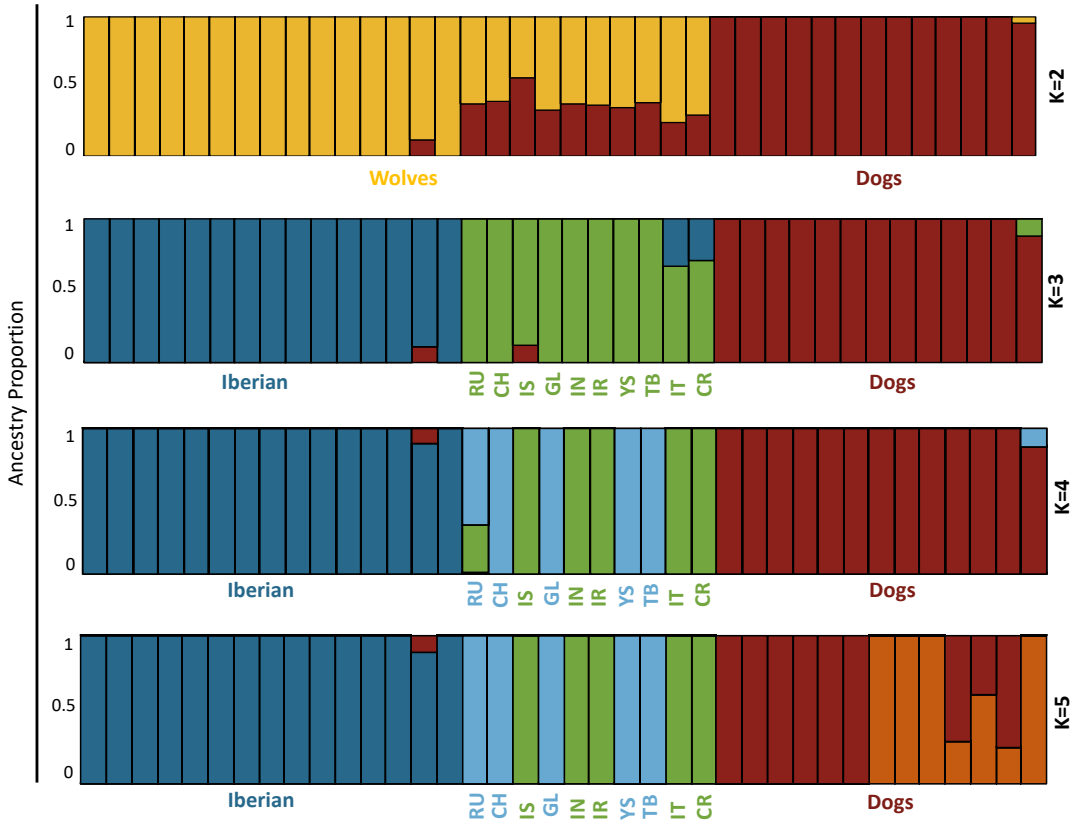
Supplementary Figure 3.1. Genetic Diversity. (**A**) Boxplots representing the proportion of heterozygous sites across the different groups. Each dot represents an individual and mean values are presented. (**B**) Inverse cumulative average number of runs of homozygosity (ROH) per individual across 1 Mb non-overlapping windows relative to their average size (Mb). (**C**) Decay of linkage disequilibrium. Each dot represents the averaged pairwise r^2 values between pairs of SNPs in non-overlapping 1 kb windows.



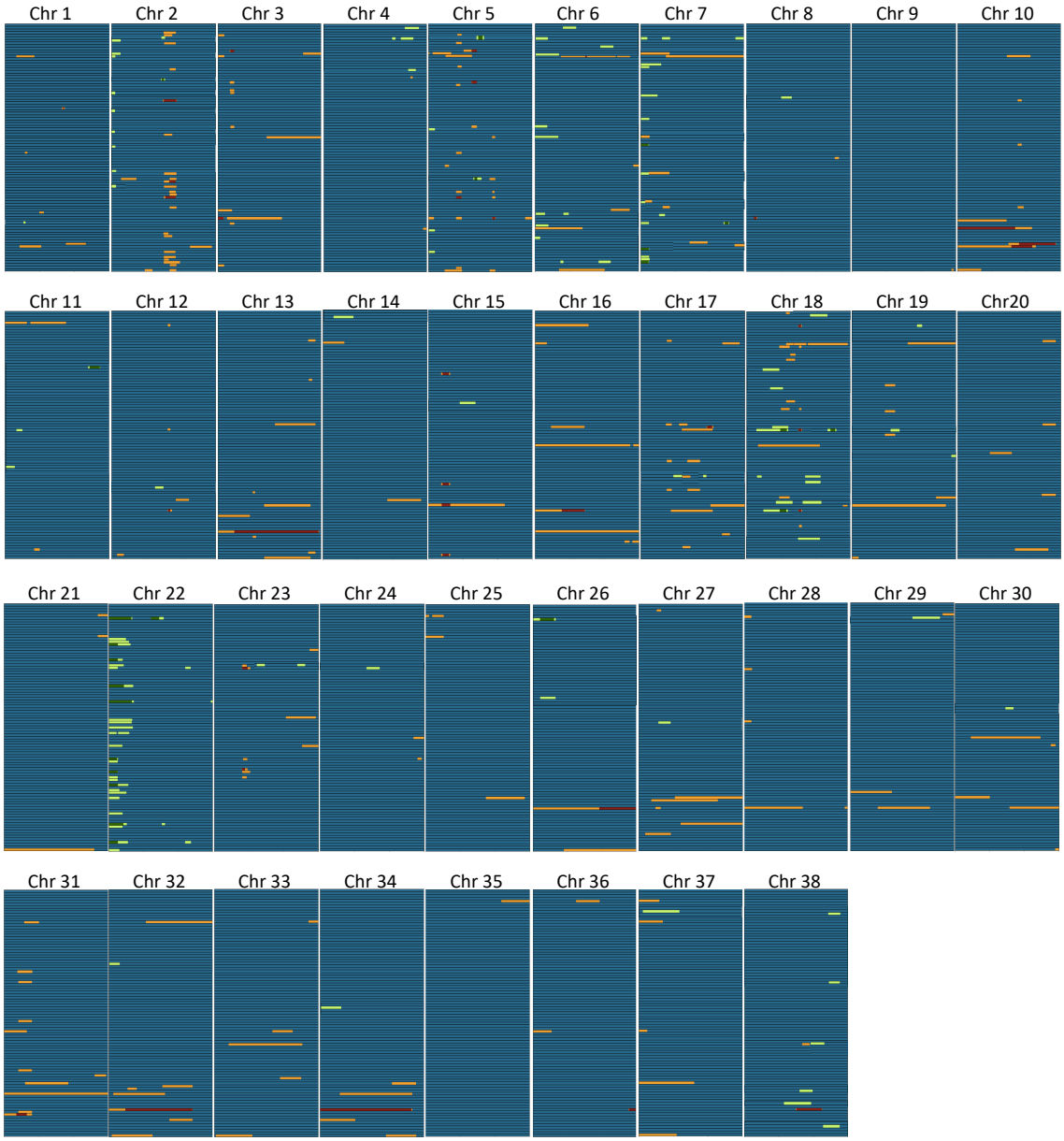
Supplementary Figure 3.2. Global ancestry analysis. Ancestry fractions of all individuals inferred using the 85K SNP data in ADMIXTURE for K = 2–5 clusters (left) and PCA (right). Each admixture graph represents one K cluster, where individuals are identified by a single bar. Different colors indicate the ancestry proportion attributed to each cluster. PCA differentiates domestic dogs from all wolves in PC1, and PC2 separates the Iberian population from the remaining Eurasian wolves.



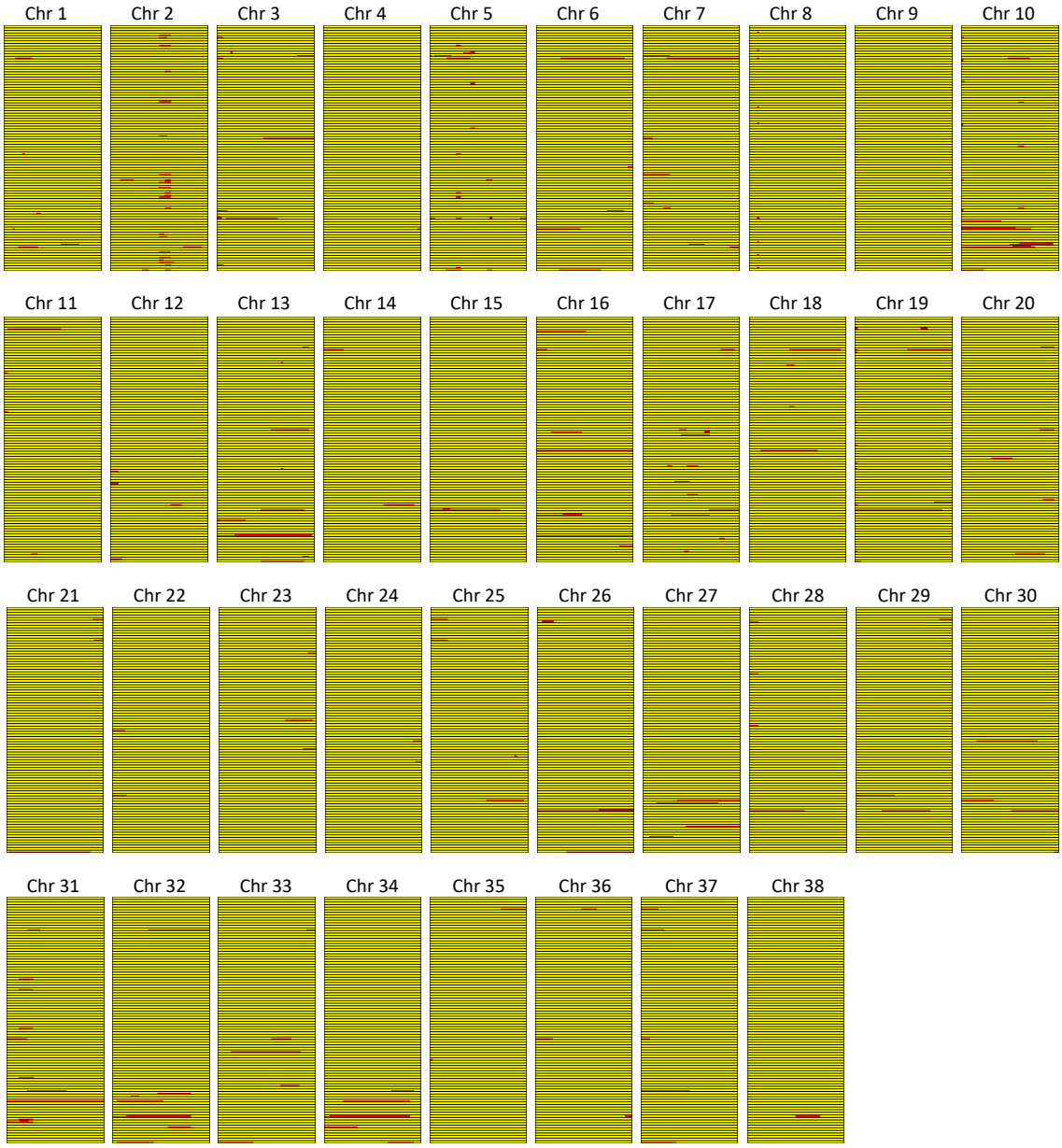
Supplementary Figure 3.3. Admixture proportions. Ancestry fractions of dogs and wolves inferred using the 85K SNP data in ADMIXTURE for K = 2. Iberian and Eurasian wolves were run independently in two separate datasets with dogs.



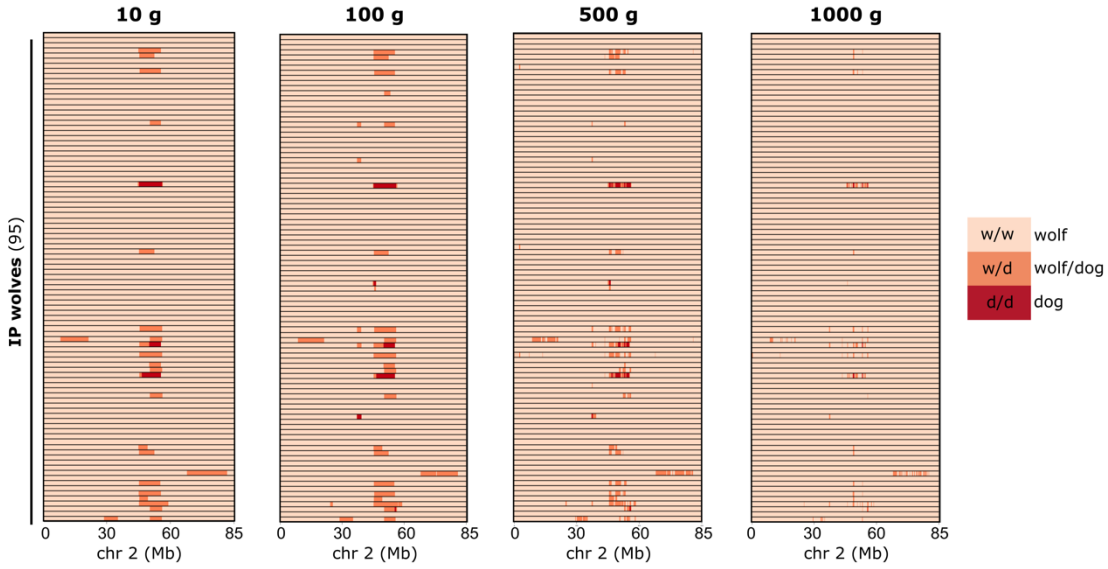
Supplementary Figure 3.4. Global ancestry analysis using WGS data. Ancestry fractions of all individual genomes inferred using ADMIXTURE in K = 2–5 clusters. Each admixture graph represents a K cluster and individuals are identified by a single bar. Different colors indicate the ancestry proportion attributed to each cluster. Abbreviation for wolf's origin: RU, Russia; CH, China; IS, Israel; GL, Great Lakes; IN, India; IR, Iran; YS, Yellowstone; TB, Tibet; IT, Italy; CR, Croatia.



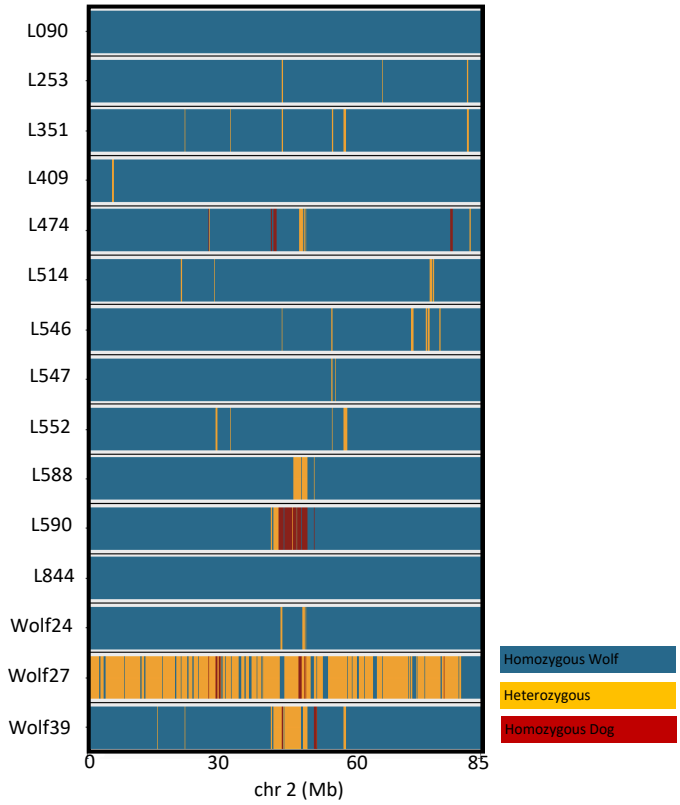
Supplementary Figure 3.5. Local ancestry analysis using ELAI. Results of local ancestry analysis in ELAI (10 mixture generations) based on 85K SNPs. Blocks were colored based on the attributed local ancestry: blue for the unsampled population (both chromosomes), dark green for Eurasian wolf (both chromosomes), red for dog (both chromosomes), light green and yellow represent a mixture of unsampled/Eurasian wolf and unsampled/dog (one chromosome of each), respectively. Each box identifies a chromosome, and every row represents an Iberian wolf sample (N = 95).



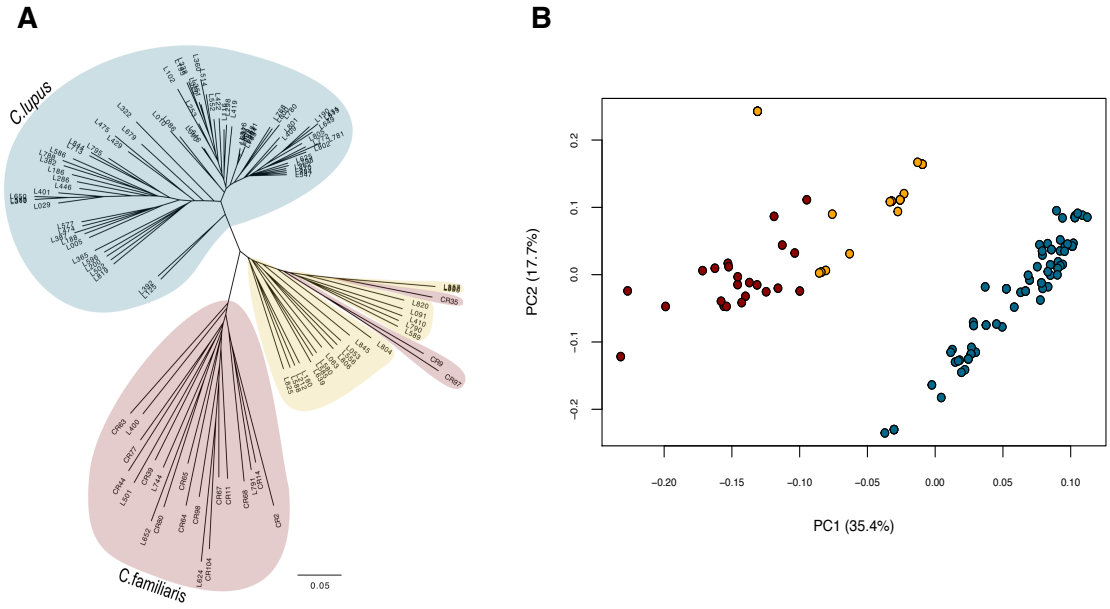
Supplementary Figure 3.6. Local ancestry analysis using LAMP. Results of local ancestry analysis in LAMP-ANC based on 85K SNPs. Blocks were colored based on the attributed local ancestry: yellow for wolf (both chromosomes), red for dog (both chromosomes), and a mixture of yellow and red for wolf/dog (one chromosome of each). Each box identifies a chromosome, and every row represents an Iberian wolf sample (N = 95).



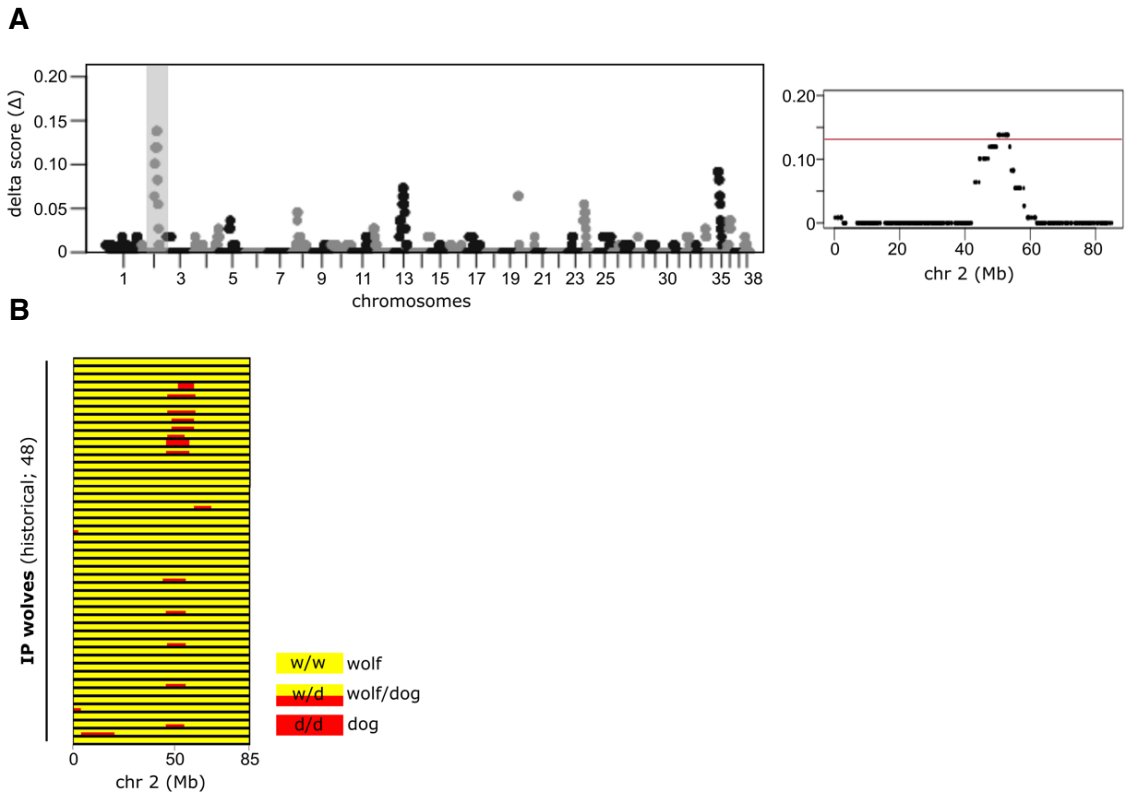
Supplementary Figure 3.7. Representation of the dog block found using 85K SNPs. Representation of the dog block found in chromosome 2 of Iberian wolves based on local ancestry analysis in ELAI, considering several mixture generations (10, 100, 500, and 1000). Each row represents a single individual and different colors represent the attributed local ancestry: light orange for wolf (both chromosomes), red for dog (both chromosomes), and orange represents a mixture of wolf/dog (one chromosome of each).



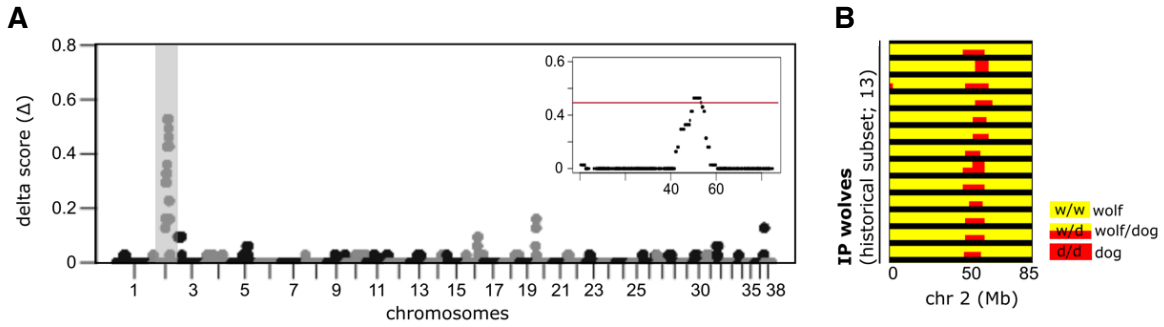
Supplementary Figure 3.8. Local ancestry analysis using WGS data. Representation of the Δblock on chromosome 2 across all Iberian wolf genomes, based on ELAI results. Each row represents an individual and different colors represent the attributed local ancestry: blue for wolf (both chromosomes), red for dog (both chromosomes), and yellow represents a mixture of wolf/dog (one chromosome of each).



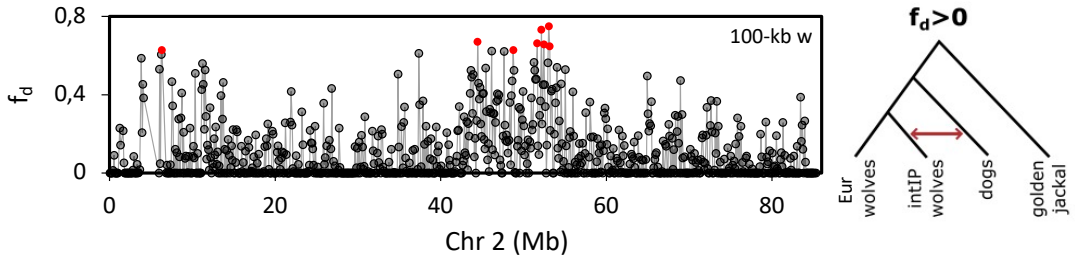
Supplementary Figure 3.9. Individual clustering within Δblock. NJ tree (**A**) and PCA (**B**) of Iberian wolves and village dogs from Iberia, constructed based on 78 SNPs spanning the Δblock on chromosome 2 defined by Δstatistics (chr 2: 50.02 – 52.62 Mb). Iberian wolves without the Δblock are represented in blue, Iberian wolves with the Δblock in yellow and village dogs in red.



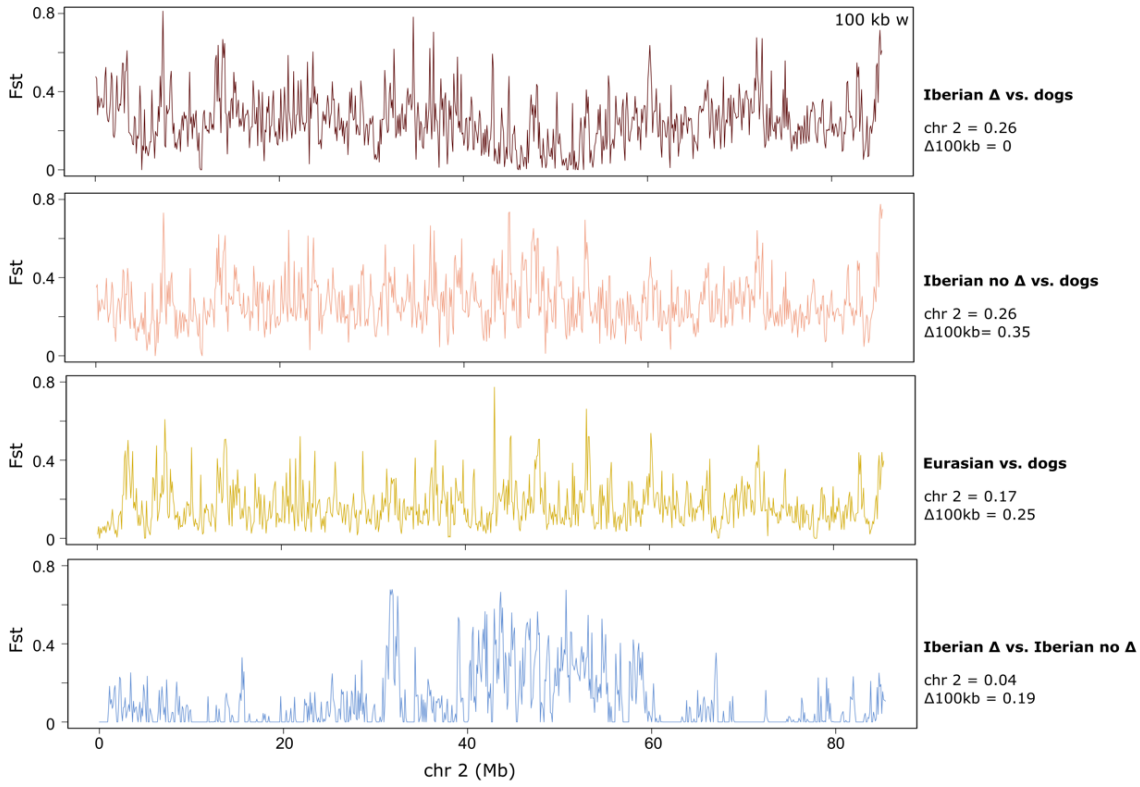
Supplementary Figure 3.10. Dog block in historical samples. (**A**) Δancestry scores indicating the excess of dog ancestry across Iberian wolf genome. Δblock on chromosome 2 is highlighted in gray and has a zoomed-in view with the red line indicating the cut-off used for detecting outlier regions. (**B**) Representation of the dog block found in chromosome 2 of Iberian wolf historical samples, based on local ancestry analysis. Each row represents a single individual and different colors represent the attributed local ancestry: yellow for wolf (both chromosomes), red for dog (both chromosomes), and a mixture of yellow and red for wolf/dog (one chromosome of each).



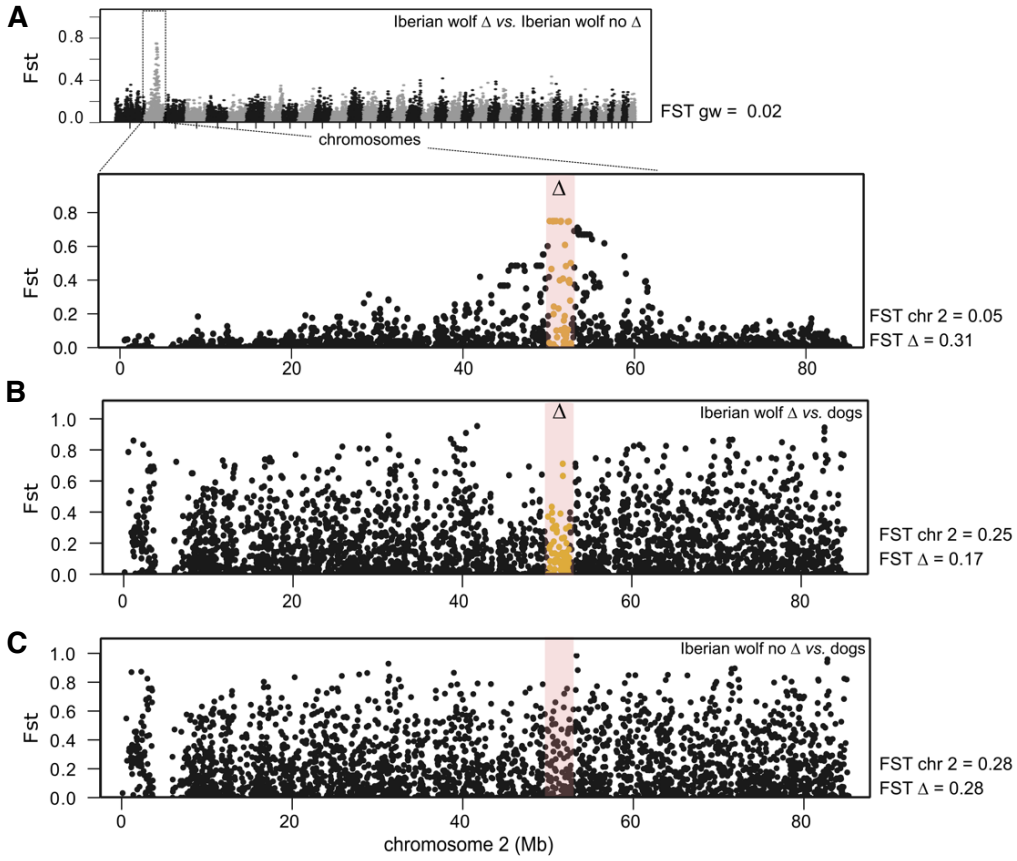
Supplementary Figure 3.11. Dog block found using a subset of historical samples. (**A**) Δancestry scores indicating the excess of dog ancestry across Iberian wolf genome. Δblock on chromosome 2 is highlighted in gray and has a zoomed-in view with the red line indicating the cut-off used for detecting outlier regions. (**B**) Representation of the dog block found in chromosome 2 of Iberian wolf historical samples (subset with > 60K SNPs), based on local ancestry analysis. Each row represents a single individual and different colors represent the attributed local ancestry: yellow for wolf (both chromosomes), red for dog (both chromosomes), and a mixture of yellow and red for wolf/dog (one chromosome of each).



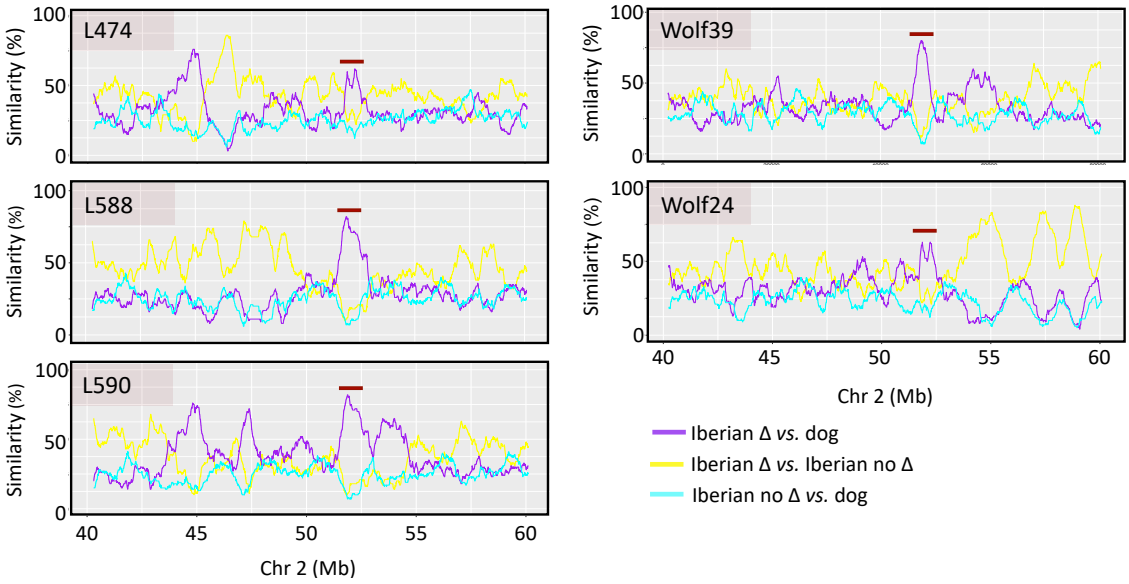
Supplementary Figure 3.12. Signatures of variant sharing. Fraction of introgression (f_d statistics) across chromosome 2, estimated using WGS data under the phylogenetic conformation displayed at the right (P1 – Eurasian wolves, P2 – Iberian wolves with the Δblock, P3 – dogs, O – golden jackal). Each dot represents a 100 kb window, and red dots represent the top 1% windows with the highest fraction of introgression.



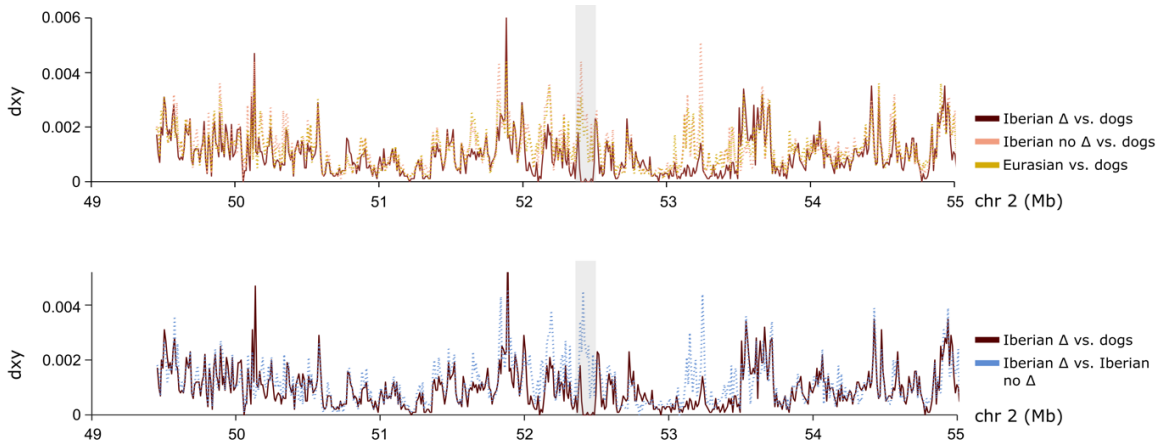
Supplementary Figure 3.13. Levels of genetic differentiation using WGS data. Weighted F_{ST} values estimated along chromosome 2 in 100 kb windows across several groups. Average values for the entire chromosome and within the Δ100 kb introgressed region (chr 2: 52.4 – 52.5) are provided on the right.



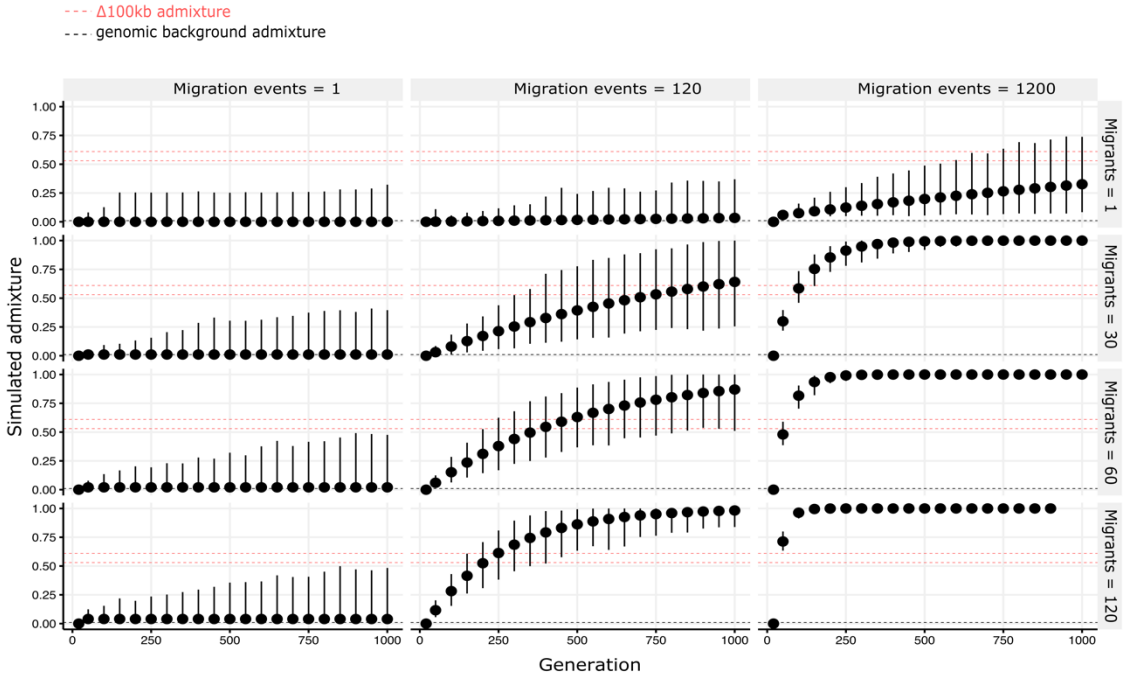
Supplementary Figure 3.14. Levels of genetic differentiation using the 85K SNPs. (**A**) Manhattan plot of the genome-wide distribution of F_{ST} values across the 38 autosomes, with a zoomed-in view over chromosome 2, between Iberian wolves with and without the Δblock. (**B**/**C**) Pairwise F_{ST} values along chromosome 2 between Iberian wolves with and without the Δblock and dogs. Average F_{ST} values of genome-wide, along chromosome 2 and within Δblock are presented on the right. The position of the Δblock is highlighted by a red bar and SNPs within it are colored in orange in **A**) and **B**). Each dot represents a SNP.



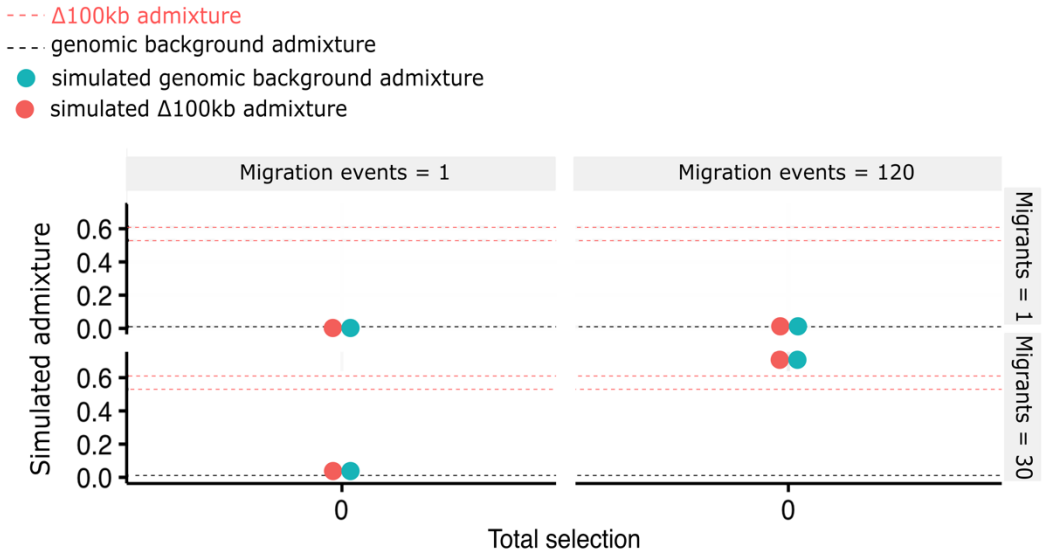
Supplementary Figure 3.15. Levels of genetic sequence similarity within Δblock. Pairwise sequence similarity analysis using HYBRIDCHECK between haplotype sequences of Iberian wolves with the Δblock, an Iberian wolf without the Δblock, and one Portuguese village dog across chromosome 2 (subset region chr 2: 40 – 60 Mb). Each panel represents the results of each triplet analysis, alternating individuals with the Δblock. The percentage of sequence similarity was estimated along the chromosome in 1 kb windows. The red horizontal bar in each panel indicates the position of Δ100 kb.



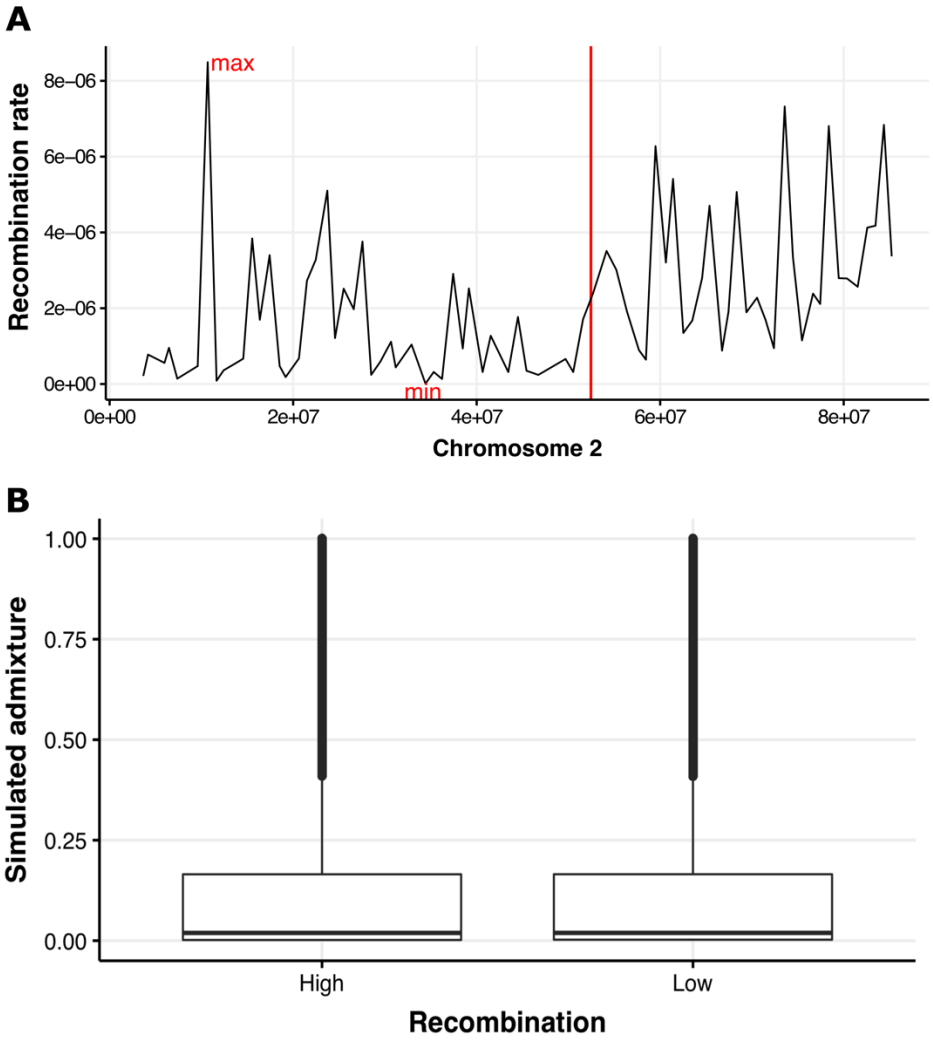
Supplementary Figure 3.16. Levels of genetic divergence. Mean pairwise sequence divergence (d_{xy}) estimated in 10 kb non-overlapping windows along 6 Mb haplotypes surrounding the Δ100 kb region, between several groups (each comparison is color coded). In the top and bottom panels, the gray bar highlights the position of Δ100 kb in the chromosome.



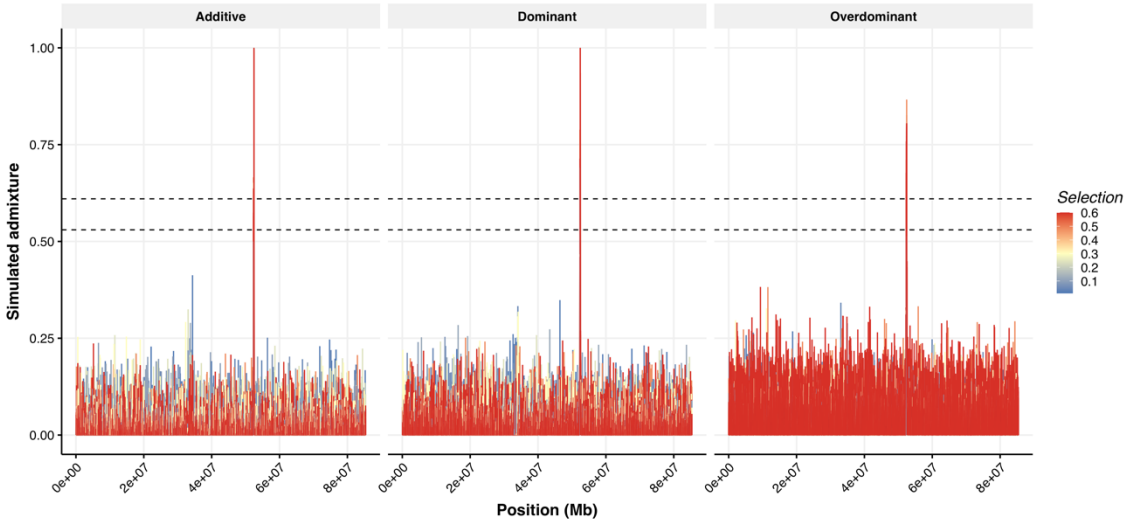
Supplementary Figure 3.17. Neutral simulations. Mean and standard deviation of 100 simulated replicates in SLiM for average chromosome-wide admixture per generation (0 represents the present time) across neutral models. A range of migration events and number of migrants were simulated. Empirical estimates of dog admixture within the Δ100 kb region (0.53-0.60; 95% CI of ELAI average estimate) and genome-wide level (0.01) are shown in red and black dashed lines, respectively.



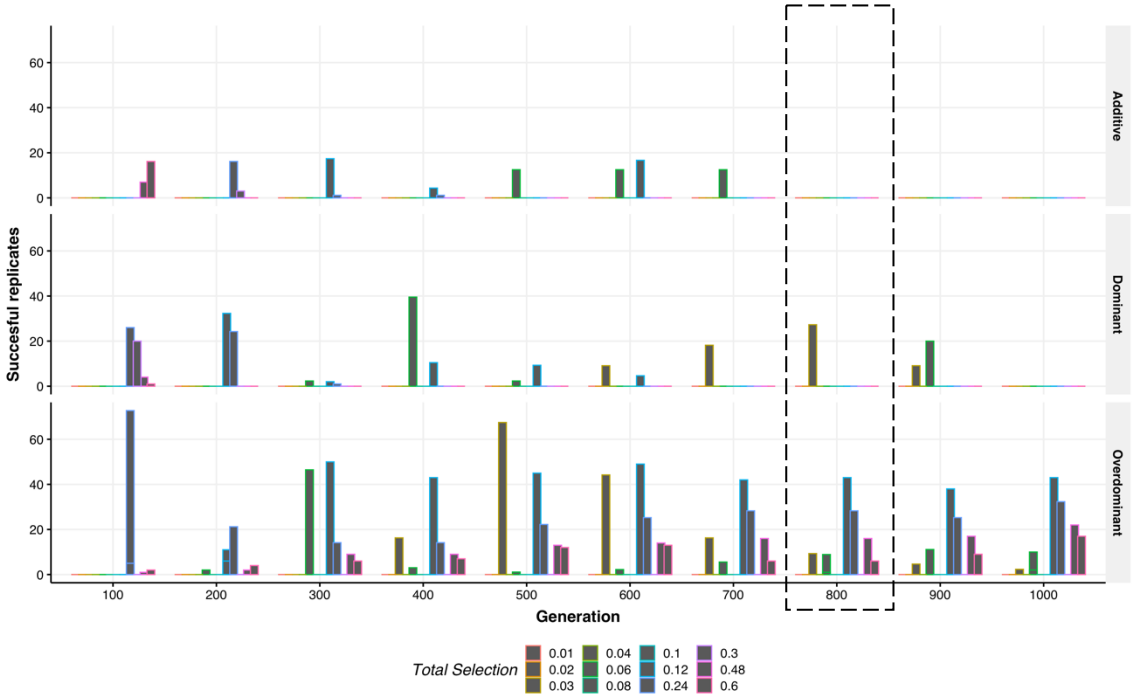
Supplementary Figure 3.18. Neutral simulations inside and outside the introgressed block. Mean of 100 simulated replicates for average background admixture (outside Δ100 kb region; blue dots) and within the introgressed Δ100 kb region (red dots) across neutral models. Values were retrieved and summarized for the last generation in the simulations (1200 generations) and for the parameters that recovered likely scenarios in neutral models of the genomic background (1 and 120 migration events; 1 and 30 migrants; Supplementary Figure 3.17). Empirical estimates of dog admixture within the Δ100 kb region and at genome-wide level are shown in red and black dashed lines, respectively.



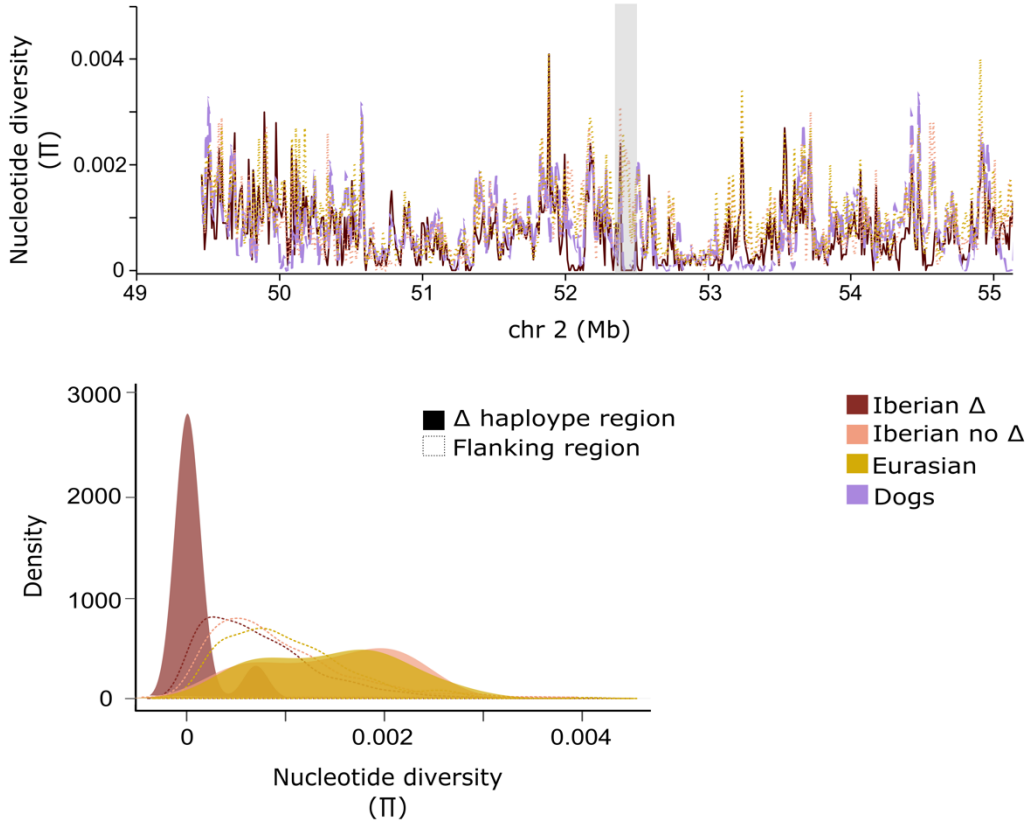
Supplementary Figure 3.19. Effects of recombination rate. (**A**) Recombination rate across chromosome 2 based on the dog genetic map. The red line indicates the position of the Δ100 kb region, and the position of the maximum and minimum values of recombination are indicated. (**B**) Boxplots representing the average simulated admixture in regions of the highest and lowest recombination rate in chromosome 2 across the neutral models at generation 800. No differences were detected.



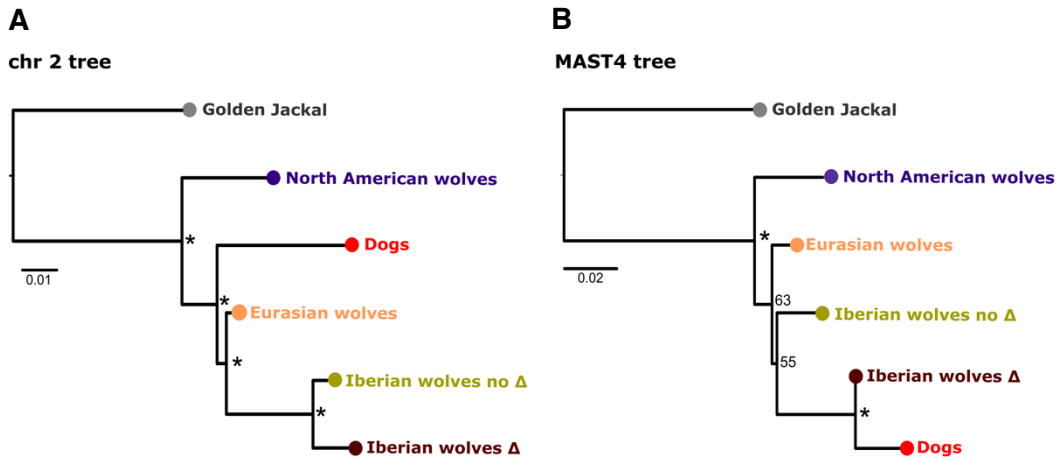
Supplementary Figure 3.20. Simulated genomic landscape of admixture. Proportion of the simulated admixture across chromosome 2 (in 50 kb windows) generated in SLiM, under a range of selection and dominance coefficients. The simulated patterns are compatible with those observed in the empirical data (empirical values of admixture inside Δ100 kb region are shown in dashed lines).



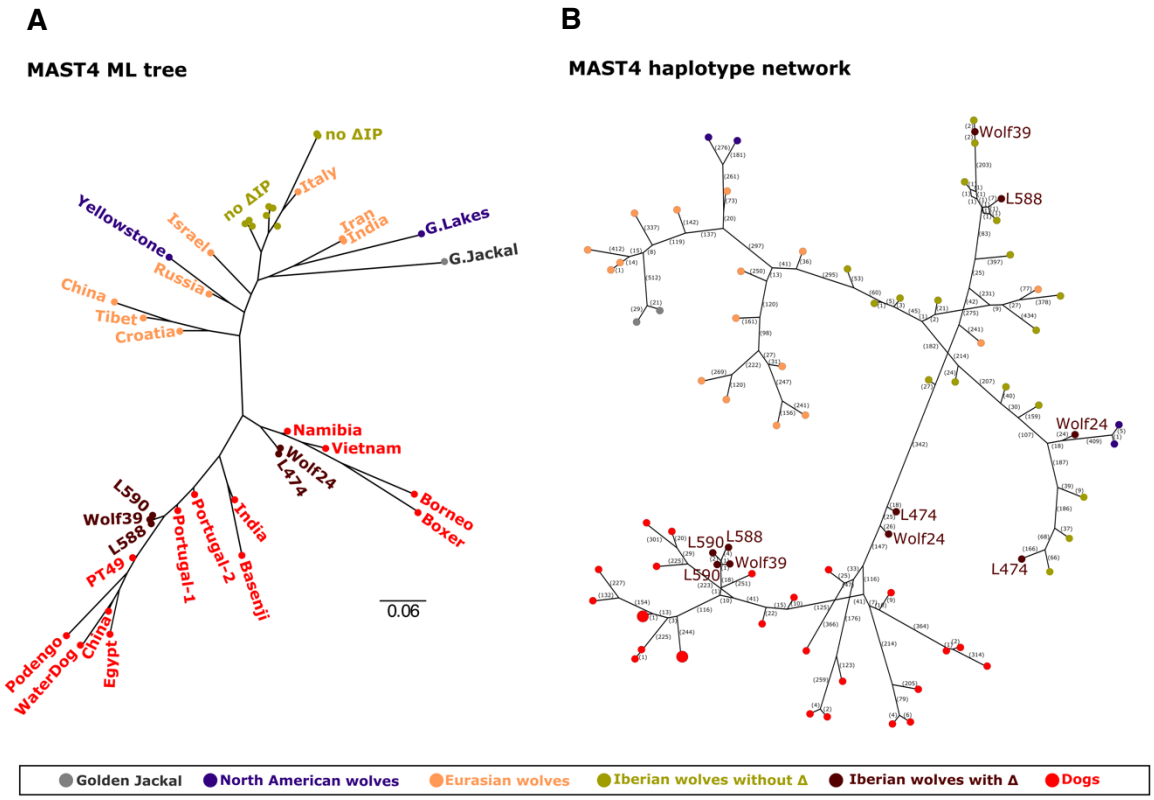
Supplementary Figure 3.21. Successful replicates in SLiM simulations. Number of replicates that successfully recovered the empirical patterns of admixture and allele frequencies per generation across 100 replicates for different models of selection. A range of simulated selection coefficients (*s* = 0-0.1) for 1, 3 or 6 introgressed SNPs resulted in a total amount of selection ranging from 0 to 0.6 (color gradient). The neutral model (selection = 0) is also represented. The dashed box highlights the results for the proposed time of the introgression (~800 generations).



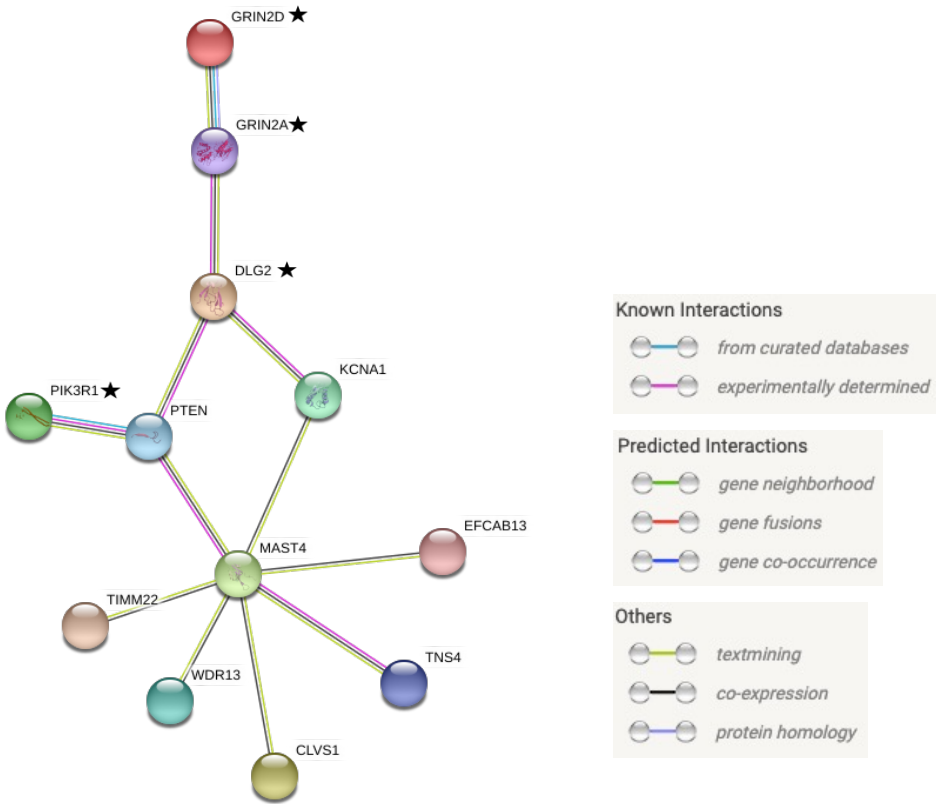
Supplementary Figure 3.22. Levels of nucleotide diversity (pi). The top panel shows the nucleotide diversity (pi) estimated in 10 kb non-overlapping windows along 6 Mb haplotypes surrounding Δ100 kb (represented by the gray bar) in each group (color coded). The bottom panel shows the distribution of pi in the Δ100 kb region (designated Δ haplotype) and in the 6 Mb flanking region in wolves.



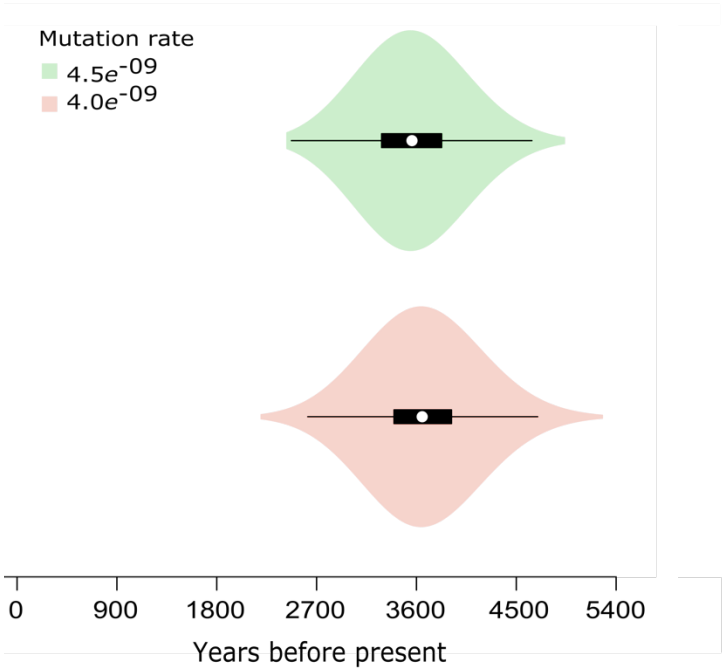
Supplementary Figure 3.23. Population trees. Maximum-likelihood population trees based on allele frequencies using SNPs spanning the (**A**) entire chromosome 2 and (**B**) the MAST4 gene, with bootstrap supports (* - 100), across the different groups.



Supplementary Figure 3.24. Phylogenetic relationships in MAST4 gene. (**A**) Maximum-likelihood tree constructed using SNPs spanning the entire MAST4 gene with single individuals. (**B**) Network of MAST4 haplotypes (only sample IDs from Iberian wolves with the Δblock are shown). The number of mutations is displayed in brackets.



Supplementary Figure 3.25. Protein-protein interaction network of MAST4. Protein associations around MAST4 in STRING, revealing several genes involved in glutamate receptor signaling (identified with a star). Colored lines between nodes indicate the type of interaction evidence (right legend) and colored nodes represent the queried proteins and first shell of interactors. Filled nodes have associated 3D structure that is known or predicted, while empty nodes do not have any known structure.



Supplementary Figure 3.26. Age distribution. Estimates of TMRCA of the 90 kb introgressed haplotype in Iberian wolves, using two mutation rates and assuming a generation time of 4.5 years. Violin plots represent the distribution of TMRCA from MCMC iterations.

Supplementary Table 3.1. Canine SNP chip data. Description of the full dataset used in this study. We genotyped 95 Iberian wolves, five Eurasian wolves, 23 Iberian village dogs and 39 purebred dogs. Additionally, we add data from 50 European wolves from Stronen et al. 2015 and 58 purebred dogs from the LUPA project.

Sample ID	Population/Breed	Group	Locality	Study
1	Maremma Sheepdog	Purebred dog	Italy	This study
2	Maremma Sheepdog	Purebred dog	Italy	This study
1308	Leonberger	Purebred dog	Germany	This study
1309	Leonberger	Purebred dog	Germany	This study
DA172	Great Dane	Purebred dog	Germany	This study
DA179	Great Dane	Purebred dog	Germany	This study
Kom660	Komondor	Purebred dog	Hungary	This study
SB513	Saint Bernard	Purebred dog	Central Europe	This study
SB514	Saint Bernard	Purebred dog	Central Europe	This study
CGT33	Cão de Gado Transmontano	Purebred dog	Iberian Peninsula	This study
CGT34	Cão de Gado Transmontano	Purebred dog	Iberian Peninsula	This study
CGT36	Cão de Gado Transmontano	Purebred dog	Iberian Peninsula	This study
CGT38	Cão de Gado Transmontano	Purebred dog	Iberian Peninsula	This study
CGT39	Cão de Gado Transmontano	Purebred dog	Iberian Peninsula	This study
CCL22	Castro Laboreiro	Purebred dog	Iberian Peninsula	This study
CCL32	Castro Laboreiro	Purebred dog	Iberian Peninsula	This study
CL04	Castro Laboreiro	Purebred dog	Iberian Peninsula	This study
CL5	Castro Laboreiro	Purebred dog	Iberian Peninsula	This study
CL11	Castro Laboreiro	Purebred dog	Iberian Peninsula	This study
ME134	Mastin Español	Purebred dog	Iberian Peninsula	This study
ME_Leon11	Mastin Español	Purebred dog	Iberian Peninsula	This study
ME257	Mastin Español	Purebred dog	Iberian Peninsula	This study
ME259	Mastin Español	Purebred dog	Iberian Peninsula	This study
ME_Avispado	Mastin Español	Purebred dog	Iberian Peninsula	This study
MoP361	Montaña del Pirineo	Purebred dog	Iberian Peninsula	This study
MoP577	Montaña del Pirineo	Purebred dog	Iberian Peninsula	This study
MoP579	Montaña del Pirineo	Purebred dog	Iberian Peninsula	This study
MoP583	Montaña del Pirineo	Purebred dog	Iberian Peninsula	This study
MoP581	Montaña del Pirineo	Purebred dog	Iberian Peninsula	This study
RA01	Rafeiro do Alentejo	Purebred dog	Iberian Peninsula	This study
RA05	Rafeiro do Alentejo	Purebred dog	Iberian Peninsula	This study
RA16	Rafeiro do Alentejo	Purebred dog	Iberian Peninsula	This study
RA22	Rafeiro do Alentejo	Purebred dog	Iberian Peninsula	This study
RA23	Rafeiro do Alentejo	Purebred dog	Iberian Peninsula	This study
SE_EMD004	Serra da Estrela	Purebred dog	Iberian Peninsula	This study
SE_EMD009	Serra da Estrela	Purebred dog	Iberian Peninsula	This study
SE_EMD016	Serra da Estrela	Purebred dog	Iberian Peninsula	This study
SE_EMD021	Serra da Estrela	Purebred dog	Iberian Peninsula	This study
SE_EMD030	Serra da Estrela	Purebred dog	Iberian Peninsula	This study
ASh_LU1	Australian Shepherd	Purebred dog	NA	[1,2]
BeT_GT25	Belgian Tervuren	Purebred dog	NA	[1,2]
BeT_GT27	Belgian Tervuren	Purebred dog	NA	[1,2]
Bgl_GT1	Beagle	Purebred dog	NA	[1,2]
Bgl_GT5	Beagle	Purebred dog	NA	[1,2]
BMD_GT30	Bernese Mountain Dog	Purebred dog	NA	[1,2]
BMD_GT31	Bernese Mountain Dog	Purebred dog	NA	[1,2]
BoC_GT55	Border Collie	Purebred dog	NA	[1,2]
BoC_GT56	Border Collie	Purebred dog	NA	[1,2]
BrS_GT76	Brittany Spaniel	Purebred dog	NA	[1,2]
BrS_GT83	Brittany Spaniel	Purebred dog	NA	[1,2]
Chi_LU135	Chihuahua	Purebred dog	NA	[1,2]
Chi_LU136	Chihuahua	Purebred dog	NA	[1,2]
CKC_GT88	Cavalier King Charles Spaniel	Purebred dog	NA	[1,2]
CKC_LU138	Cavalier King Charles Spaniel	Purebred dog	NA	[1,2]
CoS_GT91	Cocker Spaniel	Purebred dog	NA	[1,2]
CoS_GT93	Cocker Spaniel	Purebred dog	NA	[1,2]
Dac_GT111	Dachshund	Purebred dog	NA	[1,2]
Dac_GT117	Dachshund	Purebred dog	NA	[1,2]
Dal_GT122	Dalmatian	Purebred dog	NA	[1,2]
Dal_GT123	Dalmatian	Purebred dog	NA	[1,2]
ECS_LU437	English Cocker Spaniel	Purebred dog	NA	[1,2]
ECS_LU438	English Cocker Spaniel	Purebred dog	NA	[1,2]
Elk_GT128	Elkhound	Purebred dog	NA	[1,2]
Elk_GT129	Elkhound	Purebred dog	NA	[1,2]
ESS_LU598	English Springer Spaniel	Purebred dog	NA	[1,2]
ESS_LU600	English Springer Spaniel	Purebred dog	NA	[1,2]
ESt_GT162	English Setter	Purebred dog	NA	[1,2]
ESt_GT163	English Setter	Purebred dog	NA	[1,2]

FcR_GT198	Flatcoated Retriever	Purebred dog	NA	[1,2]
FcR_GT199	Flatcoated Retriever	Purebred dog	NA	[1,2]
GoS_LU604	Gordon Setter	Purebred dog	NA	[1,2]
GoS_LU606	Gordon Setter	Purebred dog	NA	[1,2]
GRe_GT214	Golden Retriever	Purebred dog	NA	[1,2]
GRe_GT217	Golden Retriever	Purebred dog	NA	[1,2]
Gry_GT239	Greyhound	Purebred dog	NA	[1,2]
Gry_GT240	Greyhound	Purebred dog	NA	[1,2]
GSh_GT200	German Shepherd	Purebred dog	NA	[1,2]
GSh_GT201	German Shepherd	Purebred dog	NA	[1,2]
IrW_LU1210	Irish Wolfhound	Purebred dog	NA	[1,2]
IrW_LU1211	Irish Wolfhound	Purebred dog	NA	[1,2]
JRT_GT250	Jack Russell Terrier	Purebred dog	NA	[1,2]
JRT_GT251	Jack Russell Terrier	Purebred dog	NA	[1,2]
LMu_GT279	Large Munsterlander	Purebred dog	NA	[1,2]
LRe_GT264	Labrador Retriever	Purebred dog	NA	[1,2]
LRe_GT266	Labrador Retriever	Purebred dog	NA	[1,2]
Mop_GT280	Mops	Purebred dog	NA	[1,2]
Mop_GT282	Mops	Purebred dog	NA	[1,2]
Pdl_GT328	Poodle	Purebred dog	NA	[1,2]
Pdl_GT345	Poodle	Purebred dog	NA	[1,2]
Rtw_GT348	Rottweiler	Purebred dog	NA	[1,2]
Rtw_GT349	Rottweiler	Purebred dog	NA	[1,2]
Scn_LU1563	Schnauzer	Purebred dog	NA	[1,2]
Scn_LU1565	Schnauzer	Purebred dog	NA	[1,2]
StP_GT376	Standard Poodle	Purebred dog	NA	[1,2]
StP_GT379	Standard Poodle	Purebred dog	NA	[1,2]
TYo_GT388	Yorkshire Terrier	Purebred dog	NA	[1,2]
TYo_GT390	Yorkshire Terrier	Purebred dog	NA	[1,2]
CR2	Village dog	Village dog	Iberian Peninsula	This study
CR9	Village dog	Village dog	Iberian Peninsula	This study
CR11	Village dog	Village dog	Iberian Peninsula	This study
CR14	Village dog	Village dog	Iberian Peninsula	This study
CR35	Village dog	Village dog	Iberian Peninsula	This study
CR39	Village dog	Village dog	Iberian Peninsula	This study
CR44	Village dog	Village dog	Iberian Peninsula	This study
CR63	Village dog	Village dog	Iberian Peninsula	This study
CR64	Village dog	Village dog	Iberian Peninsula	This study
CR65	Village dog	Village dog	Iberian Peninsula	This study
CR67	Village dog	Village dog	Iberian Peninsula	This study
CR68	Village dog	Village dog	Iberian Peninsula	This study
CR77	Village dog	Village dog	Iberian Peninsula	This study
CR80	Village dog	Village dog	Iberian Peninsula	This study
CR97	Village dog	Village dog	Iberian Peninsula	This study
CR98	Village dog	Village dog	Iberian Peninsula	This study
CR104	Village dog	Village dog	Iberian Peninsula	This study
L400	Village dog	Village dog	Iberian Peninsula	This study
L501	Village dog	Village dog	Iberian Peninsula	This study
L624	Village dog	Village dog	Iberian Peninsula	This study
L652	Village dog	Village dog	Iberian Peninsula	This study
L744	Village dog	Village dog	Iberian Peninsula	This study
L791	Village dog	Village dog	Iberian Peninsula	This study
L005	Iberian Wolf	Iberian Wolf	Bragança, PT	This study
L010	Iberian Wolf	Iberian Wolf	Bragança, PT	This study
L029	Iberian Wolf	Iberian Wolf	Vila Real, PT	This study
L053	Iberian Wolf	Iberian Wolf	Bragança, PT	This study
L063	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L086	Iberian Wolf	Iberian Wolf	Viseu, PT	This study
L090	Iberian Wolf	Iberian Wolf	Bragança, PT	This study
L091	Iberian Wolf	Iberian Wolf	Bragança, PT	This study
L102	Iberian Wolf	Iberian Wolf	Bragança, PT	This study
L116	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L125	Iberian Wolf	Iberian Wolf	Viana do Castelo, PT	This study
L180	Iberian Wolf	Iberian Wolf	Asturias, Spain	This study
L186	Iberian Wolf	Iberian Wolf	Asturias, Spain	This study
L188	Iberian Wolf	Iberian Wolf	Asturias, Spain	This study
L190	Iberian Wolf	Iberian Wolf	Asturias, Spain	This study
L195	Iberian Wolf	Iberian Wolf	Asturias, Spain	This study
L200	Iberian Wolf	Iberian Wolf	Asturias, Spain	This study
L212	Iberian Wolf	Iberian Wolf	Asturias, Spain	This study
L253	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L262	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L286	Iberian Wolf	Iberian Wolf	Castilla y Leon, Spain	This study
L298	Iberian Wolf	Iberian Wolf	Vila Real, PT	This study
L316	Iberian Wolf	Iberian Wolf	Viana do Castelo, PT	This study

L319	Iberian Wolf	Iberian Wolf	Castilla y Leon, Spain	This study
L320	Iberian Wolf	Iberian Wolf	Castilla y Leon, Spain	This study
L322	Iberian Wolf	Iberian Wolf	Viana do Castelo, PT	This study
L337	Iberian Wolf	Iberian Wolf	Viana do Castelo, PT	This study
L338	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L351	Iberian Wolf	Iberian Wolf	Asturias, Spain	This study
L357	Iberian Wolf	Iberian Wolf	Asturias, Spain	This study
L360	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L363	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L364	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L365	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L367	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L370	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L382	Iberian Wolf	Iberian Wolf	Asturias, Spain	This study
L384	Iberian Wolf	Iberian Wolf	Asturias, Spain	This study
L387	Iberian Wolf	Iberian Wolf	Asturias, Spain	This study
L392	Iberian Wolf	Iberian Wolf	Viana do Castelo, PT	This study
L401	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L409	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L410	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L419	Iberian Wolf	Iberian Wolf	Asturias, Spain	This study
L422	Iberian Wolf	Iberian Wolf	Viseu, PT	This study
L429	Iberian Wolf	Iberian Wolf	Guarda, PT	This study
L431	Iberian Wolf	Iberian Wolf	Asturias, Spain	This study
L446	Iberian Wolf	Iberian Wolf	Asturias, Spain	This study
L474	Iberian Wolf	Iberian Wolf	Bragança, PT	[3]
L475	Iberian Wolf	Iberian Wolf	Guarda, PT	This study
L502	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L509	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L511	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L514	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L546	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L547	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L552	Iberian Wolf	Iberian Wolf	Asturias, Spain	This study
L556	Iberian Wolf	Iberian Wolf	Asturias, Spain	This study
L577	Iberian Wolf	Iberian Wolf	Castilla y Leon, Spain	This study
L580	Iberian Wolf	Iberian Wolf	Castilla y Leon, Spain	This study
L582	Iberian Wolf	Iberian Wolf	Castilla y Leon, Spain	This study
L583	Iberian Wolf	Iberian Wolf	Castilla y Leon, Spain	This study
L585	Iberian Wolf	Iberian Wolf	Castilla y Leon, Spain	[3]
L586	Iberian Wolf	Iberian Wolf	Castilla y Leon, Spain	This study
L588	Iberian Wolf	Iberian Wolf	Castilla y Leon, Spain	This study
L589	Iberian Wolf	Iberian Wolf	Castilla y Leon, Spain	This study
L590	Iberian Wolf	Iberian Wolf	Castilla y Leon, Spain	This study
L596	Iberian Wolf	Iberian Wolf	Vila Real, PT	This study
L619	Iberian Wolf	Iberian Wolf	Asturias, Spain	This study
L625	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L639	Iberian Wolf	Iberian Wolf	Asturias, Spain	This study
L649	Iberian Wolf	Iberian Wolf	Asturias, Spain	This study
L650	Iberian Wolf	Iberian Wolf	Asturias, Spain	This study
L679	Iberian Wolf	Iberian Wolf	Vila Real, PT	This study
L713	Iberian Wolf	Iberian Wolf	Castilla y Leon, Spain	This study
L768	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L772	Iberian Wolf	Iberian Wolf	Viseu, PT	This study
L780	Iberian Wolf	Iberian Wolf	Viana do Castelo, PT	This study
L781	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L783	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L788	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L790	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L795	Iberian Wolf	Iberian Wolf	Viana do Castelo, PT	This study
L798	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L800	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L801	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L802	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L804	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L805	Iberian Wolf	Iberian Wolf	Galicia, Spain	This study
L806	Iberian Wolf	Iberian Wolf	Asturias, Spain	This study
L819	Iberian Wolf	Iberian Wolf	Castilla y Leon, Spain	This study
L820	Iberian Wolf	Iberian Wolf	Castilla y Leon, Spain	This study
L825	Iberian Wolf	Iberian Wolf	Asturias, Spain	This study
L844	Iberian Wolf	Iberian Wolf	Castilla y Leon, Spain	This study
L845	Iberian Wolf	Iberian Wolf	Castilla y Leon, Spain	This study
5	European Wolf	Eurasian Wolf	Poland	[4]
15	European Wolf	Eurasian Wolf	Belarus	[4]

L112	European Wolf	Eurasian Wolf	Romania	This study
L114	European Wolf	Eurasian Wolf	Romania	This study
L525	European Wolf	Eurasian Wolf	Slovenia	This study
L527	European Wolf	Eurasian Wolf	Slovenia	This study
L743	Asiatic Wolf	Eurasian Wolf	Iran	This study
27	European Wolf	Eurasian Wolf	Belarus	[4]
52	European Wolf	Eurasian Wolf	Belarus	[4]
57	European Wolf	Eurasian Wolf	Belarus	[4]
62	European Wolf	Eurasian Wolf	Latvia	[4]
103	European Wolf	Eurasian Wolf	Belarus	[4]
112	European Wolf	Eurasian Wolf	Belarus	[4]
128	European Wolf	Eurasian Wolf	Latvia	[4]
132	European Wolf	Eurasian Wolf	Latvia	[4]
133	European Wolf	Eurasian Wolf	Belarus	[4]
163	European Wolf	Eurasian Wolf	Russia	[4]
203	European Wolf	Eurasian Wolf	Belarus	[4]
206	European Wolf	Eurasian Wolf	Belarus	[4]
378	European Wolf	Eurasian Wolf	Ukraine	[4]
424	European Wolf	Eurasian Wolf	Russia	[4]
444	European Wolf	Eurasian Wolf	Russia	[4]
527	European Wolf	Eurasian Wolf	Ukraine	[4]
548	European Wolf	Eurasian Wolf	Ukraine	[4]
562	European Wolf	Eurasian Wolf	Ukraine	[4]
563	European Wolf	Eurasian Wolf	Ukraine	[4]
689	European Wolf	Eurasian Wolf	Poland	[4]
846	European Wolf	Eurasian Wolf	Poland	[4]
987	European Wolf	Eurasian Wolf	Poland	[4]
988	European Wolf	Eurasian Wolf	Poland	[4]
1003	European Wolf	Eurasian Wolf	Bulgaria	[4]
1135	European Wolf	Eurasian Wolf	Latvia	[4]
1148	European Wolf	Eurasian Wolf	Latvia	[4]
1150	European Wolf	Eurasian Wolf	Latvia	[4]
1174	European Wolf	Eurasian Wolf	Croatia	[4]
1219	European Wolf	Eurasian Wolf	Croatia	[4]
1262	European Wolf	Eurasian Wolf	Finland	[4]
1263	European Wolf	Eurasian Wolf	Finland	[4]
1267	European Wolf	Eurasian Wolf	Finland	[4]
1271	European Wolf	Eurasian Wolf	Finland	[4]
1282	European Wolf	Eurasian Wolf	Finland	[4]
1344	European Wolf	Eurasian Wolf	Ukraine	[4]
1345	European Wolf	Eurasian Wolf	Ukraine	[4]
1382	European Wolf	Eurasian Wolf	Greece	[4]
1387	European Wolf	Eurasian Wolf	Greece	[4]
895	European Wolf	Eurasian Wolf	Bulgaria	[4]
1032	European Wolf	Eurasian Wolf	Bulgaria	[4]
1224	European Wolf	Eurasian Wolf	Croatia	[4]
833	European Wolf	Eurasian Wolf	Poland	[4]
897	European Wolf	Eurasian Wolf	Poland	[4]
1381	European Wolf	Eurasian Wolf	Poland	[4]
574	European Wolf	Eurasian Wolf	Russia	[4]
584	European Wolf	Eurasian Wolf	Russia	[4]
553	European Wolf	Eurasian Wolf	Ukraine	[4]
347	European Wolf	Eurasian Wolf	Slovakia	[4]

[1] – Lequarré et al. 2011; [2] – Vaysse et al. 2011; [3] – Pacheco, Lobo et al. 2022; Chapter II; [4] – Stronen et al. 2015.
Abbreviation: PT, Portugal; NA, not available.

Supplementary Table 3.2. WGS data. Description of the 39 canid genomes used in this study. We generated 12 Iberian wolf genomes and added additional publicly available data from 11 Eurasian wolves, 2 North American wolves, 13 dogs and 1 golden jackal, from several studies. Individual information about the mean depth and missing data across the final dataset is also provided.

Sample ID	Population	Locality	BioSample	Study	Mean Depth	Missing (%)
L090	Iberian Wolf	Portugal		This study	12.58	0.02
L351	Iberian Wolf	Spain		This study	11.29	0.08
L474	Iberian Wolf	Portugal		This study	11.13	0.10
L546	Iberian Wolf	Spain		This study	14.80	0.06
L552	Iberian Wolf	Spain		This study	13.29	0.06
L590	Iberian Wolf	Spain		This study	13.87	0.07
L253	Iberian Wolf	Spain		This study	13.09	0.08
L409	Iberian Wolf	Spain		This study	14.23	0.06
L514	Iberian Wolf	Spain		This study	13.38	0.07
L547	Iberian Wolf	Spain		This study	14.21	0.08
L588	Iberian Wolf	Spain		This study	15.65	0.07
L844	Iberian Wolf	Spain		This study	15.42	0.01
Wolf24	Iberian Wolf	Portugal	SAMN02921316	[7]	26.37	0.15
Wolf27	Iberian Wolf	Spain	SAMN02921319	[7]	25.93	0.20
Wolf39	Iberian Wolf	Spain	SAMN04851099	[9]	37.99	0.05
Wolf21	European Wolf	Italy	SAMN02921313	[7]	6.01	2.50
Wolf06	European Wolf	Croatia	SAMN03366712	[6]	6.74	3.54
Wolf01	Asian Wolf	Russia	SAMN01974486	[5]	11.56	1.11
Wolf05	Asian Wolf	China	SAMN03366711	[6]	28.09	0.08
Wolf07	Asian Wolf	Israel	SAMN03366710	[6]	5.84	3.53
WolfTibetan	Asian Wolf	Tibet	SAMN03653004	[8]	24.12	0.17
Wolf19	Asian Wolf	India	SAMN02921311	[7]	27.11	0.06
Wolf20	Asian Wolf	Iran	SAMN02921312	[7]	28.40	0.08
Wolf28	N-A Wolf	Yellowstone	SAMN02921320	[7]	27.54	0.08
Wolf18	N-A Wolf	Great Lakes	SAMN02921310	[7]	26.56	0.09
PT49	Village dog	Portugal	SAMN04851101	[10]	14.88	0.18
VillDog_PT01	Village dog	Portugal	SAMN02485604	[10]	15.30	0.18
VillDog_PT02	Village dog	Portugal	SAMN02485605	[10]	13.90	0.26
PT_Podengo	Podengo	Portugal	SAMN03580388	[11]	28.04	0.04
PT_WaterDog	Water Dog	Portugal	SAMN08873219	[12]	42.61	0.04
Boxer	Boxer	NA	SAMN02921305	[6]	23.83	12.17
CFA.118003	Basenji	China	SAMN06159689	[11]	39.81	0.06
VillDog_Borneo	Village dog	Borneo	SAMN02485592	[10]	4.74	5.60
VillDog_China	Village dog	China	SAMN01974490	[5]	17.81	5.51
VillDog_Egypt	Village dog	Egypt	SAMN02485582	[10]	6.89	1.40
VillDog_India	Village dog	India	SAMN02485586	[10]	12.85	0.64
VillDog_Namibia	Village dog	Namibia	SAMN02485598	[10]	7.69	1.80
VillDog_Vietnam	Village dog	Vietnam	SAMN02485609	[10]	8.44	2.44
Golden Jackal	Asian Jackal	Israel	SAMN03366713	[6]	4.84	5.24

[5] – Wang et al. 2013; [6] – Freedman et al. 2014; [7] – Fan et al. 2016; [8] – Zhang et al. 2014; [9] – Botigue et al. 2017; [10] – Auton et al. 2013; [11] – Marsden et al. 2016; [12] – Plassais et al. 2019.
Abbreviation: N-A, North American; NA, not available.

Supplementary Table 3.3. Genes within Δblock. Description of the 22 protein-coding genes located within Δblock. This information is accordingly to the CanFam3.1 Ensembl annotation.

Gene ID	Gene Description	Start	End
HTR1A	5-hydroxytryptamine receptor 1A	50,002,740	50,006,548
RNF180	ring finger protein 180	50,197,137	50,352,946
RGS7BP	regulator of G protein signaling 7 binding protein	50,430,844	50,528,673
SHISAL2B	shisa like 2B	50,590,487	50,606,452
SREK1IP1	srek1 interecting protein 1	50,612,272	50,654,746
CWC27	cwc27 spliceosome associated cyclophillin	50,655,037	50,846,604
ENSCAFG00000042469		50,828,607	50,831,060
ADAMTS6	ADAM metallopeptidase with thrombospondin type 1 motif 6	50,955481	51,236,421
CENPK	centromere protein K	51,263,185	51,292,425
PPWD1	peptidylpropyl isomerase domain and WD repeat containing 1	51,300,182	51,320,989
TRIM23	tripartite motif containing 23	51,325,038	51,360,117
TRAPPC13	trafficking protein particle complex 13	51,360,593	51,398,210
SGTB	small glutamine rich tetratricopeptide repeat	51,402,856	51,445,278
NLN	neurolysin	51,444,566	51,538,028
ERBIN	erbb2 interacting protein	51,613,230	51,729,021
ENSCAFG00000043906		51,670,843	51,671,704
SREK1	splicing regulatory glutamic acid and lysine rich protein 1	51,761,763	51,802,694
MAST4	microtubule associated serine threonine kinase family member 4	52,185,407	52,633,243
ENSCAFG00000042414		52,307,006	52,285,621
ENSCAFG00000046880		52,432,389	52,459,478
CD180	CD180 molecule	52,645,766	53,660,790
ENSCAFG00000049004		52,910,756	52,912,978

Supplementary Table 3.4. Historical samples. Description of the 48 historical Iberian wolf samples used and generated in a parallel study conducted by Pacheco, Lobo et al. 2022 (Chapter II).

Sample ID	Population	Year	Locality	Study
CL-H010	Iberian Wolf	1975	Castilla y Leon, Spain	[3]
CL-H017	Iberian Wolf	1972	Castilla y Leon, Spain	[3]
CL-H020	Iberian Wolf	1974	Castilla y Leon, Spain	[3]
CL-H022	Iberian Wolf	1976	Castilla y Leon, Spain	[3]
CL-H027	Iberian Wolf	1983	Castilla y Leon, Spain	[3]
CL-H029	Iberian Wolf	1983	Castilla y Leon, Spain	[3]
CL-H034	Iberian Wolf	1984	Castilla y Leon, Spain	[3]
CL-H035	Iberian Wolf	1986	Castilla y Leon, Spain	[3]
CL-H036	Iberian Wolf	1984	Castilla y Leon, Spain	[3]
CL-H038	Iberian Wolf	1984	Castilla y Leon, Spain	[3]
CL-H040	Iberian Wolf	1983	Castilla y Leon, Spain	[3]
CL-H042	Iberian Wolf	1982	Castilla y Leon, Spain	[3]
CL-H052	Iberian Wolf	1979	Guarda, PT	[3]
CL-H083	Iberian Wolf	1972	Castelo Branco, PT	[3]
CL-H084	Iberian Wolf	1972	Castelo Branco, PT	[3]
CL-H086	Iberian Wolf	1912	Castelo Branco, PT	[3]
CL-H093	Iberian Wolf	1930	Castelo Branco, PT	[3]
CL-H100	Iberian Wolf	1962	Beja, PT	[3]
CL-H103	Iberian Wolf	1959	Beja, PT	[3]
CL-H105	Iberian Wolf	1959	Portalegre, PT	[3]
CL-H116	Iberian Wolf	1927	Jaén, Spain	[3]
CL-H117	Iberian Wolf	1970	Castilla la Mancha, Spain	[3]
CL-H119	Iberian Wolf	1944	Extremadura, Spain	[3]
CL-H126	Iberian Wolf	1973	Setubal, PT	[3]
CL-H129	Iberian Wolf	1991	Braga, PT	[3]
CL-H131	Iberian Wolf	1987	Viana do Castelo, PT	[3]
CL-H132	Iberian Wolf	1955	Braga, PT	[3]
CL-H135	Iberian Wolf	1945	Vila Real, PT	[3]
CL-H137	Iberian Wolf	NA	Vila Real, PT	[3]
CL-H139	Iberian Wolf	1980	Bragança, PT	[3]
CL-H140	Iberian Wolf	1980	Bragança, PT	[3]
CL-H142	Iberian Wolf	1960	Bragança, PT	[3]
CL-H143	Iberian Wolf	1970	Bragança, PT	[3]
CL-H145	Iberian Wolf	1980	Bragança, PT	[3]
CL-H146	Iberian Wolf	1980	Bragança, PT	[3]
CL-H148	Iberian Wolf	1980	Bragança, PT	[3]
CL-H149	Iberian Wolf	1970	Braga, PT	[3]
CL-H177	Iberian Wolf	1984	Castilla y Leon, Spain	[3]
CL-H182	Iberian Wolf	1964	Jaén, Spain	[3]
CL-H188	Iberian Wolf	1970	Asturias, Spain	[3]
CL-H197	Iberian Wolf	1969	Castilla y Leon, Spain	[3]
CL-H200	Iberian Wolf	1974	Galicia, Spain	[3]
CL-H220	Iberian Wolf	1976	Castilla y Leon, Spain	[3]
CL-H234	Iberian Wolf	1969	Galicia, Spain	[3]
CL-H241	Iberian Wolf	1967	Castilla y Leon, Spain	[3]
CL-H261	Iberian Wolf	1991	Sevilla, Spain	[3]
CL-H268	Iberian Wolf	1989	Castilla y Leon, Spain	[3]
CL-H272	Iberian Wolf	2005	Castilla y Leon, Spain	[3]

[3] – Pacheco, Lobo et al. 2022; Chapter II.
Abbreviation: PT, Portugal; NA, not available

Supplementary Table 3.5. Outlier regions in f_d analysis. Top 1% outlier windows in f_d statistics (fraction of introgression) along chromosome 2, using Iberian wolves without the Δblock or Eurasian wolves as P1, Iberian wolves with the Δblock as P2, and dogs as P3 (one golden jackal set as outgroup). Windows overlapping Δ100 kb are highlighted in gray. Results using 50 and 100 kb windows are presented showing the start and end position in chromosome 2 (Mb) and f_d value.

P1 = Iberian wolves without Δblock						**P1 = Eurasian wolves**					
50 kb windows			100 kb windows			50 kb windows			100 kb windows		
START	END	f_d	START	END	f_d	START	END	f_d	START	END	f_d
52.40	52.45	0.81	52.40	52.50	0.77	52.10	52.15	0.81	53.00	53.10	0.75
48.55	48.60	0.81	55.00	55.10	0.69	44.45	44.50	0.80	52.10	52.20	0.73
50.05	50.10	0.74	48.50	48.60	0.68	45.40	45.45	0.80	44.40	44.50	0.67
55.05	55.10	0.74	53.10	53.20	0.67	38.00	38.50	0.75	51.60	51.70	0.66
48.65	48.70	0.73	54.50	54.60	0.67	51.65	51.70	0.72	52.40	52.50	0.66
52.10	52.15	0.73	52.10	52.20	0.65	68.90	68.95	0.71	53.10	53.20	0.65
52.45	52.50	0.72	45.00	45.10	0.63	53.15	53.20	0.71	48.70	48.80	0.63
44.45	44.50	0.71				46.10	46.15	0.71	63.00	64.00	0.63
53.25	53.30	0.70				51.35	51.40	0.70			
54.55	54.60	0.68				52.40	52.45	0.69			
53.15	53.20	0.68				53.25	53.30	0.67			
45.10	45.15	0.68				52.15	52.20	0.66			
45.05	45.10	0.67				48.70	48.75	0.66			

Supplementary Table 3.6. SNPs with high allele sharing between Iberian wolves with the ∆block and dogs. SNPs presenting increased allele sharing between Iberian wolves with the ∆block and dogs within the ∆100 kb region. The position of each SNP is provided as the allele found in high frequency in dogs and Iberian wolves carrying the ∆block (designated "∆ allele"). The alternative allele is also described. Frequencies for the ∆allele are presented.

Position	∆ allele (dog)	Alternative allele	Freq dogs (N=13)	Freq Iberian wolves ∆ (N=5)	Freq wolves no ∆ (N=19)
52,410,241	A	G	0.62	0.60	0.05
52,412,955	C	T	0.73	0.70	0.05
52,415,653	A	G	0.73	0.50	0.05
52,439,949	T	C	0.88	0.70	0.03
52,439,986	A	G	0.88	0.70	0.05
52,442,711	A	G	0.81	0.60	0.08
52,444,107	T	C	0.77	0.60	0.05
52,449,886	T	C	0.73	0.60	0.03
52,458,657	T	G	0.58	0.60	0.05
52,465,058	T	C	0.58	0.60	0.05
52,465,071	A	G	0.58	0.60	0.05
52,470,773	A	G	0.62	0.60	0.03
52,474,634	T	C	0.58	0.60	0.05
52,474,659	G	C	0.58	0.60	0.05
52,474,724	G	C	0.62	0.60	0.03
52,474,727	C	G	0.62	0.60	0.05
52,478,735	T	C	0.54	0.60	0.03
52,486,200	G	T	0.79	0.60	0.05

Chapter IV.

Brain transcriptomes uncover a possible genetic basis for the behavioral responses of Iberian wolves to human-dominated landscapes

Lobo D. et al.

(in preparation to be submitted)

Abstract

Following the advent of agricultural societies at least 10 thousand years ago (ka), humans have been drastically altering landscapes, which has challenged species to find evolutionary mechanisms to cope with these changes. The Iberian wolf subspecies is an example of a large carnivore that has managed to persist in highly human-dominated landscapes despite intense human persecution. The Iberian wolf population presents behavioral traits, such as reduced dispersal ability and increased tolerance levels towards humans and their activities, which we believe may be the product of adaptation to co-exist with humans through a process of self-domestication. Given that phenotypic evolution during animal domestication is believed to be strongly linked to changes in gene regulation, we addressed the self-domestication hypothesis by generating complete brain transcriptomes of four Iberian wolves. Their gene expression profiles were then compared to those of non-Iberian wolves and domestic dogs. Expression analysis revealed 4204 differentially expressed genes (DEGs) between Iberian and non-Iberian wolves, including candidate genes associated with key pathways for neurological processing, such as serotonin and glutamate genes, and genes involved in locomotory behavior. These genes have been widely associated with the domestic behavioral phenotype in several mammalian species. The direction of expression of some of these candidate genes was similar in Iberian wolves and dogs, pointing to reduced anxiety, increased cognitive abilities, and reduced exploratory tendencies. In addition, we also found overlapping signatures of allele frequency changes and expressional variation in genes implicated in the metabolism of glutamate and in genes with the potential to affect mobility. These results suggest an important role of neurotransmission in the possible adaptation of Iberian wolves to human-dominated landscapes, which seems to have favored behavioral traits similar to those selected during dog domestication, leading to a process of self-domestication.

4.1. Introduction

The development of human agricultural societies, fostered by the domestication of plants and animals during the Neolithic period, triggered extensive and enduring consequences for species as a result of dramatically altered landscapes (Zeder 2008). Species have been challenged to find mechanisms to adapt to new ecological, evolutionary, and environmental circumstances (Sullivan et al. 2017), and anthropogenic selection has rapidly altered their distribution, behavior, and morphology (Yeh and Price 2004; Hunter 2007; Beckman et al. 2022). Thus, understanding the mechanisms that drove species adaptation to increasingly human-modified landscapes is of utmost urgency.

Among mammals, gray wolves from the Iberian Peninsula (*C. l. signatus*) are a particularly good model to address this question. Iberian wolves co-exist with humans for millennia which may have resulted in the unique set of behavioral traits they exhibit (e.g., low dispersal ability, Silva et al. 2018; diet shift towards human-related resources, Llaneza and López-Bao 2015; Newsome et al. 2016; and high tolerance towards humans and their activities, Llaneza et al. 2016; Dennehy et al. 2021), and which possibly have facilitated their persistence in the highly human-dominated landscape of Iberian Peninsula. Despite the intensive human persecution suffered in the last centuries, and contrary to most of the Central European wolf populations, Iberian wolves have never been on the brink of extinction. To date, little is known about the specific ecological, environmental, and evolutionary mechanisms that equipped Iberian wolves differently. In previous work, we have observed that the Iberian wolf genome has been enriched by the introgression of domestic dog variants in a gene associated with several neurological processes (Lobo et al. Chapter III), which we believe can be partially contributing to their adaptation to co-exist with humans. Adaptation could additionally result from a natural evolutionary process towards traits favoring persistence. These could ultimately resemble some of the traits selected during dog domestication, such as the absence of dispersal behavior, leading to an ongoing process of self-domestication (i.e., a selective process against aggressive behavior in mammals, applied to wolves as favoring behaviors to co-exist with humans, Hare et al. 2012).

Domestication itself can be considered a special case of adaptation to human-induced environmental change (Beckman et al. 2022). Behavioral shifts towards reduced aggressiveness, as well as increased social tolerance, towards humans were most certainly a primary target of selection during domestication and the first phenotypic

changes to evolve in domestic animals (Sato et al. 2020; Andersson and Purugganan 2022). So far, no obvious “domestication gene” has been functionally associated with the behavior of domestic animals. There are, however, parallel changes in sets of genes, gene families, and functional pathways in several domestic bird and mammal species (e.g., chickens, dogs, cattle, pigs, rabbits, horses, and cats) associated with neurotransmission, synaptic plasticity and exploratory behavior (Hou et al. 2020; O’Rourke and Boeckx 2020). The genomes of several pairs of domestic and wild species have been extensively investigated and compared to detect signatures of shifts in allele frequencies that could be indicative of artificial selection (Rubin et al. 2010; Kijas et al. 2012; Rubin et al. 2012; Axelsson et al. 2013). However, highly differentiated genomic regions have been repeatedly located within noncoding regions, suggesting that gene regulation, and not mutations within protein-coding sequences, are likely to be underlying the phenotypic changes observed across domesticated animals (Li et al. 2014; Sato et al. 2020; Andersson and Purugganan 2022).

To investigate whether the evolutionary process underlying the behavioral traits observed in Iberian wolves mirrors the process of domestication, we used RNA-seq to generate the first-ever brain transcriptomes of four wild wolves living in their natural habitat. The four Iberian wolf brain transcriptomes were then used in conjunction with data from the only available current study in previous literature, where the complete brain transcriptomes of (captive) wolves were analyzed (Albert et al. 2012). We evaluated gene expression profiles in the frontal cortex of Iberian wolves and estimated patterns of differential gene expression using the available RNA-seq data from non-Iberian wolves and dogs (Broad Institute 2019). Differentially expressed genes (DEGs) were then characterized and intersected with genomic signatures of selection in Iberian wolves. With this comparative approach, we pinpointed a possible genetic basis for the behavioral responses in Iberian wolves to human-dominated landscapes.

4.2. Results and Discussion

4.2.1. Gene expression profiles in the brain of wolves and dogs

We sequenced the complete frontal cortex transcriptome of four wild-caught Iberian wolves at an average of 30 million reads (Supplementary Tables 4.1 and 4.2). As a control, we used a set of publicly available RNA-seq data generated from frontal cortex tissues of six captive gray wolves (hereafter non-Iberian; Supplementary Table 4.1;

Albert et al. 2012). To have a reference for the direction of expression in candidate genes associated with the domestic behavioral phenotype, we also added to our study publicly available RNA-seq data from the frontal cortex of five dogs (Supplementary Table 4.1; Broad Institute 2019). All the samples were aligned to the dog reference genome CanFam3.1 (Lindblad-Toh et al. 2005) and gene abundance was estimated for a reference set of 30,951 genes.

A Principal Component Analysis (PCA) using normalized count data showed that individuals from the three reference groups cluster together in their respective groups. The highest amount of variation was found between Iberian and non-Iberian wolves (87% in PC1; **Figure 4.1A**), while similar levels of variation were found in the expression profiles between dogs and Iberian and non-Iberian wolves, 74% and 73% in PC1, respectively (**Figure 4.1A**). These results seem to suggest that gene expression profiles in the frontal cortex were not mostly explained by evolutionary distance in wolves and dogs, as previously suggested (Saetre et al. 2004), and might be instead greatly influenced by the environment where individuals were inserted (Carroll 2005; Gibson 2008; Pavey et al. 2010; Rivera et al. 2021). Given information regarding the original population of non-Iberian wolves was unknown, we reconstructed the phylogenetic relationships of our samples using 79,267 exome-wide SNPs that resulted from merging our genotypes to those of publicly available whole-genomes (WGS) of worldwide wolves, dogs, and a golden jackal (see Materials and Methods for details; Supplementary Table 4.3). The topology of the inferred neighbor-joining (NJ) tree followed that described for the genus *Canis* (Gopalakrishnan et al. 2018). Four out of the six non-Iberian wolves from our dataset grouped with North American wolves, while the remaining two non-Iberian wolves grouped with the Central European wolf population (**Figure 4.1B**). The four Iberian wolves and the five dogs from our dataset clustered within their respective group (**Figure 4.1B**). These results suggest that despite the great evolutionary distance between North American and European wolves (divergence of 30-20 ka; Fan et al. 2016) the expression profiles of the six non-Iberian wolves were remarkably similar, which allows their grouping as a control population for our comparisons of gene expression. We note, however, that non-Iberian wolves were born and kept in captivity in zoos and parks, which could have some effect on their brain gene expression profiles. Having gene expression profiles from wild non-Iberian wolves would enhance the accuracy of our comparisons, but such was not possible due to bureaucratic and sampling constraints.

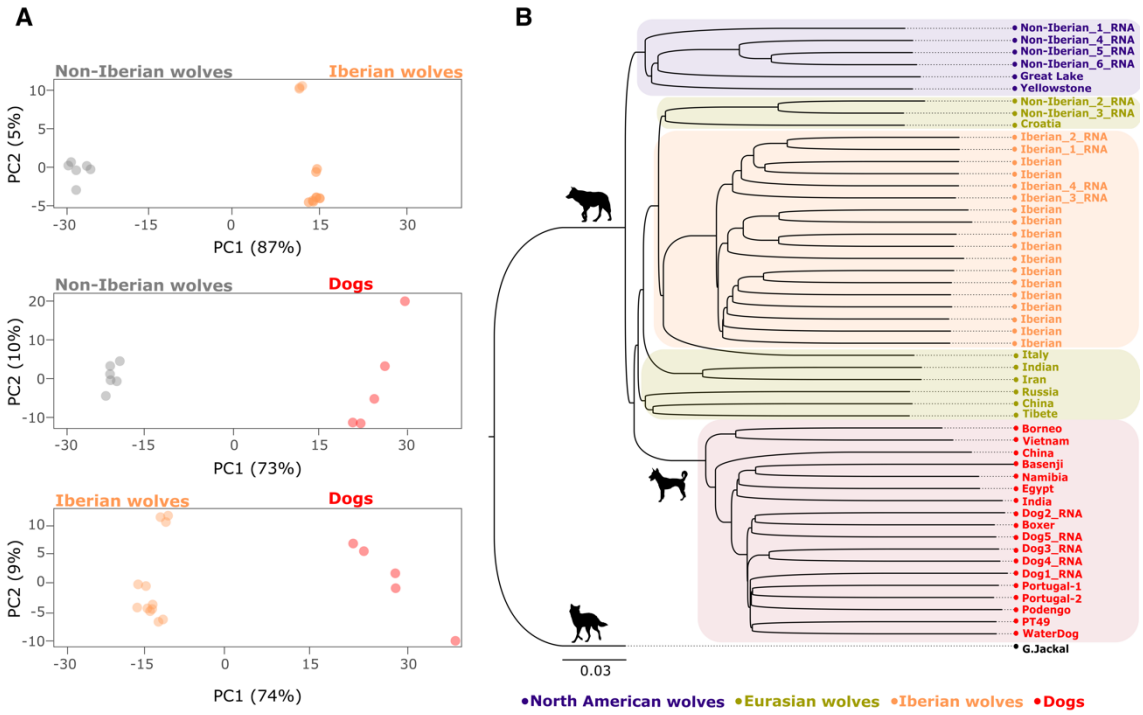


Figure 4.1. PCA of gene expression profiles and exome-wide NJ tree. (**A**) PCA plots for each of the three comparisons performed for differential expression analysis, constructed using normalized counts (rlog transformed) of all the genes in the reference set. Each dot represents a sample (for each of the four Iberian wolves three replicates/individual libraries were performed), and each group is represented by a different color. (**B**) NJ tree constructed with genetic distances estimated between all individuals using 79,267 exome-wide SNPs. Individuals used for the RNA-seq data are indicated with "RNA" in their sample ID (for Iberian wolf samples the three replicates were merged for genotyping). Each species/subspecies is represented by a different color.

Comparing gene expression profiles between Iberian and non-Iberian wolves resulted in 4204 DEGs (p-adj < 0.05 and |fold change| > 1), of which 1699 were up-regulated (hereafter referred to as "+") and 2505 were down-regulated (hereafter referred to as "-") in the former (**Figure 4.2A**). To investigate whether there was significant enrichment for specific terms within the 4204 DEGs, we performed a Gene Ontology Enrichment analysis that uncovered an overrepresentation of genes involved in protein binding, response to stress, system development, and regulation of response to stimulus, among others (detailed results of significantly enriched terms are summarized in Supplementary Figure 4.1). Differential expression analysis between non-Iberian wolves and dogs resulted in 1477 DEGs (+ 661, - 816 in dogs) (**Figure 4.2B**) and in 3460 DEGs (+ 2136, - 1324 in dogs) between Iberian wolves and dogs (**Figure 4.2C**). Of those, 420 (9.3%)

were found to be commonly differentially expressed in the two groups of wolves when compared with dogs.

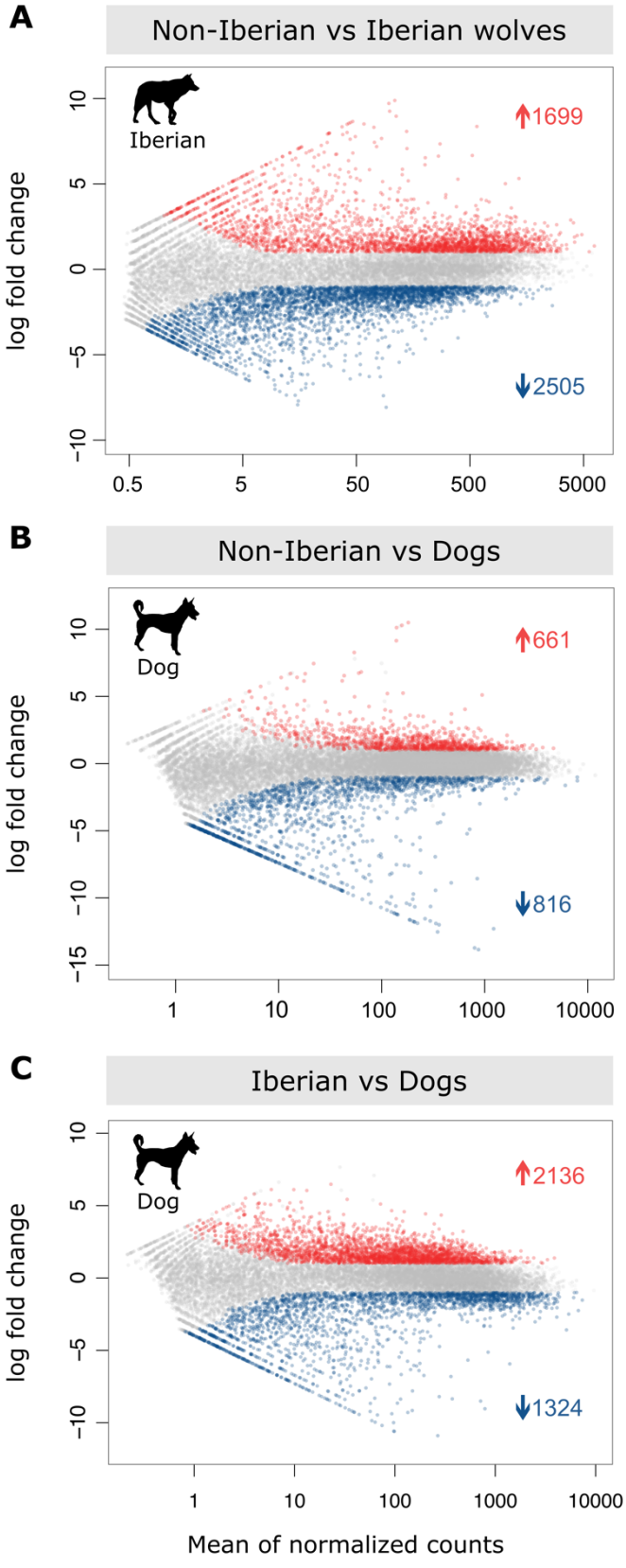


Figure 4.2. Levels of differential expression. MA plots showing the number of differentially expressed genes in each comparison (**A**) non-Iberian vs. Iberian wolves; (**B**) non-Iberian wolves vs. dogs; (**C**) Iberian wolves vs. dogs. The direction of expression is presented always in relation to the second condition (indicated within each plot by the dark silhouette). Up and down-regulated genes are colored in red and blue, respectively, and the corresponding number of genes in each category is presented.

4.2.2. Expressional changes in candidate genes for the domestic behavioral phenotype

Next, we investigated patterns of gene expression between Iberian and non-Iberian wolves in a gene set that we defined by their previous association with the domestic behavioral phenotype (i.e., genes displaying convergent signatures of selection and/or gene expression in domestic animals; Supplementary Table 4.4). More specifically, we investigated differences in expression in genes involved in neurologic receptor pathways, such as the glutamate and serotonin neurotransmitters, and in genes associated with GO categories of behavior and locomotory behavior found to be under positive selection in domestic dogs.

Serotonin and glutamate are neurotransmitters known to regulate aggressive and fear response behaviors (Wang et al. 2018), as well as memory and learning abilities (Ko et al. 2005; Li et al. 2014). More interestingly, genes in serotonin and glutamate pathways are thought to have been repeatedly selected across the domestication of animal species, more prominently regarding the glutamate receptor genes (Hou et al. 2020; O'Rourke and Boeckx 2020). Among the serotonin genes, we found differential expression between Iberian and non-Iberian wolves in HTR1A (+; direction in Iberian wolves), HTR3A (-), and HTR6 (-) (**Figure 4.3A**). Significant changes in the expression of serotonin receptors have previously been uncovered in brain regions of tame foxes (Kukekova et al. 2011; Wang et al. 2018), and fixed differences in allele frequencies of these genes between dogs and wolves have also been reported (Li et al. 2014). However, we found no significant differences in the expression profiles of these genes in dogs when compared to non-Iberian wolves that could associate their expression with the dog domestic behavioral phenotype (Supplementary Table 4.4).

We identified nine glutamate receptor genes differentially expressed between Iberian and non-Iberian wolves – GRIK1 (+), GRIK3 (-), GRM2 (-), GRM6 (-), GRM3 (+), GRIN2A (-), GRIN2D (-), GRIA4 (+) and GRID2 (+) (**Figure 4.3A**). Notably, the ionotropic kainite glutamate receptor gene GRIK3 was found to be one of the few targets of selection across several domestic mammals (dog, cattle, and sheep; O'Rourke and Boeckx 2020), and its decreased brain expression has been associated with a significant reduction in anxiety and fear memory (Ko et al. 2005; Hou et al. 2020). The NMDA glutamate receptor gene GRIN2A is also very relevant, as it was previously associated with dog domestication (Wang et al. 2015), and we found in Iberian wolves the same reduced expression observed in dogs when compared to non-Iberian wolves (**Figure 4.3B**;

Supplementary Table 4.4). Several other NMDA glutamate receptor genes were found in genomic regions of extreme differentiation between tame and aggressive foxes that also displayed levels of differential brain expression between these two fox strains (Kukekova et al. 2018; Wang et al. 2018). Of note is also the shared expression pattern within Iberian wolves and dogs of the MAST4 (-) gene (**Figures 4.3A** and **4.3B**; Supplementary Table 4.4), which interacts with several genes that activate NMDA glutamate receptors. MAST4 was previously identified as a potential candidate gene for behavioral adaptation in Iberian wolves, as it was inherited through dog introgression and showed signatures compatible with selection in both species (Lobo et al. Chapter III). Convergent signatures of gene expression in glutamate-related genes between Iberian wolves and dogs may suggest that the Iberian wolf population could be under a self-domestication process.

Domestication-related behaviors associated with glutamate receptors, such as reduced anxiety, harm avoidance, and increased memory and cognitive abilities, may contribute to explaining the increased tolerance towards humans and their activities known in Iberian wolves. Following this hypothesis, reduced levels of anxiety and increased learning and memory abilities would facilitate exploration behaviors by avoiding sources of human disturbance, but not fearing movements through high human-related risk constraints (Rio-Maior et al. 2019), including approximation to villages, and exploitation of human-derived food resources. Such behavioral alterations could also result in a reduction in long dispersal movements, ultimately increasing the likelihood of survival (Rio-Maior et al. 2019; Morales-González et al. 2021). Concordantly, we found differential levels of expression between Iberian and non-Iberian wolves in DOCK2 (+), CXCL10 (+), and ERC2 (+) (**Figure 4.3A**), genes associated with locomotory and exploration behavior (Freedman et al. 2016; Hou et al. 2020). The first two were suggested to be under positive selection in domestic dogs (Freedman et al. 2016), and we found the same increased gene expression in dogs as in Iberian wolves when compared to non-Iberian wolves (**Figure 4.3B**; Supplementary Table 4.4). These results shed light on a possible genetic basis for the remarkably low dispersal propensity seen in Iberian wolves (Silva et al. 2018). Yet, our results represent indirect evidence of the role of these genes in the possible adaptation of Iberian wolves, as extensive additional testing will be needed to prove their functional implications.

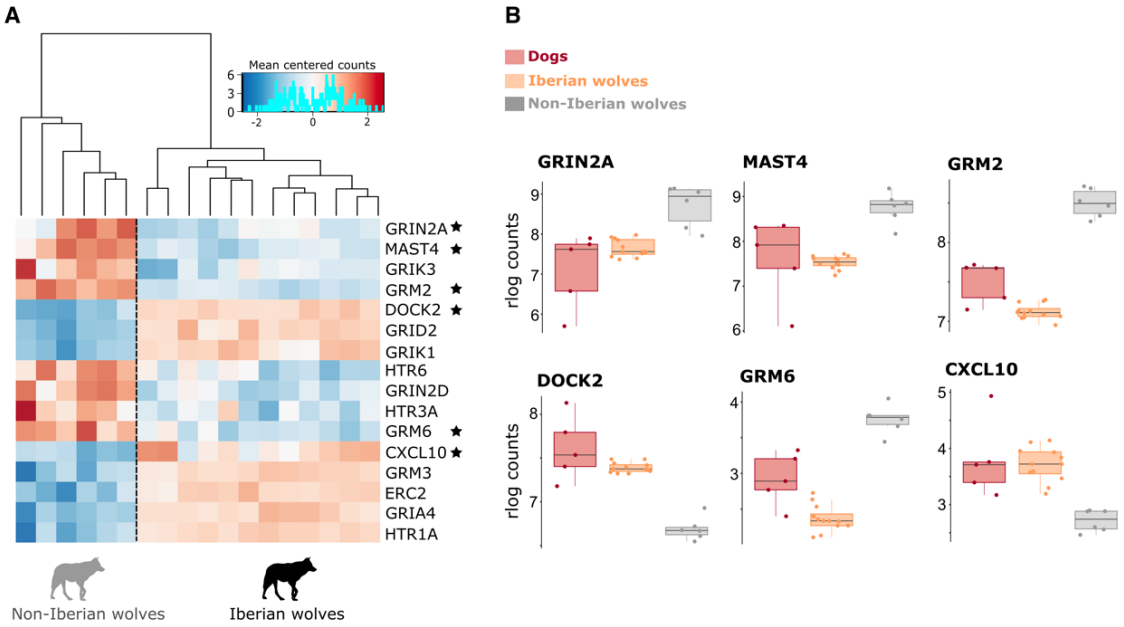


Figure 4.3. Expression profiles in the set of candidate genes. (**A**) Heatmap of rlog transformed mean centered counts reflecting levels of expression for genes associated with the domestic behavioral phenotype (rows) between Iberian and non-Iberian wolves. Genes that presented the same significant direction of expression in dogs as in Iberian wolves are identified with a star. Hierarchical clustering of Pearson's correlation coefficient between samples is represented in the dendrogram. (**B**) Expression profiles (average of rlog normalized counts) of genes marked with a star in (A) across the three groups. Boxplots display the mean value and standard errors, and each dot represents a single sample. The tree replicates per individual in Iberian wolves are presented in both figures.

4.2.3. Overlap between DEGs and allele frequency changes in Iberian wolves

To understand if differential expression patterns were associated with mutations in the protein-coding sequence, we identify 172,845 exome-wide SNPs in the RNA-seq data and calculated allele frequency differences (ΔAF) at these positions between Iberian and non-Iberian wolves. We found 1728 SNPs with extreme levels of differentiation (|ΔAF| > 0.83), of which ~36% mapped to protein-coding genes accordingly to the dog reference genome CanFam3.1 Ensembl v.99 annotation. About one-quarter of those (N = 181 genes) overlapped with the list of 4204 DEGs previously identified (Supplementary Table 4.5). Values of ΔAF were further tested using publicly available WGS data of 14 Iberian wolves and seven Eurasian wolves (Supplementary Table 4.3). This genomic data does not comprise the four individuals used for the RNA-seq data, allowing independent cross-validation. Apart from the GRIA4 gene where we found a SNP with high ΔAF associated with a synonymous mutation (Supplementary Table 4.6), no other gene from the candidate gene set was associated with high levels of differentiation in protein-coding

regions. These results indicate that the differential levels of gene expression we detected were not caused by variation in their protein-coding sequence, but likely through *cis*-acting regulatory mutations or indirectly by the action of distinct genes in a *trans*-regulatory process. Similar to what was previously proposed for the genetic mechanisms that underly phenotypic differentiation between wolves and domestic dogs (Li et al. 2014), the behavioral differences observed between Iberian and non-Iberian wolves may be mostly underlined by expressional variation rather than fixed mutations in the protein-coding sequence. Furthermore, alterations in the transcriptome may have allowed for rapid and immediate responses to the human-modified landscapes, as the accumulation of advantageous mutations usually occurs at a slower rate (Via and Lande 1985; Seebacher et al. 2015; Rivera et al. 2021).

Within the 181 genes displaying both high ΔAF and differential gene expression between Iberian and non-Iberian wolves, we found 20 genes associated with neurological processes that could be interesting candidates (Supplementary Table 4.6). For this set of genes, we further investigate the SNPs associated with high ΔAF in order to understand their possible functional roles. Except for SNPs found in GPRC5C, SZT2, and NHLH2 genes (**Table 4.1**), most were translated into synonymous mutations.

Table 4.1. ΔAF. Allele frequency differences in candidate genes overlapping DEGs between Iberian and non-Iberian wolves. Information about the direction of gene expression in Iberian wolves, chromosome and chromosome position, derived allele, frequency of the derived allele calculated for the RNA-seq and WGS data in Iberian and non-Iberian wolves, and the genomic consequence associated with each SNP is presented. Data absent from WGS is represented with NA.

Gene*	Direction	Chr	Position	Derived allele	Freq Iberian (RNA)	Freq Iberian (WGS)	Freq non-Iberian (RNA)	Freq non-Iberian (WGS)	Genomic consequence
GPRC5C	Down (-)	9	5986419	C	1	0.57	0	0.25	3' UTR
			5986577	C	1	0.57	0	0.21	3' UTR
SZT2	Down (-)	15	16743755	T	1	0.32	0	0	Synonymous
			16747446	T	1	NA	0	NA	Missense
NHLH2	Up (+)	17	53360278	G	0.88	0.46	0	0	3' UTR

* Gene description:
GPRC5C: G Protein-Coupled Receptor Class C Group 5 Member C (class C metabotropic glutamate)
SZT2: Seizure Threshold 2 Protein Homolog
NHLH2: Nescient Helix-Loop-Helix 2

GPRC5C exhibited two consecutive SNPs (T-to-C) positioned in a 3' prime UTR whose derived alleles were fixed in Iberian wolves (with a C frequency of 100%) but were not found in non-Iberian wolves (**Table 4.1**). These allele frequency changes were validated using WGS data, confirming a higher frequency of the C allele in Iberian wolves (frequency of 57% vs. 23% in non-Iberian wolves). Most importantly, GPRC5C is a G

protein-coupled receptor that belongs to the glutamate receptor subfamily (Robbins et al. 2000), and its expression was significantly decreased in the Iberian wolf's brain compared to non-Iberian wolves. It is thus possible that this mutation in the UTR region of the gene could be associated with its down-regulation in Iberian wolves. These overlapping signatures of shifts in allele frequencies and gene expression differences in glutamate-related genes may indicate this neurotransmitter played an important role in the evolution of Iberian wolves in human-dominated landscapes.

In the SZT2, we found a C-to-T change in the protein-coding sequence, causing an arginine-to-leucine missense mutation (reverse strand translated), with a T frequency of 100% in the Iberian wolves using exome-wide SNP data (**Table 4.1**). This SNP was missing in the WGS data, but an immediately upstream SNP with which it is likely linked, showed the same frequency in the exome-wide SNP dataset and a 32% frequency using WGS data (**Table 4.1**). The T allele was not present in non-Iberian wolves. Because this mutation is not yet described in the dog genome, it is difficult to predict its impact on the function of the protein, however, we know that SZT2 is down-regulated in Iberian wolves, which may imply some effect. Mutations in SZT2 have been associated with juvenile epilepsy and neurodevelopment delay in both humans (Sun et al. 2019) and mice (Frankel et al. 2009).

Finally, similarly to GPRC5C, NHLH2 also exhibited a SNP (C-to-G) in the 3' prime UTR in high frequency in Iberian wolves (G frequency of 88% in the exome-wide SNP dataset; 46% in WGS) comparative to non-Iberian wolves where the G allele was not found (**Table 4.1**). This gene showed increased expression in the Iberian wolf's brain compared to non-Iberian wolves. Notably, previous studies in mice have shown that NHLH2 has the potential to influence either the motivation or ability to exercise, and individuals with mutations in this gene showed reduced physical activity (Good et al. 2008). The parallels with the reduced dispersal ability, or tendency to disperse, known in Iberian wolves are particularly striking, making NHLH2 a very interesting candidate for this behavioral phenotype. There were no significant differences in the expression of these three genes in dogs compared to non-Iberian wolves, nor change in allele frequencies, that could indicate that these alterations are associated with the dog domestic phenotype. However, these expressional and allele frequency changes might reflect a direct response to adaptation to human-induced environmental change in Iberian wolves, which could be similar, but not necessarily the same, to those selected during dog domestication.

4.3. Conclusions

In the face of an increasingly human-modified planet, it is of utmost urgency to understand the mechanisms allowing species adaptation. In this work, we used RNA-seq to generate the first brain transcriptomes of a wild wolf population and uncover a possible genetic basis for behavioral responses to human-dominated landscapes. We found expressional changes in the brain of Iberian wolves in genes that were previously associated with the domestic behavioral phenotype, including serotonin and glutamate receptor genes, as well as in genes associated with locomotory behavior. The direction of expression in some of these genes points to reduced anxiety, increased cognitive abilities, and reduced exploratory tendencies, which may contribute to explaining the increased tolerance towards humans and their activities known in Iberian wolves. Concordantly, our findings also revealed significant allele frequency changes in differentially expressed genes, such as the GPRC5C, which belongs to the glutamate receptor subfamily, and the NHLH2, which is associated with reduced physical activity. These overlapping signatures of changes in allele frequencies and gene expression suggest these could be potential candidate genes for the behavioral responses described in the Iberian wolf. This study provides a first glimpse of the evolutionary mechanisms that could have allowed Iberian wolves to persist in highly human-modified landscapes, which may be under a process of self-domestication. These mechanisms seem to have comprised rapid and immediate responses involving the alteration of the transcriptome and a subsequent slower accumulation of point mutations. Future studies will be required to investigate and characterize functional implications of these genes, particularly those involved in the metabolism of glutamate, on the behavioral phenotypes of Iberian wolves and to assess their adaptive potential.

4.4. Materials and Methods

4.4.1. Brain tissue collection

Samples from the brain frontal cortex were collected from four wild-caught adult Iberian wolves (three males, one female) in Asturias, North of Spain, between 2017 and 2020 (Supplementary Table 4.1). Wolves were killed by local authorities in the scope of population density control programs approved by the Principado de Asturias. No animals were injured or killed purposefully for this study. To avoid stress-induced bias, animals were not chased, and brain samples were dissected immediately after death.

Specifically, we targeted the frontal cortex, which is the brain region responsible for complex behaviors in mammals, as memory and learning abilities. Frontal cortex was cut in small pieces following a standard procedure across the four individuals and preserved in RNAlater solution (Qiagen). Samples were then stored at -80ºC until RNA extraction.

4.4.2. RNA-seq experiments

Total RNA was isolated using the RNeasy Mini kit (Qiagen) according to manufacturer's instructions and purified (removing DNA contamination) by treatment with RNase-Free DNase II (Qiagen). We measured RNA integrity (RIN), concentration and quality in each sample using the TapeStation RNA ScreenTape (Agilent Technologies), Nanodrop, and Qubit RNA BR assay kit (Thermo Fisher Scientific). All samples had RIN values above 8. Prior to library preparation, total RNA in a sample was normalized to 2.5 µg. Three independent strand-specific Illumina libraries (replicates) were prepared for each individual using the LM-seq protocol (Hou et al. 2015). Libraries (~300 bp long) were pooled and sequenced on the Illumina Hiseq X instrument, using a PE 150 bp chemistry to an average of 30 million reads per library, at a sequencing service provide (Macrogen, Rep. of Korea).

In the absence of brain samples from other wolves from wild populations, we added to our dataset a set of publicly available RNA-seq data generated from frontal cortex tissues of six gray wolves kept in captivity (designated as "non-Iberian"; Albert et al. 2012), and also from five dogs (several breeds; Broad Institute 2019) (Supplementary Table 4.1). These samples were essential to perform differential expression analysis and compare gene expression profiles. We used the online tool CurSE (Vaneechoutte and Vandepoele 2019) (http://bioinformatics.psb.ugent.be/webtools/Curse/) to search for all the publicly available RNA-seq data derived from the frontal cortex of dogs and wolves.

Raw reads were trimmed using TRIMMOMATIC v.0.39 (Bolger et al. 2014) to remove low quality reads (<20), short reads (<20 bp), and adapter sequences under the following parameters: *ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:2:keepBothReads LEADING:3 TRAILING:3 MINLEN:20 SLIDINGWINDOW:4:20*. Reads passing the filtering step were aligned to the dog reference genome CanFam3.1 using a spliced-aware aligner, STAR v.2.7.5a (Dobin et al. 2013), and multi-aligned reads were discarded. We then used featureCounts (Liao et al. 2014) to obtain gene abundance estimates based on CanFam3.1 Ensembl v.99 annotation file. Reads were counted as a meta-feature by

aggregating the counts at a gene level. The final set was comprised by 30,951 reference genes.

4.4.3. Differential expression analysis

Expression profiles in the frontal cortex of Iberian and non-Iberian wolves were compared through differential expression gene analysis (comparison 1). We also performed differential expression analysis between non-Iberian wolves and dogs (comparison 2), to have a reference for genes associated with the domestic behavioral phenotype. As a control, we also compared expression profiles between Iberian wolves and dogs (comparison 3).

Prior to differential expression analysis, we used TVscript (Lobo et al. 2022; Chapter VI) to characterize and remove genes that were associated with high levels of intra-condition variation. As input, TVscript requires raw count data per sample, and samples must be grouped into two conditions. The absolute pairwise differences between normalized counts of a given gene for each condition are calculated, and the variance within these differences is determined. We ran TVscript for the three comparisons and exclude genes associated with the top 5% levels of intra-condition variation, resulting in 2590, 2313 and 2568 genes removed in each comparison 1,2 and 3, respectively. The removal of such genes associated with high levels of intra-condition variation was shown to increase the reliability of the differentially expressed genes identified (Lobo et al. 2022; Chapter VI).

After filtering, we used raw counts to perform differential expression analysis with DESEQ2 (Love et al. 2014) in each of the three comparisons. Genes located within sex chromosomes, without chromosome location or annotation, and microRNA genes, were not considered for differential expression analysis. Genes with low counts (< 10 counts across all samples) were also discarded. DEGs were determined following two criteria: i) p-adj < 0.05 (corrected by the Benjamini and Hochberg method); and ii) |log2-fold changes| > 1. The overall effects of experimental covariates in each comparison were checked by performing a PCA with the *plotPCA* function from DESEQ2 with normalized count values. Most of the variation in our datasets was explained between, not within, conditions (**Figure 4.1A**), validating datasets for our comparisons.

DEGs detected between Iberian and non-Iberian wolves were subjected to a GO analysis in g:Profiler (Raudvere et al. 2019) to find possible enrichment for specific GO terms and biological pathways (KEGG). The analysis was performed using the Ensembl annotation v.99 of CanFam 3.1 reference genome and a p-value < 0.05 (after correction

with g:SCS for false discovery rate) to define enriched terms. GO terms were classified into the three main categories of biological process, cellular component, and molecular function.

4.4.4. Exome-wide SNP calling and allele frequency estimate

To investigate allele frequency changes between Iberian and non-Iberian wolves, and associate those with differential expression results, we called 172,845 genome-wide exome-wide SNPs from the RNA-seq alignments of all individuals (Iberian wolves, non-Iberian wolves, and dogs). Before genotype calling, we merged the bam files of the three replicates for each Iberian wolf using SAMtools merge tool (Li et al. 2009). Then, we followed GATK "Best Practices Guide" to identify short variants in RNA-seq data. Briefly, duplicates were marked with PICARD MarkDuplicates, alignments spanning introns were reformatted with SplitNCigarReads tool from GATK v.4.1.8.1 (McKenna et al. 2010), base quality scores were recalibrated, and genotypes were called using HaplotypeCaller tool from GATK. HaplotypeCaller was firstly run per sample under the -ERC GVCF mode with the minimum Phred-scaled threshold of 20 (*-stand-call-conf 20*), as suggested. Lastly, the joint genotyping with all samples was conducted with GenotypeGVCFs tool. The resulting raw VCF file was first subdivided to keep only SNPs (exclude indels), and posteriorly filtered with GATK VariantFiltration tool as recommended: *QD < 2.0, QUAL < 30.0, SOR > 3.0, FS > 60.0, MQ < 20.0 and DP < 4.0*. A final set of filters were applied with VCFtools v.0.1.15 (Danecek et al. 2011), to i) keep only bi-allelic and autosomal SNPs with the filter flag *PASS*, ii) have less than 10% of missing data (*--max-missing 0.9*) and iii) were supported by at least four reads (*--min-meanDP 4*). Finally, we estimated missingness and depth statistics across all 15 samples using VCFtools and verified that all individuals had less than 20% of missing data and mean coverage above 10x.

We estimated the frequency of the derived allele across the 172,845 SNPs for Iberian wolves, non-Iberian wolves and dogs using VCFtools. We further calculated ΔAF (allele frequency difference) between Iberian and non-Iberian wolves and selected SNPs with $|\Delta AF| \geq 0.83$ (corresponding to the 99th percentile). From those, we evaluated the possible association between highly differentiated SNPs and differences in gene expression by selecting SNPs that overlapped with protein-coding genes and also showing significant differences in expression between Iberian and non-Iberian wolves. To ensure that allele frequency changes in candidate regions were not biased by the low sample size, ΔAF was further confirmed using ~10.8 million SNPs from WGS data of 14 Iberian wolves and seven Eurasian wolves (Supplementary Table 4.3). These individuals

were distinct from those used in the RNA-seq data to allow independent cross-validation. This data was obtained from a VCF file we previously generated in Chapter III containing several others canid genomes.

Because the geographical origin of the wild ancestors of the non-Iberian wolves used in this study was unknown, although these individuals were known to be unrelated and thought to have born in captivity in Germany and Austria (Dr. Frank W. Albert, pers. comm.), we investigated their position in a phylogenetic tree containing gray wolves from populations worldwide. To do that, we first merged our dataset with the VCF file containing the canid genomes using BCFtools merge tool (Li 2011). In addition to the 14 Iberian wolves and the seven Eurasian wolves previously mentioned, we added WGS data from two North American wolves, 13 dogs and one golden jackal (Supplementary Table 4.3). The merged VCF file was posteriorly filtered with VCFtools to keep sites with less than 10% missing (*--max-missing 0.9*) and with minor allele frequency (MAF) greater or equal to 1% (*--maf 0.01*), resulting in 79,267 exome-wide SNPs. Using this dataset we constructed a neighbor-joining (NJ) tree based on pairwise genetic distances estimated in TASSEL 5 (Bradbury et al. 2007) as 1 – IBS (identity by state) across all individuals. The NJ was built also in Tassel 5 setting the golden jackal as root and drawn with FIGTREE v.1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/).

4.5. References

Albert FW, Somel M, Carneiro M, Aximu-Petri A, Halbwax M, Thalmann O, Blanco-Aguiar JA, Plyusnina IZ, Trut L, Villafuerte R, et al. 2012. A Comparison of Brain Gene Expression Levels in Domesticated and Wild Animals. PLoS Genet. 8:e1002962.

Andersson L, Purugganan M. 2022. Molecular genetic variation of animals and plants under domestication. Proc. Natl. Acad. Sci. 119.

Axelsson E, Ratnakumar A, Arendt M-L, Maqbool K, Webster MT, Perloski M, Liberg O, Arnemo JM, Hedhammar Å, Lindblad-Toh K. 2013. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature 495:360–364.

Beckman AK, Richey BMS, Rosenthal GG. 2022. Behavioral responses of wild animals to anthropogenic change: insights from domestication. Behav. Ecol. Sociobiol. 76:105.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. 2007. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635.

Carroll SB. 2005. Evolution at Two Levels: On Genes and Form. PLoS Biol. 3:e245.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

Dennehy E, Llaneza L, López-Bao JV. 2021. Contrasting wolf responses to different paved roads and traffic volume levels. Biodivers. Conserv. 30:3133–3150.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21

Fan Z, Silva P, Gronau I, Wang S, Armero AS, Schweizer RM, Ramirez O, Pollinger J, Galaverni M, Ortega Del-Vecchyo D, et al. 2016. Worldwide patterns of genomic variation and admixture in gray wolves. Genome Res. 26:163–173.

Frankel WN, Yang Y, Mahaffey CL, Beyer BJ, O'Brien TP. 2009. Szt2 , a novel gene for seizure threshold in mice. Genes, Brain Behav. 8:568–576.

Freedman AH, Schweizer RM, Ortega-Del Vecchyo D, Han E, Davis BW, Gronau I, Silva PM, Galaverni M, Fan Z, Marx P, et al. 2016. Demographically-Based Evaluation of Genomic Regions under Selection in Domestic Dogs. PLoS Genet. 12(3):e100.

Gibson G. 2008. The environmental contribution to gene expression profiles. Nat. Rev. Genet. 9:575–581.

Good DJ, Coyle CA, Fox DL. 2008. Nhlh2: A Basic Helix-Loop-Helix Transcription Factor Controlling Physical Activity. Exerc. Sport Sci. Rev. 36:187–192.

Gopalakrishnan S, Sinding MHS, Ramos-Madrigal J, Niemann J, Samaniego Castruita JA, Vieira FG, Carøe C, Montero M de M, Kuderna L, Serres A, et al. 2018. Interspecific Gene Flow Shaped the Evolution of the Genus Canis. Curr. Biol. 28:3441-3449.e5.

Hare B, Wobber V, Wrangham R. 2012. The self-domestication hypothesis: evolution of bonobo psychology is due to selection against aggression. Anim. Behav. 83:573–585.

Hou Y, Qi F, Bai X, Ren T, Shen X, Chu Q, Zhang X, Lu X. 2020. Genome-wide analysis reveals molecular convergence underlying domestication in 7 bird and mammals. BMC Genomics 21:1–20.

Hou Z, Jiang P, Swanson SA, Elwell AL, Nguyen BKS, Bolin JM, Stewart R, Thomson JA. 2015. A cost-effective RNA sequencing protocol for large-scale gene expression studies. Sci. Rep. 5:9570.

Hunter P. 2007. The human impact on biological diversity. EMBO Rep. 8:316–318.

Kijas JW, Lenstra JA, Hayes B, Boitard S, Porto Neto LR, San Cristobal M, Servin B, McCulloch R, Whan V, Gietzen K, et al. 2012. Genome-Wide Analysis of the World's Sheep Breeds Reveals High Levels of Historic Mixture and Strong Recent Selection. PLoS Biol. 10:e1001258.

Ko S, Zhao MG, Toyoda H, Qiu CS, Zhuo M. 2005. Altered behavioral responses to noxious stimuli and fear in glutamate receptor 5 (GluR5)- or GluR6-deficient mice. J. Neurosci. 25:977–984.

Kukekova A, Johnson J, Teiling C, Li L, Oskina I, Kharlamova A, Shikhevich S, Sun Q, Ponnala L, Temnykh S, et al. 2011. Sequence comparison of prefrontal cortical brain transcriptome from a tame and an aggressive silver fox (Vulpes vulpes). BMC Genomics 12:482.

Kukekova A, Johnson J, Xiang XY, Feng SH, Liu S, Rando H, Kharlamova A, Herbeck Y, Natalya S, Xiong Z, et al. 2018. The red fox genome assembly identifies genomic regions associated with tame and aggressive behaviors. Nat. Ecol. Evol. 2:1479–1491.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987–2993.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Li Y, Wang GD, Wang MS, Irwin DM, Wu DD, Zhang YP. 2014. Domestication of the dog from the Wolf was promoted by enhanced excitatory synaptic plasticity: A hypothesis. Genome Biol. Evol. 6:3115–3121.

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30:923–930.

Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ, Zody MC, et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438:803–819.

Llaneza L, García EJ, Palacios V, Sazatornil V, López-Bao JV. 2016. Resting in risky environments: the importance of cover for wolves to cope with exposure risk in human-dominated landscape. Biodivers. Conserv. 25:1515–1528.

Llaneza L, López-Bao JV. 2015. Indirect effects of changes in environmental and agricultural policies on the diet of wolves. Eur. J. Wildl. Res. 61:895–902.

Lobo D, Linheiro R, Godinho R, Archer JP. 2022. On taming the effect of transcript level intra-condition count variation during differntial expression analysis: A story of dogs, foxes and wolves. PLoS ONE 17:e0274591.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15:550.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303.

Morales-González A, Fernández-Gil A, Quevedo M, Revilla E. 2021. Patterns and determinants of dispersal in grey wolves (Canis lupus). Biol. Rev.97:466-480.

Newsome TM, Boitani L, Chapron G, Ciucci P, Dickman CR, Dellinger JA, López-Bao J V., Peterson RO, Shores CR, Wirsing AJ, et al. 2016. Food habits of the world's grey wolves. Mamm. Rev. 46:255–269.

O'Rourke T, Boeckx C. 2020. Glutamate receptors in domestication and modern human evolution. Neurosci. Biobehav. Rev. 108:341–357.

Pavey SA, Collin H, Nosil P, Rogers SM. 2010. The role of gene expression in ecological speciation. Ann. N. Y. Acad. Sci. 1206:110–129.

Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, Vilo J. 2019. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 47:W191–W198.

Rio-Maior H, Nakamura M, Álvares F, Beja P. 2019. Designing the landscape of coexistence: Integrating risk avoidance, habitat selection and functional connectivity to inform large carnivore conservation. Biol. Conserv. 235:178–188.

Rivera HE, Aichelman HE, Fifer JE, Kriefall NG, Wuitchik DM, Wuitchik SJS, Davies SW. 2021. A framework for understanding gene expression plasticity and its influence on stress tolerance. Mol. Ecol. 30:1381–1397.

Robbins MJ, Michalovich D, Hill J, Calver AR, Medhurst AD, Gloger I, Sims M, Middlemiss DN, Pangalos MN. 2000. Molecular Cloning and Characterization of Two Novel Retinoic Acid-Inducible Orphan G-Protein-Coupled Receptors (GPRC5B and GPRC5C). Genomics 67:8–18.

Rubin CJ, Megens HJ, Barrio AM, Maqbool K, Sayyab S, Schwochow D, Wang C, Carlborg Ö, Jern P, Jørgensen CB, et al. 2012. Strong signatures of selection in the domestic pig genome. Proc. Natl. Acad. Sci. 109:19529–19536.

Rubin CJ, Zody MC, Eriksson J, Meadows JRS, Sherwood E, Webster MT, Jiang L, Ingman M, Sharpe T, Ka S, et al. 2010. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature 464:587–591.

Saetre P, Lindberg J, Ellegren H, Vila C, Jazin E, Leonard JA, Olsson K, Pettersson U, Bergstrom TF. 2004. From wild wolf to domestic dog: Gene expression changes in the brain. Mol. Brain Res. 126:198–206.

Sato DX, Rafati N, Ring H, Younis S, Feng C, Blanco-Aguiar JA, Rubin C-J, Villafuerte R, Hallböök F, Carneiro M, et al. 2020. Brain Transcriptomics of Wild and Domestic Rabbits Suggests That Changes in Dopamine Signaling and Ciliary Function Contributed to Evolution of Tameness. Genome Biol. Evol. 12:1918–1928.

Seebacher F, White CR, Franklin CE. 2015. Physiological plasticity increases resilience of ectothermic animals to climate change. Nat. Clim. Chang. 5:61–66.

Silva P, López-Bao JV, Llaneza L, Álvares F, Lopes S, Blanco JC, Cortés Y, García E, Palacios

V, Rio-Maior H, et al. 2018. Cryptic population structure reveals low dispersal in Iberian wolves. Sci. Rep. 8:14108.

Sullivan AP, Bird DW, Perry GH. 2017. Human behaviour as a long-term ecological driver of non-human evolution. Nat. Ecol. Evol. 1:0065.

Sun X, Zhong X, Li T. 2019. Novel SZT2 mutations in three patients with developmental and epileptic encephalopathies. Mol. Genet. Genomic Med. 7:e926.

Vaneechoutte D, Vandepoele K. 2019. Curse: building expression atlases and co-expression networks from public RNA-Seq data. Bioinformatics 35:2880–2881.

Via S, Lande R. 1985. Genotype-Environment Interaction and the evolution of phenotypic plasticity. Evolution 39:505–522.

Wang GD, Zhai W, Yang HC, Wang L, Zhong L, Liu YH, Fan RX, Yin TT, Zhu CL, Poyarkov AD, et al. 2015. Out of southern East Asia: the natural history of domestic dogs across the world. Cell Res. 26:21–33.

Wang X, Pipes L, Trut L, Herbeck Y, Vladimirova A, Gulevich R, Kharlamova A, Johnson J, Acland G, Kukekova A, et al. 2018. Genomic responses to selection for tame/aggressive behaviors in the silver fox (Vulpes vulpes). Proc. Natl. Acad. Sci. 115:10398–10403.

Yeh PJ, Price TD. 2004. Adaptive Phenotypic Plasticity and the Successful Colonization of a Novel Environment. Am. Nat. 164:531–542.

Zeder MA. 2008. Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. Proc. Natl. Acad. Sci. 105:11597–11604.

4.6. Supplementary Materials



Molecular Function (n=4)	
GO:0005488	binding
GO:0005515	protein binding
GO:0042803	protein homodimerization activity
GO:0042802	identical protein binding

Biological Process (n=45)	
GO:0043412	macromolecule modification
GO:0071840	cellular component organization or biogenesis
GO:0016043	cellular component organization
GO:0048518	positive regulation of biological process
GO:0032501	multicellular organismal process
GO:0048522	positive regulation of cellular process
GO:0032502	developmental process
GO:0007155	cell adhesion
GO:0006950	response to stress
GO:0048856	anatomical structure development
GO:0043588	skin development
GO:0022610	biological adhesion
GO:0033554	cellular response to stress
GO:0048519	negative regulation of biological process
GO:0009892	negative regulation of metabolic process
GO:0036211	protein modification process
GO:0006464	cellular protein modification process
GO:0044248	cellular catabolic process
GO:0007275	multicellular organism development
GO:0007166	cell surface receptor signaling pathway
GO:0070925	organelle assembly
GO:0008283	cell population proliferation
GO:0042127	regulation of cell population proliferation
GO:0031324	negative regulation of cellular metabolic process
GO:0048523	negative regulation of cellular process
GO:0009893	positive regulation of metabolic process
GO:0006281	DNA repair
GO:0006996	organelle organization
GO:0048513	animal organ development
GO:0006259	DNA metabolic process
GO:0006974	cellular response to DNA damage stimulus
GO:0010629	negative regulation of gene expression
GO:0048731	system development
GO:0009056	catabolic process
GO:0010605	negative regulation of macromolecule metabolic process
GO:0009451	RNA modification
GO:2000113	negative regulation of cellular macromolecule biosynthetic process
GO:0051383	kinetochore organization
GO:0008544	epidermis development
GO:0048583	regulation of response to stimulus
GO:0051382	kinetochore assembly
GO:0009888	tissue development
GO:0051179	localization
GO:0030198	extracellular matrix organization
GO:0031327	negative regulation of cellular biosynthetic process

Cellular Component (n=22)	
GO:0043227	membrane-bounded organelle
GO:0005654	nucleoplasm
GO:0070013	intracellular organelle lumen
GO:0043233	organelle lumen
GO:0031974	membrane-enclosed lumen
GO:0031981	nuclear lumen
GO:0043231	intracellular membrane-bounded organelle
GO:0005737	cytoplasm
GO:0071944	cell periphery
GO:0005634	nucleus
GO:0005886	plasma membrane
GO:0005813	centrosome
GO:0005622	intracellular
GO:0005856	cytoskeleton
GO:0015630	microtubule cytoskeleton
GO:0005815	microtubule organizing center
GO:0005829	cytosol
GO:0043226	organelle
GO:0000151	ubiquitin ligase complex
GO:1990234	transferase complex
GO:0005657	replication fork
GO:0005730	nucleolus

Human Phenotype (n=1)	
HP:0000007	autosomal recessive inheritance

Supplementary Figure 4.1. Gene Ontology Analysis. Gene ontology terms significantly enriched (p-value < 0.05) within the 4204 DEGs between Iberian and non-Iberian wolves. The GO term ID and definition are provided.

Supplementary Table 4.1. Brain transcriptomic data. Description of the 15 canid brain transcriptomes used in this study. We generated brain transcriptomes of four wild-caught Iberian wolves and added additional data from six wolves (non-Iberian) and five dogs from different studies.

Sample ID	Population	Sex	Age	Collection date	BioSample	Study
Wolf-1 (3x)	Iberian wolf (Spain)	F	Adult	11 October 2017		This study
Wolf-2 (3x)	Iberian wolf (Spain)	M	Adult	11 October 2017		This study
Wolf-3 (3x)	Iberian wolf (Spain)	M	Adult	13 February 2018		This study
Wolf-4 (3x)	Iberian wolf (Spain)	M	Adult	18 May 2020		This study
Non-IP-1	NA	M	Old adult	NA	SAMEA1876817	[1]
Non-IP-2	NA	F	Old adult	NA	SAMEA1876823	[1]
Non-IP-3	NA	F	Old adult	NA	SAMEA1876888	[1]
Non-IP-4	NA	M	Old adult	NA	SAMEA1876902	[1]
Non-IP-5	NA	M	Old adult	NA	SAMEA1876839	[1]
Non-IP-6	NA	M	Old adult	NA	SAMEA1876829	[1]
Dog-1	Yorkshire Terrier	M	11 ya	NA	SAMN11526752	[2]
Dog-2	Newfoundland	M	11 ya	NA	SAMN11526831	[2]
Dog-3	Belgian Malanois	M	3 ya	NA	SAMN11526804	[2]
Dog-4	Belgian Malanois	M	2 ya	NA	SAMN11526778	[2]
Dog-5	Labrador	M	16 ya	NA	SAMN07419804	[2]

[1] – Albert et al. 2012; [2] – Broad Institute 2019.

Abbreviation: F, female; M, male; ya, years old; NA, not available

Supplementary Table 4.2. RNA-seq statistics. General sequencing statistics for the Iberian wolf transcriptomes generated in this study (for each replicate, R1-R3), and for the transcriptomes of non-Iberian wolves and dogs generated in other studies. Information provided refers to the number of raw reads, reads passing trimming, reads uniquely mapped to the dog reference genome and reads that mapped to genes. Additionally, it is also shown the percentage of missing and average depth across the 172,842 exome-wide SNPs (in Iberian wolf samples the three replicates were merged for genotyping).

Sample ID	Raw reads	Trimmed reads	Reads uniquely mapped	Reads mapped to genes	Missing (%)	Mean depth
Wolf-1 (R1)	34834971	27926136	21902195	12713720		
Wolf-1 (R2)	38836339	31493797	25171724	15087201	0.18	64.85
Wolf-1 (R3)	31591559	25578251	20964353	12265632		
Wolf-2 (R1)	32133614	24729847	18843560	10728812		
Wolf-2 (R2)	29080887	23670128	19650446	11709383	0.31	60.22
Wolf-2 (R3)	37027185	27347239	20509556	11779203		
Wolf-3 (R1)	29813639	23935113	19114346	11176474		
Wolf-3 (R2)	32856989	26633290	21558655	12742638	0.23	58.95
Wolf-3 (R3)	30999364	22836366	17761403	10249257		
Wolf-4 (R1)	40191079	29226417	21020148	11939686		
Wolf-4 (R2)	35627661	25078023	18567772	10733918	0.37	56.20
Wolf-4 (R3)	34606162	22934969	16514438	9529548		
Dog-1	70078608	62862729	59942566	40817088	0.06	95.36
Dog-2	49489034	47160028	44174616	29395522	0.92	46.99
Dog-3	33948940	30060012	28418871	18727437	0.69	40.42
Dog-4	47266325	40611676	37704428	25948695	0.99	49.64
Dog-5	59298584	56628826	52560063	33658211	8.45	34.82
Non-IP-1	21400074	18683067	16264815	10157359	6.26	15.30
Non-IP-2	21860980	19571683	17188275	10525022	4.16	16.23
Non-IP-3	18597874	14447078	12821218	8735681	17.41	10.62
Non-IP-4	20069118	16986524	14309603	9239906	7.53	14.15
Non-IP-5	16250247	14942950	13163242	8236266	6.64	13.84
Non-IP-6	21544048	18991266	16178388	10053779	4.81	15.69

Supplementary Table 4.3. Whole-genome sequencing data. Description of the 37 canid genomes used in this study from publicly available data.

Sample ID	Population	Locality	BioSample	Study
L090	Iberian Wolf	Portugal		[3]
L351	Iberian Wolf	Spain		[3]
L474	Iberian Wolf	Portugal		[3]
L546	Iberian Wolf	Spain		[3]
L552	Iberian Wolf	Spain		[3]
L590	Iberian Wolf	Spain		[3]
L253	Iberian Wolf	Spain		[3]
L409	Iberian Wolf	Spain		[3]
L514	Iberian Wolf	Spain		[3]
L547	Iberian Wolf	Spain		[3]
L588	Iberian Wolf	Spain		[3]
L844	Iberian Wolf	Spain		[3]
Wolf24	Iberian Wolf	Portugal	SAMN02921316	[6]
Wolf39	Iberian Wolf	Spain	SAMN04851099	[8]
Wolf21	Eurasian Wolf	Italy	SAMN02921313	[6]
Wolf06	Eurasian Wolf	Croatia	SAMN03366712	[5]
Wolf01	Eurasian Wolf	Russia	SAMN01974486	[4]
Wolf05	Eurasian Wolf	China	SAMN03366711	[5]
WolfTibetan	Eurasian Wolf	Tibet	SAMN03653004	[7]
Wolf19	Eurasian Wolf	India	SAMN02921311	[6]
Wolf20	Eurasian Wolf	Iran	SAMN02921312	[6]
Wolf28	N-A Wolf	Yellowstone	SAMN02921320	[6]
Wolf18	N-A Wolf	Great Lakes	SAMN02921310	[6]
PT49	Village dog	Portugal	SAMN04851101	[9]
VillDog_PT01	Village dog	Portugal	SAMN02485604	[9]
VillDog_PT02	Village dog	Portugal	SAMN02485605	[9]
PT_Podengo	Podengo	Portugal	SAMN03580388	[10]
PT_WaterDog	Water Dog	Portugal	SAMN08873219	[11]
Boxer	Boxer	NA	SAMN02921305	[5]
CFA.118003	Basenji	China	SAMN06159689	[10]
VillDog_Borneo	Village dog	Borneo	SAMN02485592	[9]
VillDog_China	Village dog	China	SAMN01974490	[4]
VillDog_Egypt	Village dog	Egypt	SAMN02485582	[9]
VillDog_India	Village dog	India	SAMN02485586	[9]
VillDog_Namibia	Village dog	Namibia	SAMN02485598	[9]
VillDog_Vietnam	Village dog	Vietnam	SAMN02485609	[9]
Golden Jackal	Asian Jackal	Israel	SAMN03366713	[5]

[3] – Lobo et al. Chapter III; [4] – Wang et al. 2013; [5] – Freedman et al. 2014; [6] – Fan et al. 2016; [7] – Zhang et al. 2014; [8] – Botigue et al. 2017; [9] – Auton et al. 2013; [10] – Marsden et al. 2016; [11] – Plassais et al, 2019.

Abbreviation: N-A, North American; NA, not available.

Supplementary Table 4.4. Expression profiles of the candidate gene set. Results from differential expression analysis for the three comparisons: non-Iberian vs. Iberian wolves, non-Iberian wolves vs. dogs, Iberian wolves vs. dogs. The direction of expression is presented in relation to the last condition in each comparison and is expressed by the log2 fold-change value. P-adj values below 0.05 are in bold. "Removed" refers to genes that were excluded from a specific comparison due to high levels of intra-condition variation.

Group	Gene	Non-Iberian vs. Iberian		Non-Iberian vs. Dogs		Iberian vs. Dogs	
		log2 FC	p-adj	log2 FC	p-adj	log2 FC	p-adj
Serotonin receptors	HTR1A	1.658	**1.9e-27**	-0.568	0.331	-2.200	**2.1e-15**
	HTR6	-1.257	**6.5e-09**	-0.968	**0.006**	0.326	0.267
	HTR3A	-1.433	**1.2e-04**	0.048	0.951	1.519	**2.9e-04**
Glutamate receptors	GRIK1	1.501	**1.7e-28**	0.518	0.079	-0.947	**2.6e-08**
	GRIK3	-1.118	**4.4e-08**	-0.798	0.155	0.344	0.332
	GRM2	-1.593	**4.1e-50**	-1.600	**1.6e-10**	0.025	0.901
	GRM6	-2.159	**2.7e-08**	-1.728	**0.006**	0.456	0.488
	GRM3	1.108	**1.2e-22**	0.470	0.171	Removed	Removed
	GRIN2A	-1.299	**5.1e-09**	-2.207	**0.002**	-0.895	0.053
	GRIN2D	-1.403	**1.6e-12**	-0.168	0.730	1.266	**1.5e-07**
	GRIA4	1.218	**2.4e-35**	-0.197	0.650	-1.389	**1.3e-12**
	GRID2	1.348	**5.8e-29**	-0.453	0.400	-1.786	**1.3e-11**
GO categories of behavior and locomotory behavior	DOCK2	1.199	**1.6e-28**	1.149	**4.8e-04**	-0.005	0.984
	CXCL10	2.719	**1.9e-07**	2.842	**0.002**	0.156	0.801
	ERC2	1.715	**2.1e-31**	-0.459	0.504	-2.156	**2.2e-10**
Introgression from dogs	MAST4	-1.421	**2.0e-20**	-1.469	**0.021**	-0.032	0.950

Supplementary Table 4.5. Genes overlapping SNPs with high ΔAF. List of the 181 genes that contained SNPs with high allele frequency change (|ΔAF| > 0.83) between Iberian and non-Iberian wolves and overlapped with the list of DEGs between the two groups.

Gene ID	Chromosome	Gene start (bp)	Gene end (bp)
ECHDC1	1	66012782	66073000
PTAR1	1	87706760	87752012
SPTLC1	1	94997290	95064468
ZNF628	1	102225674	102228781
VSIG10L	1	105630790	105638922
ALDH16A1	1	107032192	107047197
DHX34	1	108521685	108634105
CCDC8	1	109352047	109353612
PPP1R37	1	110352697	110394757
CLASRP	1	110412340	110435198
SPTBN4	1	113168484	113240934
GRAMD1A	1	117515393	117537943
MINDY3	2	20865946	21014429
SETD9	2	44046605	44073536
SMN	2	54596008	54636804
ZNF319	2	58551659	58557037
KIAA0825	3	14893203	15249730
NIPA2	3	32062708	32094491
MPHOSPH10	3	38060743	38080242
BLM	3	53408217	53473146
DOK7	3	60869206	60897318
SLC26A1	3	91485270	91534115
AIFM2	4	21087339	21104172
PLCXD3	4	68014162	68198283
ZMIZ1	4	28854908	29045986
NT5M	5	42080645	42117741
PLEKHN1	5	56194583	56202293
MON1B	5	73976545	73998255
LEPROT	5	44845581	44877144
GRIA4	5	26272970	26642100
WFDC1	5	68067885	68096067
CA7	5	82446246	82455590
PITPNM3	5	30754963	30819443
VWA5A	5	10325266	10362641
DNTTIP2	6	55235542	55263728
ZSCAN25	6	9931501	9939584
CTBS	6	63308420	63342266
ZNF646	6	17148403	17156226
ADGRL4	6	68078246	68220743
RCSD1	7	30756645	30852818
TBX19	7	30185850	30365512
CENPF	7	12555143	12617996
ROCK1	7	66862608	67142853
MYBPH	7	36428	43803
TOR3A	7	20738161	20750635
ZFYVE21	8	71451999	71461938
TEDC1	8	72788592	72795326
GBGT1	9	51274686	51284122
MRPS2	9	51211627	51214475
SERPINF1	9	45891410	45902765
MRM1	9	36474268	36478562
GPRC5C	9	5983542	5999068
DPH7	9	48207392	48220047
LRRC8A	9	54734505	54756780
NLE1	9	38222051	38365621
TEPSIN	9	854319	863290
FAM222B	9	42944687	43017420
FIBCD1	9	53073027	53103937
NDUFA4L2	10	1362577	1364515
YEATS4	10	11364553	11382645
TGFBRAP1	10	38095547	38157506
NAGA	10	23304354	23396068
ARFGAP3	10	22658610	22724377
JADE2	11	22648221	22698448

TNC	11	69077535	69170281
ARG1	12	242714	255305
ITPR3	12	2805187	3048938
RGL2	12	2731983	2739032
AFG1L	12	65313140	65515992
PFDN6	12	2729562	2737163
QRSL1	12	63972426	63991309
MLIP	12	21502561	21757764
ASCC3	12	58592025	58932615
EXOSC4	13	37574527	37577380
NUDCD1	13	9815957	9895359
SLC45A4	13	35707692	35748183
COMMD8	13	43369173	43383094
ARHGAP39	13	37947888	37993351
TEC	13	43978573	44062795
CAV2	14	55437622	55445469
AKR1B1	14	2991888	3006940
CYP51A1	14	17746656	17766658
LSM6	15	44651830	44666418
MARCHF1	15	59753216	60036060
HPDL	15	15027472	15028671
CCDC59	15	24364779	24373419
SZT2	15	16728758	16780325
P3H1	15	603329	621568
TNS3	16	278957	431704
PRAG1	16	36053967	36103764
CHPF2	16	15271805	15280024
UBXN8	16	33591273	33679673
REPIN1	16	14669587	14690502
CD160	17	58555851	58563341
CEBPZ	17	29565722	29584874
TPC3	17	35163732	35203659
NHLH2	17	53359258	53458795
ZNF512	17	21505655	21646364
ZNF513	17	21332621	21335577
AGBL5	17	21048423	21064868
SLC35F6	17	20793964	20806775
ZNF638	17	51251520	51318590
DUSP11	17	49245496	49278767
SLC16A1	17	62622902	62787481
RAB3IL1	18	54484177	54547523
LDLRAD3	18	31903095	32061171
RNF150	19	1817164	2053846
RAB3GAP1	19	37861804	37963501
TNNC1	20	37311966	37317321
CHDH	20	36150229	36187068
SLC35E1	20	45979369	45996469
FEM1A	20	54973083	54975104
KIF15	20	43359206	43532219
PNPLA6	20	52318408	52348963
SBNO2	20	57267724	57640750
KLHL26	20	44404352	44432472
PPP4R2	20	19118124	19167264
UPF1	20	44236264	44270916
NAV2	21	41458984	42177813
KLHL1	22	24191778	24555120
PCID2	22	60602888	60623719
SIAH2	23	45341199	45385546
GMPS	23	49569531	49699199
TPD52L2	24	47353573	47369356
BPI	24	26767746	26800762
SRC	24	25999048	26018077
KAT14	24	4933069	4968216
FAM210B	24	41574327	41588282
MPHOSPH8	25	18392409	18443241
DOCK5	25	32046527	32283658
TRIM35	25	30276763	30300769
NUP58	25	14106187	14158464
SNED1	25	51057376	51135412
CAPN10	25	50555425	50566135
SLC2A11	26	28652974	28673724
AP1B1	26	22630670	22689996
GAS2L1	26	22615319	22621311
RNLS	26	38083371	38414033

HPS4	26	19933281	20157058
TPCN1	26	10599780	10660767
SRRM4	26	15057392	15318325
LRTM2	27	43877684	43887595
YBX3	27	35043699	35066227
KCNA5	27	39756363	39811035
RARG	27	1958204	1978476
STRAP	27	30675247	30817388
STK38L	27	20429372	20510059
REP15	27	20029545	20030255
CALHM1	28	15814429	15825924
SFRP5	28	10993930	10999564
SEMA4G	28	13662686	13672359
ZNF511	28	40881766	40886090
INPP5A	28	40405882	40566643
RGS20	29	5521641	5605405
TSPYL5	29	41503312	41503956
MSC	29	20800062	20802657
LRRCC1	29	31678261	31715611
ZFAND1	29	28808423	28825748
ARMC1	29	15277020	15314695
ASPH	29	11814554	12022680
LEO1	30	17598076	17651290
CGNL1	30	22612828	22718493
ANKDD1A	30	29309093	29353793
CLK3	30	37706645	37719684
ETFA	30	39848270	39964823
URB1	31	27087570	27158164
PSMG1	31	33914668	33925883
C2CD2	31	36274539	36348657
WDR4	31	37085779	37106788
FRAS1	32	2558354	2991063
METAP1	32	21192948	21239996
PAPSS1	32	28136909	28235484
SNX4	33	28419128	28469488
SKIL	34	34871574	34894854
TPMT	35	16929240	16953229
RFTN2	37	7074586	7152714
SGO2	37	9611939	9767044
ATG9A	37	25768275	25890811
OBSL1	37	26059713	26079488
CDC73	38	5805916	5938111
RRP15	38	13377571	13421810

Supplementary Table 4.6. Candidate genes overlapping SNPs with high ΔAF. List of the 20 genes within the 181 genes (Supplementary Table 4.5) associated with neurological processes that could be potential candidates. Information about the position of the SNP with high ΔAF, derived allele, frequency of the derived allele calculated for the RNA-seq in Iberian and non-Iberian wolves, and the genomic consequence associated with each SNP is presented. SNPs with allele frequencies changes not confirmed with WGS data, or those missing in the WGS dataset, are indicated.

Gene ID	SNP position	Derived allele	Freq Iberian wolves	Freq non-Iberian wolves	Genomic consequence
GRIA4	26384147	C	1.00	0.17	Synonymous
CA7	82446473	T	0.00	0.83	UTR - not confirmed with WGS
ZSCAN25	9935192	C	0.00	0.92	Synonymous - not confirmed with WGS
GPRC5C	5986419	C	1.00	0.00	3' UTR variant
	5986577	C	1.00	0.00	3' UTR variant
LRRC8A	54741758	T	0.88	0.00	Missense - not found in WGS
TNC	69089523	T	0.88	0.00	Synonymous - not confirmed with WGS
SLC45A4	35709056	C	0.13	1.00	Synonymous - not found in WGS
ARHGAP39	37961783	A	1.00	0.00	Synonymous
HPDL	15027586	C	1.00	0.10	Synonymous - not confirmed with WGS
	15028307	C	1.00	0.13	Missense - not found in WGS but LD with previous did not confirm
SZT2	16743755	T	1.00	0.00	Synonymous
	16747446	T	1.00	0.00	Missense - not found in WGS but LD with previous confirmed
PRAG1	36094832	A	0.88	0.00	Synonymous - not found in WGS
	36095078	G	0.88	0.00	Synonymous - not found in WGS
DUSP11	49261830	A	1.00	0.17	Intronic
ZNF638	51282781	G	0.88	0.00	Synonymous - not found in WGS
NHLH2	53360278	G	0.88	0.00	3' UTR variant
RAB3GAP1	37957963	G	1.00	0.17	Intronic
PNPLA6	52327832	A	0.88	0.00	Intronic
NAV2	42109680	G	1.00	0.00	Synonymous
KLHL1	24197929	A	0.88	0.00	Synonymous
SIAH2	45361964	A	1.00	0.00	Intronic UTR - not found in WGS
	45364609	A	0.00	0.90	Intronic UTR - not confirmed with WGS
	45364843	G	0.88	0.00	Intronic UTR - not confirmed with WGS
WDR4	37097057	A	1.00	0.08	Synonymous

Chapter V.

Linking effects of a wolf population bottleneck to historical and contemporary genomic signatures of diversity and admixture with domestic dogs

Lobo D. et al.

(in preparation to be submitted)

Abstract

Ecosystem degradation and human persecution have been major global threats for large carnivores. Iberian wolves despite having survived the massive extirpation in Europe over the last two centuries, their population decreased considerably during the 1970s, leading to almost complete extinction of wolves in the South of the Iberian Peninsula. During such declines, events of hybridization with domestic dogs are believed to increase in frequency. Low rates of contemporary wolf-dog hybridization have been documented in Iberia; however, little is known about its frequency in historical times. In this study, we analyzed 68 contemporary and 54 historical samples, before and after the population bottleneck, to better understand how disturbance has affected the genomic composition of the Iberian wolf population, focusing on the levels of genetic diversity and dynamics of wolf-dog hybridization. A genome-wide assessment revealed that hybridization rates (3.7%) in the historical population did not increase in response to the population bottleneck and that wolves and dogs formed two well-defined groups. An increased number of hybridization events was found in the contemporary population (19.1%), showing a high prevalence towards the limits of the wolf distribution range. We estimated a low amount of dog ancestry (< 2%) across the genome of wolves from the historical population, suggesting that dog introgression did not play a role in the extinction of wolves in the South. Our findings also indicated a loss of genetic diversity in the contemporary population, suggestive of a rapid and human-induced decline. Still, we found active maintenance of variation in genes associated with immune response in the contemporary population, indicating that the loss of genetic diversity did not occur randomly. Given the increased human-induced environmental changes, our findings can provide important insights to predict future evolutionary responses in other large carnivores.

5.1. Introduction

Large carnivores have a long history of human direct persecution and habitat fragmentation unraveled by their current reduced abundance and fragmented distribution (Chapron et al. 2014; Ripple et al. 2014). Gray wolves (*Canis lupus*), which were once widely distributed across all Holarctic realm (Boitani 2003), suffered such a dramatic decline in the European continent that vanished from most of its Western/Central range, remaining only in small populations localized in Southern European Peninsulas (Blanco et al. 1992; Boitani 2003; Randi 2011). Smaller and isolated populations are expected to experience more severe effects of genetic drift and inbreeding, causing genetic impoverishment that can ultimately drive a species to extinction (Lande 1994; Charlesworth and Willis 2009).

The Iberian Peninsula, which includes Portugal and Spain, holds one of these isolated wolf populations that live today in a highly human-dominated landscape. The Iberian wolf population, ascribed to its own subspecies *C. l. signatus*, became isolated from the other European populations circa 10.3 thousands of years ago (ka) (Silva et al. 2020). The population is essentially distributed in the Northwestern region of Iberia (**Figure 5.1A**) and is currently recognized as at lower risk/conservation dependent (Rede Natura 2000, ICNF). However, the Iberian wolf population represents today a relic of a population that was historically distributed across most of the Peninsula (Clavero et al. 2022; Nores and López-Bao 2022). Indeed, the Iberian wolf suffered a dramatic decline during the last two centuries, reaching its minimum in the 1970s (Valverde 1971; Garzón 1979; Sastre et al. 2011). Intensive human persecution led to the extinction of the wolf in the entire Central and Southern regions of the Iberian Peninsula, with only two isolates persisting: i) a small population at the South of Douro River in Central Portugal, and ii) a currently functionally extinct population in Sierra Morena, Southern Spain (Álvares et al. 2005; Blanco and Cortés 2007; Gómez-Sánchez et al. 2018; López-Bao et al. 2018). In the North, wolves have persisted and have been in the last two decades expanding Southeastwards in Spain (Blanco and Cortés 2009; Chapron et al. 2014; López-Bao et al. 2018). Particularly, wolves have successfully persisted and expanded in intensively modified agricultural areas, such as those found in Castilla y León region, in Central Spain (Blanco and Cortés 2009).

Genome-wide studies have shown evidence of low levels of genetic diversity in the Iberian wolf in relation to Eastern European wolves (Lobo et al. Chapter III; Pilot et al.

2014; Fan et al. 2016), which is consistent with historical records of a population bottleneck (Clavero et al. 2022; Nores and López-Bao 2022). Additionally, historical and contemporary events of wolf-dog hybridization have also been documented in this wolf population, either considered a conservation threat or a promoter of adaptation (Lobo et al. Chapter III; Godinho et al. 2011; Godinho et al. 2015; Fan et al. 2016; Pacheco et al. 2017; Torres et al. 2017; Gómez-Sánchez et al. 2018). Notably, an in-deep genomic analysis of one of the last wolves from Sierra Morena revealed massive levels of introgression from domestic dog (*C. l. familiaris*), suggesting that the population decline may have favored recurrent events of wolf-dog hybridization (Gómez-Sánchez et al. 2018).

Anthropogenic hybridization between wolves and dogs has been a topic of long scientific debates focused essentially on its possible major conservation concerns (Hindrikson et al. 2017; Donfrancesco et al. 2019; Salvatori et al. 2020). The introgression of dog genes into the wolf gene pool is generally seen as a threat as it may dilute the genetic distinctiveness of wolves and disrupt behavior and local adaptations to the environment, increasing the risk of extinction through introgressive hybridization (Allendorf et al. 2001). There is evidence that hybridization between wolves and dogs correlates with disruption of social structure in wolf packs and that it is more frequent in peripheral areas of the wolf range, particularly in human-dominated landscapes where wolf populations are generally fragmented and at low densities (Godinho et al. 2011; Galaverni et al. 2017; Hindrikson et al. 2017) being outnumbered by free-ranging dogs (Randi and Lucchini 2002; Pilot et al. 2021). Several cases of wolf-dog hybridization have been reported in Europe over the past years (Vilà et al. 2003; Godinho et al. 2011; Hindrikson et al. 2012; Randi et al. 2014; Galaverni et al. 2017; Pacheco et al. 2017; Kusak et al. 2018; Pilot et al. 2018; Smeds et al. 2021) but little is known about the rate and the extent of hybridization in historical times.

In this work, we aim to understand how the genomic composition of Iberian wolves, particularly the levels of diversity and dynamics of hybridization with dogs, was affected following the bottleneck in the 20th-century. We hypothesize that the dramatic decline and population contraction within Central and Southern regions of Iberia have facilitated an increase in hybridization events and depleted genetic variability, which ultimately contributed to wolf extinction. On the other hand, the persistence and ensuing expansion of wolves in the Northern range of the species would have allowed the population to maintain stable levels of genetic diversity in this region. To approach this, we analyzed thousands of single nucleotide polymorphisms (SNPs) across contemporary and

historical Iberian wolf samples representing the periods before and following the bottleneck. We first inferred rates of wolf-dog hybridization within both historical and contemporary populations and then characterized the amount of dog introgression in the genome of historical samples over multiple timescales. We also quantified levels of genetic diversity in both populations, searched for hotspots of genetic diversity within the genome, and identified genes within high polymorphic regions to provide empirical evidence of non-random loss of genetic diversity under sharp population declines.

5.2. Materials and Methods

5.2.1. Historical and contemporary SNP dataset

In this work we leveraged the availability of two comprehensive genome-wide datasets of Iberian wolf consisting of i) 145 historical samples genotyped for 100,000 genome-wide SNPs targeted based on the Canine HD BeadChip microarray (Illumina, San Diego, CA, USA) using an in-solution target capture enrichment approach (Pacheco, Lobo et al. 2022; Chapter II), and ii) 68 contemporary wolf samples and 43 dogs (23 village dogs and 20 livestock guarding dogs) from the Iberian Peninsula that were genotyped for 170,000 genome-wide SNPs with the Canine HD BeadChip microarray (Lobo et al. Chapter III).

First, we filtered the historical dataset to keep only samples with more than 6,000 SNPs, and high call rates per sample (> 0.85) and per locus (> 0.90) using PLINK v.1.9 (Purcell et al. 2007). This resulted in 54 individuals with 5,173 SNPs. Then, we merged historical and contemporary datasets also using PLINK and the same filters for high call rates were applied. Finally, we removed loci with minor allele frequency (MAF) below 0.01. This final dataset comprises 54 historical and 68 contemporary non-related Iberian wolves and 43 dog samples genotyped to 4,470 SNPs (hereafter 5K dataset; Supplementary Table 5.1). Historical samples cover the former wolf range distribution in the Iberian Peninsula (N = 19 and N = 35 from the contracting population range in Central and Southern regions and the stable/expanding population range in Northwestern Iberia, respectively; **Figure 5.1A**), and the collection date ranged from 1912 to 1991. This dataset was used throughout the analyses described below unless it is stated otherwise.

To increase the resolution over the genome in different analyses (see below), we created two additional datasets containing a higher number of SNPs. One was set to a minimum

threshold of 20,000 SNPs in historical samples, resulting in a dataset with 47 historical wolf samples (15 from the contracting and 32 from the stable/expanding population ranges), 68 contemporary wolf samples, and 43 dogs, with 19,754 SNPs (hereafter 20K dataset). The other was limited to historical samples with more than 80,000 SNPs, resulting in a dataset with 22 historical wolf samples (10 from the contracting and 12 from the stable/expanding population ranges), 68 contemporary wolf samples, and 43 dogs, with 61,896 SNPs (hereafter 60K dataset).

5.2.2. Global ancestry estimation and hybrid identification

Levels of genome-wide differentiation (F_{ST}; Weir and Cockerham 1984) between Iberian wolves and dogs, and between contemporary and historical Iberian wolf samples, were assessed with VCFtools v.0.1.15 (Danecek et al. 2011). Average genome-wide global ancestries across all samples were firstly explored by a Principal Component Analysis (PCA) using PLINK. We then used the maximum likelihood model implemented in ADMIXTURE v.1.3 (Alexander et al. 2009) to estimate the ancestry proportions (q) in K = 2 across historical/contemporary wolf populations and dogs. ADMIXTURE was run separately for historical and contemporary wolf samples in 2000 interactions using a cross-validation procedure of 10-fold (Alexander and Lange 2011). We performed 10 independent runs and used q values from the run presenting the highest likelihood.

To establish the q thresholds for hybrid identification, we simulated both parental and hybrid genotypes in HYBRIDLAB v.1.0 (Nielsen et al. 2006). Genotypes of individuals with q values ≥ 0.98 (estimated by ADMIXTURE) were selected as reference for each parental group (historical/contemporary wolf populations and dogs) for the training data set, to generate 200 of each parental genotype. The number of reference individuals was uniformized across each parental group (N = 31) to avoid bias during allele frequencies estimation, as recommended in HYBRIDLAB documentation. Input ped files were converted into genepop format using PGDSpider v.2.1.1.5 (Lischer and Excoffier 2012). Simulated parental genotypes were then used to simulate 200 of each F1, F2, first and second backcross genotypes in both wolves and dogs. This procedure was performed independently for historical and contemporary wolf samples. The q values of the simulated genotypes were estimated in ADMIXTURE as previously described, and the threshold for hybrid identification was defined based on the lower confidence interval observed.

We used the *f3* test in THREEPOP (TREEMIX v.1.13 package; Pickrell and Pritchard 2012) to statistically test if the identified hybrid individuals were the product of admixture between wolves and dogs using the 20K dataset (this dataset did not exclude any the hybrids identified using the 5K dataset in the historical population). *f3* statistics is a population-based test and uses minor allele frequency estimates as input. We defined three groups to calculate minor allele frequency in PLINK: parental wolves (historical and contemporary), parental dogs, and hybrids (historical and contemporary). The python script *plink2treemix.py* that is provided by the TREEMIX package was subsequently used to create the input file for the test. *f3* test was run under the following phylogenetic conformation (A, B; C): A – parental wolves, B – parental dogs, C – hybrids; using blocks of 200 SNPs (N = 98) and considering Z-score values < –3 as indicative of admixture.

5.2.3. Local ancestry analysis across hybrid genomes

To estimate with higher accuracy the proportion of dog ancestry within hybrid genomes, we used the local ancestry method LAMP-ANC v.2.5 (Sankararaman et al. 2008). LAMP-ANC is a non-LD-based method that estimates within windows the most likely ancestry per SNP given the reference allele frequencies. We ran the analysis with the 20K dataset, considering as admixed population all the identified hybrids and using both parental wolves and dogs as reference populations. We set the number of generations since admixture to 10, considering a wolf generation time of 4.5 years (Mech et al. 2016); a recombination rate of 9.7×10^{-9} (Wong et al. 2010; Campbell et al. 2016); a mixture proportion of 0.81:0.19 to the wolf and dog population, respectively (based on ADMIXTURE estimates); and a linkage disequilibrium (LD) cut-off of $r^2 > 0.1$, because LAMP-ANC assumes unlinked markers.

To analyze chromosomal permeability to dog introgression within hybrid genomes, we applied the SNP-specific delta (Δ) statistics, that for each hybrid individual it subtracts locus-specific ancestry from average genome-wide ancestry. Values are then average across all hybrid individuals. Positive values are indicative of genomic regions with higher dog ancestry than the genome-wide, and therefore can be indicative of higher permeability to, or retention of, dog introgression. We calculated Δ ancestry across all chromosomes, using local estimates of dog ancestry determined by LAMP. Outlier regions were defined as those with Δ scores above two standard deviations from the mean chromosome (mean+2SD). Only hybrids resulting from backcrossing with wolves (i.e., not F1s) were used in this analysis.

5.2.4. Local ancestry analysis across the historical population genomes

To understand whether population decline and concomitant extinction in Central and Southern Iberian wolves were associated with increased homogenization of the wolf genome from extensive wolf-dog hybridization followed by introgression, we used LAMP-ANC to estimate the proportion of dog ancestry in the historical population using the 60K dataset. Given that we were interested in estimating levels of dog introgression in the wolf genome and not in recent events of hybridization, the previously identified hybrids were excluded from this analysis. LAMP-ANC was run considering the Iberian wolves from the historical population as admixed, and the contemporary wolf samples and dogs as reference populations. An equal sample size of reference wolves and dogs was used (N = 43). We set all the parameters as previously described, except for the mixture proportion that was defined as 0.99:0.01 to the wolf and dog population, respectively.

5.2.5. Genetic diversity analysis

To explore levels of genetic diversity in the historical and contemporary Iberian wolf populations, we determined the proportion of heterozygous sites, LD decay curves, and number and length of runs of homozygosity (ROH) as indicative of recent inbreeding. Hybrids were not included in these analyses.

The proportion of heterozygous sites was determined with the 5K dataset (without the MAF filter applied; 4,613 SNPs) to maximize the number of individuals included in the analysis, using the –het function in PLINK. We used an equal sample size for historical and contemporary populations to avoid bias (N = 52). Significant differences ($p < 0.05$) between population heterozygosity average values were assessed through a Mann-Whitney/Wilcoxon rank-sum test in R (R Development Core Team 2017). We also estimated the distribution of heterozygosity across the genome of individuals from both populations, using the –hardy function in PLINK. For this, we used the 60K dataset without the MAF filter applied (63,167 SNPs), which included 22 historical samples. We randomly selected the same number of contemporary samples to avoid bias. Values of heterozygosity were averaged in windows of 10 SNPs (covering ~350 kb), and high variation windows were defined as those above the 99^{th} percentile. High variation windows were checked for possible enrichment in Gene Ontology (GO) terms and biological pathways (KEGG) in g:Profiler (Raudvere et al. 2019). The analysis was carried out considering the Ensembl annotation version 104 of CanFam3.1 reference genome, the ordered query set as true and the significance threshold used was > 0.05

after g:SCS (Set Counts and Sizes) false-discovery rate correction that accounts for multiple testing in the overlap of functional terms.

Estimates of LD decay were calculated independently for each population using the 60K dataset without the MAF filter applied, based on pairwise r^2 values between all pairs of SNPs within a 1 Mb window. This analysis was carried out in PLINK using the –r2 function. Values were then averaged in non-overlapping 1 kb and 100 bp windows for plotting.

We considered ROH segments longer than 300 kb spanning at least 50 SNPs and allowing for a maximum of one heterozygous SNP (Ceballos et al. 2018). Given that ROH detection can be highly biased by LD pruning and MAF removal (Meyermans et al. 2020), we used the same 60K dataset as used previously for heterozygosity and LD analysis. ROH detection was performed in PLINK using the function –homozyg with the parameters described above and set the options that control for SNP density, sliding-window length, window gap, missing calls, and window threshold as default. Inverse cumulative average counts of ROHs per individual were calculated for every 1 Mb non-overlapping windows up to 60 Mb, for plotting. All the plots were generated in R with the *ggplot2* package.

5.3. Results

5.3.1. Levels of admixture and frequency of hybridization

The genome-wide genetic differentiation between Iberian wolves and dogs was estimated at F_{ST} = 0.29. Global ancestry analysis in ADMIXTURE and the PCA clearly differentiated Iberian wolves and dogs into two distinct groups. The first component in the PCA segregated dogs from wolves, explaining 65.9% of the variation (**Figure 5.1B**). Historical and contemporary populations group together in the PCA with no signs of differentiation between the two, which is compatible with the negligible levels of genome-wide differentiation between them (F_{ST} = 0.02). Using ADMIXTURE, the average assignments of Iberian wolves to the corresponding cluster at K = 2 were q_W = 0.99 and q_W = 0.96 for historical and contemporary populations, respectively (**Figure 5.1C**). Dogs were assigned to their corresponding cluster with an average ancestry proportion of q_D = 0.98 (**Figure 5.1C**). Simulated genotypes of parental individuals were assigned to their corresponding cluster with q values of [0.99 – 0.99] for wolves and [0.96 – 0.99] for dogs.

Based on the minimum value observed, we defined as hybrids all the individuals with average ancestry values below 0.99 and 0.96 for wolves and dogs, respectively.

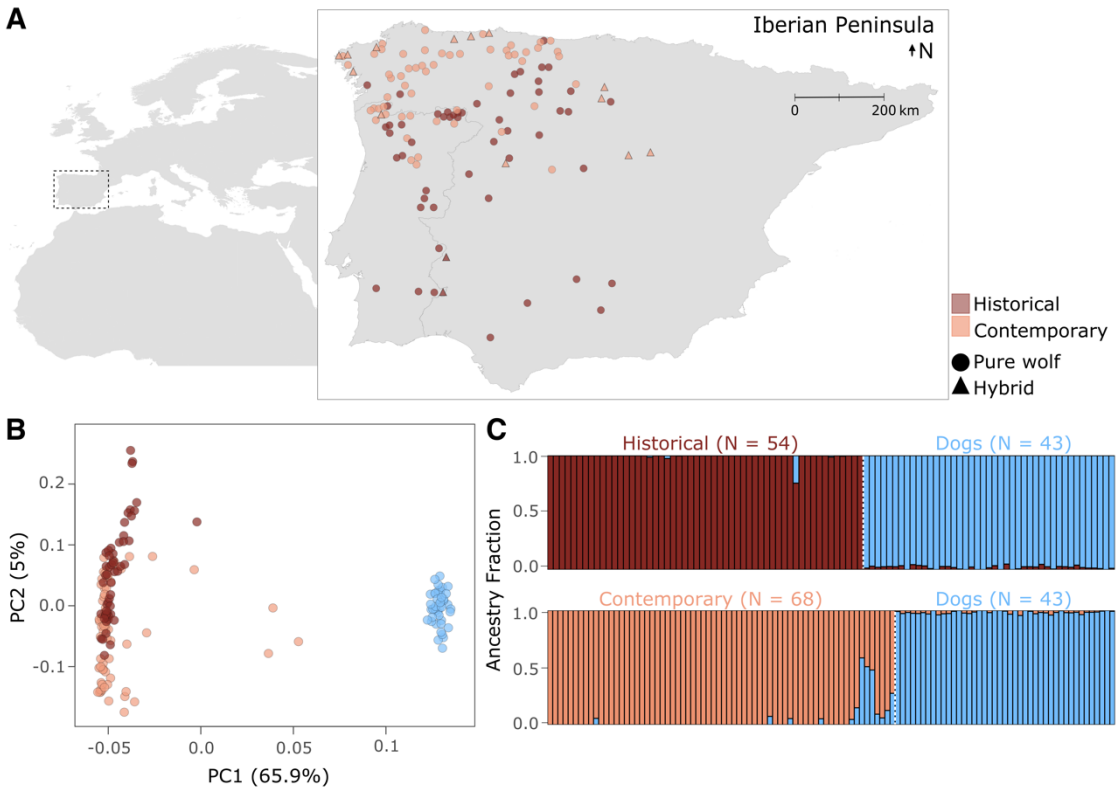


Figure 5.1. Sampling location and admixture levels. (**A**) Geographical location of the analyzed wolf samples in the Iberian Peninsula (delimited by the dashed line in the left map). Historical (N = 54) and contemporary (N = 68) samples are represented in red and orange, respectively (the same color scheme was followed in panels B and C). Circles and triangles denote pure wolves and hybrids, respectively. Further details about samples can be assessed in Supplementary Table 5.1. (**B**) Principal component analysis (PCA) for historical and contemporary wolves and dogs (N = 43; represented in blue). Each dot denotes an individual. Principal components 1 and 2 are displayed with the corresponding variance values. (**C**) Ancestry fractions of all individuals inferred in ADMIXTURE for historical/contemporary wolves and dogs in K = 2 clusters. Historical and contemporary wolf samples were analyzed independently. Individuals are represented by a single bar and different colors indicate the ancestry proportion attributed to each cluster. White dotted lines separate wolves from dogs.

Among 54 historical samples, we detected two hybrids (3.7%) with a dog ancestry proportion of q_D = 0.03 and q_D = 0.24 (**Figure 5.2A**; Supplementary Table 5.2). These two individuals were localized in the South of Iberia where the population is currently extinct (**Figure 5.1A**) but were originally collected sparse in space and time in the districts of Beja (Barrancos 1959) in Portugal and in the province of Badajoz (Villanueva del Fresno 1972) in Spain. The latter was already recorded as a putative wolf-dog hybrid

in the Doñana Biological Station collection. Within the contemporary population, we found a higher prevalence of hybrids (N = 13; 19.12%), with dog ancestries ranging from q_D = 0.05 – 0.58 (**Figure 5.2A**; Supplementary Table 5.2). Three samples exhibiting admixture proportions of ~0.50 represent F1 hybrids. All hybrids from the contemporary sampling, except one in Portugal, were found in the species marginal distribution range in Spain. Hybrid individuals were scattered along the first axis in the PCA between the wolf and dog clusters (**Figure 5.1B**), in accordance with the admixture proportion detected. Results of the *f3* test confirmed a highly significant admixture between wolves and dogs across the 15 hybrids (*f3* = -0.015; Z-score = -41.92).

5.3.2. Local ancestry analysis of hybrids

Overall, local ancestry analysis attributed identical values of dog ancestry in hybrids as those estimated by ADMIXTURE (**Figure 5.2A**). In the case of the three F1 hybrids identified in the contemporary population, we observed the expected pattern of homologous chromosomes having ancestry attributed to each parental species (Supplementary Figure 5.1). As expected, hybrids of further generations (backcrosses) showed fewer and shorter dog blocks in their genome (Supplementary Figure 5.1). Among the latter, we observed differences in the amount of dog genomic blocks retained per chromosome following backcrossing within the wolf genetic background (Supplementary Figure 5.2). The Δ ancestry statistics showed that chromosomes 32, 13, 31, 10, 18 and 26 were overrepresented with dog ancestry, whereas chromosomes 9, 5, 36 and 22 showed the lowest proportions of dog ancestry (**Figure 5.2B**). The average length of these overrepresented dog genomic regions was 9.2 Mb, and the longest fragment was found on chromosome 32 with 23.7 Mb.

5.3.3. Proportion of dog introgression in the historical population

Historical “pure” wolf samples from the contracting and stable/expanding population ranges showed low levels of dog introgression in their genome (average of 0.02 and 0.01, and maximum values of 0.04 and 0.03, respectively; Supplementary Figures 5.3, 5.4 and Supplementary Table 5.3), with few and short dog genomic blocks scattered across the chromosomes. Interestingly, the genome of four wolves captured in 1927, 1964, 1970, and 1991 in Sierra Morena carried a low amount of dog ancestry (0.04, 0.03, 0.03, and 0.02 respectively; **Figure 5.2C;** Supplementary Figure 5.3), with few and short dog blocks spread across the chromosomes, similar to those observed in the other

historical samples. Thus, there is no evidence suggesting an increase or retention of specific dog genomic blocks in the genome of Sierra Morena wolves up to the 1990s.

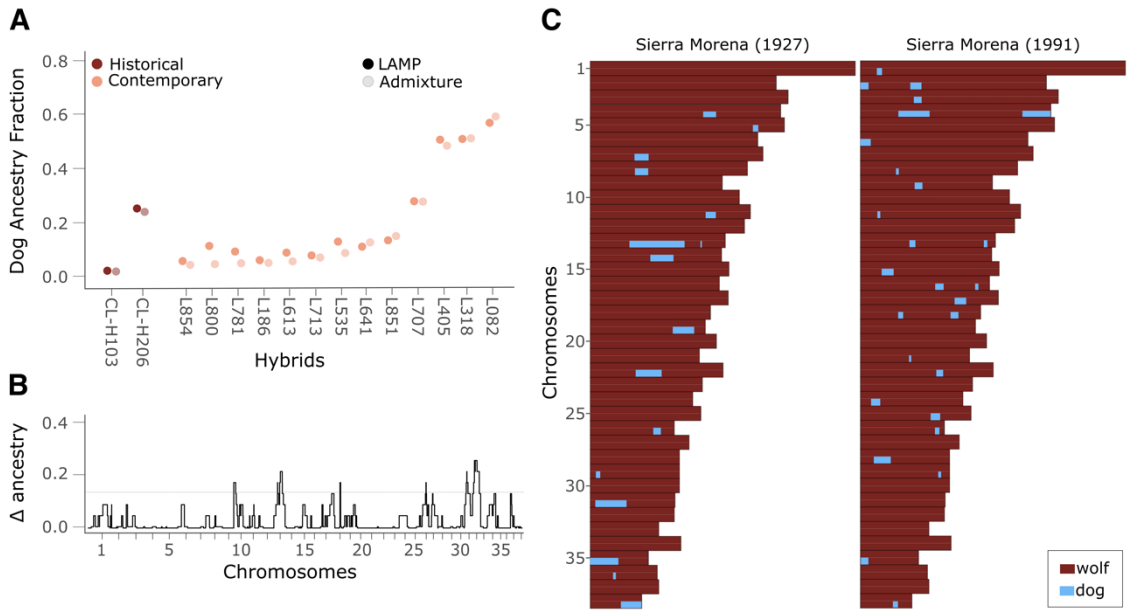


Figure 5.2. Local ancestry analysis and introgression levels. (**A**) Fraction of the global dog ancestry estimated for the 15 hybrids using LAMP-ANC (full dots) and ADMIXTURE (light dots). Historical and contemporary samples are represented in red and orange, respectively. Sample IDs are presented on the x-axis. (**B**) Δ ancestry scores indicating the excess of dog ancestry across the hybrid chromosomes (backcrosses only). The dashed line identifies the cut-off used to consider outlier regions (0.14; mean + 2SD). (**C**) Local ancestry in the genome of two wolves sampled in Sierra Morena in 1927 and 1991. Genomic blocks were colored based on the attributed local ancestry: red for wolf (both chromosomes), blue for dog (both chromosomes), and half red/blue for wolf/dog (one chromosome of each).

5.3.4. Patterns of genetic diversity

The historical population of Iberian wolves showed a significantly higher proportion of heterozygous sites relative to the contemporary population (het_{HIST} = 0.20 and het_{CONT} = 0.17, p-value = 3.65e-07; **Figure 5.3A**). Although the distribution of diversity in historical samples revealed a decrease trend period around 1970, there is no evidence of a linear decrease in heterozygosity across time or space (**Figure 5.3A**). However, a close inspection of the heterozygosity levels within samples from two geographical regions where wolves have persisted, one South of Douro, Central Portugal, and the other in Castilla y León, Northwestern Spain, revealed an apparent severe loss of diversity where the Iberian wolf population has contracted its range relative to areas where the population has been stable/expanding (heterozygosity loss of 0.04 and 0.02 for wolves

from South of Douro and Castilla y León, respectively; Supplementary Table 5.4). Unfortunately, the unbalance sample size (N = 6 for South of Douro; N = 23 for Castilla y León) hampers further statistical testing to assess the significance of these differences. The lowest levels of heterozygosity in the historical population were found in one sample from Sierra Morena collected in 1970 (het = 0.14). However, as the last sample in our dataset from this population is from 1991, which displayed high levels of heterozygosity (het = 0.21), it is difficult to predict how severe the loss of diversity was in the decades that followed the bottleneck without more recent samples.

We identified 94 and 90 high variation windows (> 99th percentile) in the genome of wolves from the historical (N = 22; het > 0.37) and contemporary (N = 22; het > 0.35) populations, respectively (**Figure 5.3B)**. Of those, 25 windows (~27%) were shared between the two populations. Within the 90 high variation windows in the contemporary population, there was a significant enrichment for GO terms associated with the major histocompatibility complex (MHC) proteins and antigen processing (genes involved: DLA-DRA, DLA-DQA1, TAP1, HLA-DQB2, and PIKFYV2; **Table 5.1**).

Table 5.1. Over-represented GO categories among the top 90 high variation windows found in the contemporary Iberian wolf population.

Category	Description	Genes (intersections)	adj p-value
	Biological process		
GO:0002504	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	PIKFYV2, DLA-DRA, DLA-DQA,1, HLA-DQB2, TAP1	0.02
	Cellular component		
GO:0042613	MHC class II protein complex	DLA-DRA, DLA-DQA1, HLA-DQB2, TAP1	1.92e-03
GO:0042611	MHC protein complex	DLA-DRA, DLA-DQA, HLA-DQB2, TAP1	3.4e-03

An elevated number of ROHs was detected in wolves from the contemporary population compared to the historical one (average per individual of 72.86 and 55.58, respectively; **Figure 5.3C**; Supplementary Table 5.5). The homozygous segments were, on average, also longer in the contemporary population (average size per segment of 7.28 Mb against the 6.02 Mb in the historical population), which also presented an elevated total amount of homozygosity in the genome (average total size of 567.38 Mb against the 351.45 Mb in the historical population). Nevertheless, almost all ROHs were shorter than 10 Mb

within both historical (83.84%) and contemporary (80.47%) populations, suggesting no strong inbreeding in the Iberian wolf. The longest homozygous segments (13.9 Mb) were detected within the wolf from Sierra Morena (1970), with a total ROH size in the genome of 847.74 Mb, which displayed the lowest levels of heterozygosity. However, no clear distinctive pattern was observed within historical samples that could indicate higher inbreeding in areas of contracting population range (**Figure 5.3D**). Increased LD levels were also found in the contemporary population (**Figure 5.3E**), where r^2 values only decayed below 0.5 at ~2 kb, contrasting with the ~1 kb in the historical population.

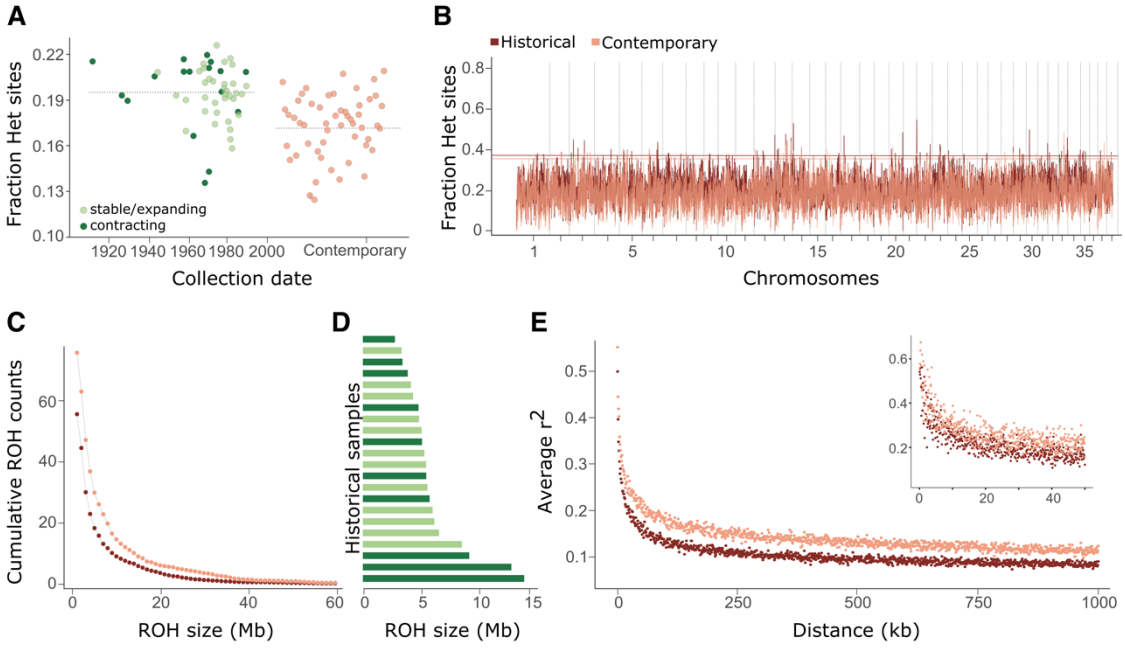


Figure 5.3. Genetic diversity in the historical and contemporary populations. (**A**) Proportion of heterozygous sites present in 52 individuals from the historical (green) and contemporary (orange) populations across time (year of the original collection). Historical samples from the contracting (N = 17) and stable/expanding (N = 35) are represented in dark and light green, respectively. Each dot denotes an individual and black dashed lines represent the average value in each population. (**B**) Distribution of heterozygosity across the genome of historical (red) and contemporary (orange) individuals (N = 22, in each), in non-overlapping windows of 10 SNPs (~350 kb). The red and orange lines identify the cut-off used to define high variation windows in each population. (**C**) Inverse cumulative average number of runs of homozygosity (ROH) per individual across 1 Mb non-overlapping windows relative to their average size, for each population. (**D**) Average size of ROH per individual (represented by a single bar) in the historical population (further information is available in Supplementary Table 5.5). (**E**) Decay of linkage disequilibrium across populations. Each dot represents the averaged pairwise r2 values between pairs of SNPs in non-overlapping 1 kb windows. The inset plot represents a zoom-in view of the first 50 kb, where each dot represents the averaged pairwise r2 values between pairs of SNPs in non-overlapping 100 bp windows.

5.4. Discussion

5.4.1. Dynamics of wolf-dog hybridization across a population bottleneck

Genome-wide analysis of Iberian wolf samples through the 20th-century revealed that the rate of wolf-dog hybridization in historical times (3.7%) did not increase in response to a population decline, contrary to what we have hypothesized. The two hybrids identified were localized in the South of Iberia in regions where the wolf population is currently extinct but were originally collected in distinct decades (1959 and 1972), suggesting a non-recurrent event nor confined to a time period of substantial population decline. The number of wolf-dog hybrids identified was higher in the contemporary population (N = 13; 19.1%), and almost all hybridization events (N = 12) took place in the limits of the distribution area in Spain, where cases of wolf-dog hybridization are known to occur (Godinho et al. 2011; Pacheco et al. 2017; Godinho et al. 2018), including five events in the expanding area in Castilla y León. Although this may not reflect an increase in hybridization over the two periods because of an unbalance effort toward hybridization cases for the contemporary sampling, the result is compatible with the dynamics of wolf-dog hybridization documented in Italy, where increased rates of hybridization were recorded two decades after the massive bottleneck in the 1970s when the population was re-expanding (Galaverni et al. 2017). The geographical distribution of the hybrids identified in the contemporary sampling towards the range border is consistent with reports of a higher incidence of wolf-dog hybridization in peripheral regions of the wolf range in other European populations (Andersone et al. 2002; Verardi et al. 2006; Godinho et al. 2011; Leonard et al. 2013).

Only five of the 15 hybrids identified had dog content above 20%, confirming that hybridization does not occur at high frequency and that Iberian wolves and dogs are two genetically well-differentiated entities (Lobo et al. Chapter III; Godinho et al. 2011). Another study covering the entire wolf contemporary distribution area has previously identified the same three F1 hybrids using microsatellite data (Godinho et al. 2011), retrieving very similar estimates of dog ancestry as those we calculated using the 5K SNP panel. However, our study extended the sampling and increased the resolution over the genome, resulting in a clear refinement in the assessment of hybridization prevalence that raised from 4% to 19% (Godinho et al. 2011).

Local ancestry analysis revealed similar levels of dog ancestry among hybrids as those estimated by global methods. However, the formation of hybrid genomes creates complex genomic signatures that cannot be fully tracked using global ancestries

approaches. As expected, we found an uneven distribution of dog genomic blocks across the 38 autosomes of backcross hybrids, suggesting a different capacity to retain or purge introgressed alleles, as has previously been shown in several mammals (Geraldes et al. 2006; Good et al. 2010; Turner and Harr 2014; Frantz et al. 2015; Pilot et al. 2021), birds (Carling and Brumfield 2008; Runemark et al. 2018) and fishes (Schumer et al. 2018). Interestingly, the chromosomes of hybrids with a significant excess of dog genomic content (e.g., chr 10, 13, 31, and 32) overlap with those previously found for the entire non-admixed Iberian wolf population (Lobo et al. Chapter III) and also coincide with those reported for Italian wolf hybrids (Galaverni et al. 2017). This result suggests that either genomic regions that are more permeable to introgression are also less exposed to purging mechanisms, or these are regions with ancestral shared variation between wolves and dogs where divergence is lower than the average. We found no differences within chromosomal rates of recombination in the dog genetic map (Campbell et al. 2016) or in levels of genetic differentiation between wolves and dogs (Pendleton et al. 2018) that could explain these results. This is in line with previous studies showing that the number of barriers to introgression between wolves and dogs is low, which can explain the lack of correlation between rates of admixture and recombination (Pilot et al. 2021). There is evidence that hybridization can produce genomes with different proportions of parental contribution due to genetic incompatibilities, epistatic interactions, and selection pressures (Runemark et al. 2018; Schumer et al. 2018). Therefore, the dog genomic content found introgressed in the wolf genome likely persisted in regions not linked to deleterious effects which limit the amount of variation in genes that play fundamental roles.

5.4.2. Dog introgression as a conservation threat to wolf populations

We did not find evidence of high levels of dog ancestry in any of the “pure” wolves from the historical population, suggesting that dog introgression was not a main driver of wolf extinction. The average proportion of 1.4% dog ancestry found in the historical population is in line with previous reports for the contemporary Iberian wolf population (Lobo et al. Chapter III) and other Eurasian populations (Fan et al. 2016; Pilot et al. 2021). Introgression patterns of few and short dog blocks in the genome of wolves from the Central and Southern regions of Iberia were similar to those observed in individuals from the Northwestern area and illustrate signatures of past admixture and not recent gene flow. Local ancestry methods, such as the one we used in this study, estimate ancestry proportions based on the reference panel (non-admixed individuals) available

(Sankararaman et al. 2008; Pasaniuc et al. 2009; Geza et al. 2019). Given that we used the contemporary non-admixed Iberian wolf population as a reference for this analysis, it is possible that not all the diversity in the historical population was represented, which can introduce bias during ancestry estimation and overestimate patterns of dog introgression.

Interestingly, the few dog blocks found in the genome of the four wolves from Sierra Morena suggest that hybridization was not occurring in high frequency in this region till the decade of 1990. Furthermore, we did not observe any increase or retention of dog genomic region from 1927 to 1991 that could link the decline of the wolf population to an increased rate of hybridization with dogs. Therefore, the fact that the genome of one of the last wolves from this population carries around a third of dog ancestry (Gómez-Sánchez et al. 2018) is not linked to a history of continuous interbreeding with dogs in the previous past decades. Instead, this might reflect signatures of recent hybridization and high inbreeding in a very small population such as Sierra Morena, before its functional collapse (López-Bao et al. 2015; Gómez-Sánchez et al. 2018; López-Bao et al. 2018). Given that we analyzed four wolf genomes, we cannot entirely exclude the possibility of not having enough representation of the population. Still, if hybridization was frequent at the time in such a small population, we would expect to find a high abundance of long dog blocks in the genome of the wolf captured in 1991, which was not the case. In this study, we did use a limited number of markers (~61K SNPs) to determine ancestry across the genome, which offers only a moderate resolution of the wolf genome. However, while ancient signatures of hybridization could have remained undetected, the SNP panel should not fail to identify hybridization of few generation with dogs (Galaverni et al. 2017; Pilot et al. 2018).

5.4.3. The non-random loss of genetic diversity

The decline of the Iberian wolf population during the 20th-century has significantly decreased the levels of genetic diversity exhibited today by the contemporary population. The distribution of individual heterozygosity across time seems to reveal a decrease concordant with the demographic bottleneck in the 1970s. However, we found no signs of a linear loss of heterozygosity throughout time across historical samples, as would be expected following a gradual population decline (Bürger and Lynch 1995; Blomqvist et al. 2010). This pattern suggests a rapid and drastic decline of the wolf population, especially in Central and Southern regions, which is compatible with estimates of the historical population range in the Iberia Peninsula (Clavero et al. 2022; Nores and López-

Bao 2022). Furthermore, contrary to the hybridization dynamics that reflect levels of gene flow at the time of sampling, it may take a few generations to see a decrease in heterozygosity following a population bottleneck (Nei et al. 1975; Shama et al. 2011). Hence, it is not surprising that samples from the Central and Southern regions collected during the 1970s displayed high levels of diversity. Still, by comparing levels of diversity in historical and contemporary samples from the same area, we found an apparent severe loss of diversity in regions where the population has contracted its range in relation to areas of expansion. Interestingly, the wolf from Sierra Morena sampled in 1991 showed high levels of heterozygosity and few ROHs in its genome. Previous studies estimating the wolf range in the past decades (Clavero et al. 2022; Nores and López-Bao 2022) have shown that the Sierra Morena population has split into small core areas around 1970 and that the most drastic decline only occurred in the 1990s (Blanco et al. 1990; López-Bao et al. 2015). This can explain the dissimilar patterns of genetic diversity we observed within samples from Sierra Morena, which likely belong to different core areas. Although our dataset does not include contemporary samples from Sierra Morena, the genome from the wolf sampled in 2003 showed genomic signatures compatible with a drastic decline, such as high levels of inbreeding and long ROHs (Gómez-Sánchez et al. 2018).

Historical levels of heterozygosity found in the Iberian wolf population were similar to those reported today for large and outbred Eastern European wolf populations that never faced major population declines (Pilot et al. 2014). This result supports the hypothesis of a recent and human-induced loss of diversity in Iberian wolves, rather than an ancient demographic cause. This result is particularly relevant, given that the Iberian wolf population is isolated from other European populations ca. 10.3 ka with negligible levels of gene flow ever since (Silva et al. 2020). This recent loss of diversity in Iberian wolves is in line with previous evidence in other Central European wolf populations, where the most severe loss in mitochondrial variability occurred in the last century due to strong human persecution (Dufresnes et al. 2018). Declines in genetic diversity following a population bottleneck are well described in several other species of mammals, birds and fishes (Lande 1993; Keller and Waller 2002; Leonard 2008; Charlesworth and Willis 2009).

The distribution of genetic diversity along the genome of historical and contemporary individuals confirmed an overall loss of heterozygosity in the latter. However, we found evidence of active maintenance of variation in the contemporary population at gene families associated with adaptive immune system. MHC-related genes are known for

retaining high variation in many species, including wolves (Niskanen et al. 2014; Rocha et al. 2019; Kloch et al. 2021), through a mechanism of balancing selection. Despite the decrease in overall diversity in the contemporary population, immune alleles were maintained in a high polymorphic state, indicating that selection operated to favor pathogen resistance. Conserving genetic variation in genes involved in immune responses might have facilitated the persistence of Iberian wolves in human-dominated landscapes by increasing their resistance in novel environments. Our empirical findings add to the evidence for a non-random loss of variation across the genome (Aguilar et al. 2004; Wood et al. 2016), reinforcing similar trends previously observed in small populations of other large carnivores (Benazzo et al. 2017; Kleinman-Ruiz et al. 2022).

5.5. Conclusions

Monitoring hybridization and genetic diversity before and following a population bottleneck is crucial to understanding how human-mediated disturbance can affect the genomic composition of a species (Frei et al. 2022). In this study, we performed one of the most comprehensive population genomic assessments over multiple timescales of a wolf population that faced massive declines in the past decades, to understand how levels of genetic diversity and wolf-dog hybridization were affected by the bottleneck. Our findings provide evidence that wolf-dog hybridization in the Iberia Peninsula did not increase in response to the population bottleneck in the 20th-century, and that dog introgression was not a driver of wolf extinction. We show that despite hybridization, Iberian wolves and dogs are two well-defined genetic entities, confirming that levels of gene flow are limited. However, in the limits of the species distribution range, hybridization seems to be more frequent and its impact on the conservation of the species should be further investigated. Additionally, our results clearly revealed a loss of genetic diversity in the contemporary population, suggestive of a rapid and human-induced decline. Still, a process of active maintenance of genetic diversity in genes associated with immune response indicates that the loss of diversity in the contemporary population did not occur randomly. Although the Iberian wolf population seems to be recovering and recolonizing areas in Spain, we encourage the maintenance of conservation measures in the Iberian Peninsula to ensure that the genetic legacy of wolves is not compromised further. The findings of this study can also contribute to predicting evolutionary responses of other large carnivores that have, or still are, facing extreme declines in response to human-induced environmental changes.

5.6. References

Aguilar A, Roemer G, Debenham S, Binns M, Garcelon D, Wayne RK. 2004. High MHC diversity maintained by balancing selection in an otherwise genetically monomorphic mammal. Proc. Natl. Acad. Sci. 101:3490–3494.

Alexander DH, Lange K. 2011. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics 12:246.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19:1655–1664.

Allendorf FW, Leary RF, Spruell P, Wenburg JK. 2001. The problems with hybrids: setting conservation guidelines. Trends Ecol. Evol. 16:613–622.

Álvares F, Barroso I, Blanco JC, Correia J, Cortés Y, Costa G, Llaneza L, Moreira L, Nascimento J, Palacios V, et al. 2005. Wolf status and conservation in the Iberian Peninsula. In: Frontiers of Wolf Recovery: Southwestern US and the World. EUA: Colorado Springs. p. 76–77.

Andersone Ž, Lucchini V, Ozoliņš J. 2002. Hybridisation between wolves and dogs in Latvia as documented using mitochondrial and microsatellite DNA markers. Mamm. Biol. 67:79–90.

Benazzo A, Trucchi E, Cahill JA, Maisano Delser P, Mona S, Fumagalli M, Bunnefeld L, Cornetti L, Ghirotto S, Girardi M, et al. 2017. Survival and divergence in a small group: The extraordinary genomic history of the endangered Apennine brown bear stragglers. Proc. Natl. Acad. Sci. 114:E9589–E9597.

Blanco JC, Cortés Y. 2007. Dispersal patterns, social structure and mortality of wolves living in agricultural habitats in Spain. J. Zool. 273:114–124.

Blanco JC, Cortés Y. 2009. Ecological and social constraints of wolf recovery in Spain. In: Musiani M, Boitani L, Paquet P, editors. A New Era for Wolves and People: Wolf Recovery, Human Attitudes, and Policy. Calgary: University of Calgary Press. p. 41–66.

Blanco JC, de la Cuesta L, Reig S. 1990. El lobo en España: una visión global. In: Blanco JC, de la Cuesta L, Reig S, editors. El lobo (Canis lupus) en España. Situación, problemática y apuntes sobre su ecología. Madrid: ICONA. p. 69–94.

Blanco JC, Reig S, de la Cuesta L. 1992. Distribution, status and conservation problems of the wolf Canis lupus in Spain. Biol. Conserv. 60:73–80.

Blomqvist D, Pauliny A, Larsson M, Flodin L-Å. 2010. Trapped in the extinction vortex? Strong genetic effects in a declining vertebrate population. BMC Evol. Biol. 10:33.

Boitani L. 2003. Wolf conservatio and recovery. In: Mech LD, Boitani L, editors. Wolves: Behaviour, Ecology and Conservation. Chicago: University of Chicago Press. p. 317–340.

Bürger R, Lynch M. 1995. Evolution and extinction in a changing environment: a quantitative-genetic analysis. Evolution 49:151–163.

Campbell CL, Bhérer C, Morrow BE, Boyko AR, Auton A. 2016. A Pedigree-Based Map of Recombination in the Domestic Dog Genome. G3 Genes|Genomes|Genetics 6:3517–3524.

Carling MD, Brumfield RT. 2008. Haldane's rule in an avian system: Using cline theory and

divergence population genetics to test for differential introgression of mitochondrial, autosomal, and sex-linked loci across the Passerina bunting hybrid zone. Evolution 62:2600–2615.

Ceballos FC, Hazelhurst S, Ramsay M. 2018. Assessing runs of Homozygosity: a comparison of SNP Array and whole genome sequence low coverage data. BMC Genomics 19:106.

Chapron G, Kaczensky P, Linnell JDC, von Arx M, Huber D, Andrén H, López-Bao JV, Adamec M, Álvares F, Anders O, et al. 2014. Recovery of large carnivores in Europe's modern human-dominated landscapes. Science 346:1517–1519.

Charlesworth D, Willis JH. 2009. The genetics of inbreeding depression. Nat. Rev. Genet. 10:783–796.

Clavero M, García-Reyes A, Fernández-Gil A, Revilla E, Fernández N. 2022. Where wolves were: setting historical baselines for wolf recovery in Spain. Anim. Conserv.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

Donfrancesco V, Ciucci P, Salvatori V, Benson D, Andersen LW, Bassi E, Blanco JC, Boitani L, Caniglia R, Canu A, et al. 2019. Unravelling the Scientific Debate on How to Address Wolf-Dog Hybridization in Europe. Front. Ecol. Evol. 7.

Dufresnes C, Miquel C, Remollino N, Biollaz F, Salamin N, Taberlet P, Fumagalli L. 2018. Howling from the past: historical phylogeography and diversity losses in European grey wolves. Proc. R. Soc. B Biol. Sci. 285:20181148.

Fan Z, Silva P, Gronau I, Wang S, Armero AS, Schweizer RM, Ramirez O, Pollinger J, Galaverni M, Del-Vecchyo DO, et al. 2016. Worldwide patterns of genomic variation and admixture in gray wolves. Genome Res. 26:163–173.

Fan Z, Silva P, Gronau I, Wang S, Armero AS, Schweizer RM, Ramirez O, Pollinger J, Galaverni M, Ortega Del-Vecchyo D, et al. 2016. Worldwide patterns of genomic variation and admixture in gray wolves. Genome Res. 26:163–173.

Frantz LAF, Schraiber JG, Madsen O, Megens H-J, Cagan A, Bosse M, Paudel Y, Crooijmans RPMA, Larson G, Groenen MAM. 2015. Evidence of long-term gene flow and selection during domestication from analyses of Eurasian wild and domestic pig genomes. Nat. Genet. 47:1141–1148.

Frei D, De-Kayne R, Selz OM, Seehausen O, Feulner PGD. 2022. Genomic variation from an extinct species is retained in the extant radiation following speciation reversal. Nat. Ecol. Evol. 6:461–468.

Galaverni M, Caniglia R, Pagani L, Fabbri E, Boattini A, Randi E. 2017. Disentangling timing of admixture, patterns of introgression, and phenotypic indicators in a hybridizing Wolf population. Mol. Biol. Evol. 34:2324–2339.

Garzón J. 1979. La apasionada geografia del lobo. Trofeo 104:26–28.

Geraldes A, Ferrand N, Nachman MW. 2006. Contrasting patterns of introgression at X-linked loci across the hybrid zone between subspecies of the European rabbit (Oryctolagus cuniculus). Genetics 173:919–933.

Geza E, Mugo J, Mulder NJ, Wonkam A, Chimusa ER, Mazandu GK. 2019. A comprehensive survey of models for dissecting local ancestry deconvolution in human genome. Brief. Bioinform. 20:1709–1724.

Godinho R, Llaneza L, Blanco JC, Lopes S, Álvares F, García EJ, Palacios V, Cortés Y, Javier T, Ferrand N. 2011. Genetic evidence for multiple events of hybridization between wolves and domestic dogs in the Iberian Peninsula. Mol. Ecol. 20:5154–5166.

Godinho R, López-Bao JV, Castro D, Llaneza L, Lopes S, Silva P, Ferrand N. 2015. Real-time assessment of hybridization between wolves and dogs: Combining noninvasive samples with ancestry informative markers. Mol. Ecol. Resour. 15:317–328.

Gómez-Sánchez D, Olalde I, Sastre N, Enseñat C, Carrasco R, Marques-Bonet T, Lalueza-Fox C, Leonard JA, Vilà C, Ramírez O. 2018. On the path to extinction: Inbreeding and admixture in a declining grey wolf population. Mol. Ecol. 27:3599–3612.

Good JM, Giger T, Dean MD, Nachman MW. 2010. Widespread over-expression of the X chromosome in sterile F-1 hybrid mice. PLoS Genet. 6:e1001148.

Hindrikson M, Männil P, Ozolins J, Krzywinski A, Saarma U. 2012. Bucking the Trend in Wolf-Dog Hybridization: First Evidence from Europe of Hybridization between Female Dogs and Male Wolves. PLoS One 7:e46465.

Hindrikson M, Remm J, Pilot M, Godinho R, Stronen AV, Baltrūnaité L, Czarnomska SD, Leonard JA, Randi E, Nowak C, et al. 2017. Wolf population genetics in Europe: a systematic review, meta-analysis and suggestions for conservation and management. Biol. Rev. 92:1601–1629.

Keller LF, Waller DM. 2002. Inbreeding effects in wild populations. Trends Ecol. Evol.17:230–241.

Kleinman-Ruiz D, Lucena-Perez M, Villanueva B, Fernández J, Saveljev AP, Ratkiewicz M, Schmidt K, Galtier N, García-Dorado A, Godoy JA. 2022. Purging of deleterious burden in the endangered Iberian lynx. Proc. Natl. Acad. Sci. 119:e2110614119.

Kloch A, Biedrzycka A, Szewczyk M, Nowak S, Niedźwiedzka N, Kłodawska M, Hájková A, Hulva P, Jędrzejewska B, Mysłajek R. 2021. High genetic diversity of immunity genes in an expanding population of a highly mobile carnivore, the grey wolf Canis lupus , in Central Europe. Divers. Distrib. 27:1680–1695.

Kusak J, Fabbri E, Galov A, Gomerčić T, Arbanasić H, Caniglia R, Galaverni M, Reljić S, Huber D, Randi E. 2018. Wolf-dog hybridization in Croatia. Vet. Arh. 88:375–395.

Lande R. 1993. Risks of Population Extinction from Demographic and Environmental Stochasticity and Random Catastrophes. Am. Nat. 142:911–927.

Lande R. 1994. Risk of population extinction from fixation of new deleterious mutations. Evolution 48:1460–1469.

Leonard JA. 2008. Ancient DNA applications for wildlife conservation. Mol. Ecol.17:4186–4196.

Leonard JA, Echegaray J, Rand E, Vilà C. 2013. Impact of hybridization with domestic dogs on the conservation of wild canids. In: Gompper M, editor. Free-Ranging Dogs and Wildlife Conservation. Oxford University Press. p. 170–184.

Lischer HEL, Excoffier L. 2012. PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. Bioinformatics 28:298–299.

López-Bao JV, Blanco JC, Rodríguez A, Godinho R, Sazatornil V, Alvares F, García EJ, Llaneza L, Rico M, Cortés Y, et al. 2015. Toothless wildlife protection laws. Biodivers. Conserv. 24:2105–2108.

López-Bao JV, Fleurke F, Chapron G, Trouwborst A. 2018. Legal obligations regarding populations on the verge of extinction in Europe: Conservation, Restoration, Recolonization, Reintroduction. Biol. Conserv. 227:319–325.

Mech LD, Barber-Meyer SM, Erb J. 2016. Wolf (Canis lupus) Generation Time and Proportion of Current Breeding Females by Age. PLoS One 11:e0156682.

Meyermans R, Gorssen W, Buys N, Janssens S. 2020. How to study runs of homozygosity using PLINK? A guide for analyzing medium density SNP data in livestock and pet species. BMC Genomics 21.

Nei M, Maruyama T, Chakraborty R. 1975. The Bottleneck Effect and Genetic Variability in Populations. Evolution 29:1–10.

Nielsen E, Bach L, Kotlicki P. 2006. hybridlab (version 1.0): a program for generating simulated hybrids from population samples. Mol. Ecol. Notes 6:971–973.

Niskanen AK, Kennedy LJ, Ruokonen M, Kojola I, Lohi H, Isomursu M, Jansson E, Pyhäjärvi T, Aspi J. 2014. Balancing selection and heterozygote advantage in major histocompatibility complex loci of the bottlenecked Finnish wolf population. Mol. Ecol. 23:875–889.

Nores C, López-Bao JV. 2022. Historical data to inform the legal status of species in Europe: An example with wolves. Biol. Conserv. 272:109639.

Pacheco C, Lobo D, Silva P, Álvares F, García EJ, Castro D, Layna JF, López-Bao JV, Godinho R. 2022. Assessing the performance of historical skins and bones for museomics using wolf specimens as a case study. Front. Ecol. Evol. 10:970249.

Pacheco C, López-Bao JV, García EJ, Lema FJ, Llaneza L, Palacios V, Godinho R. 2017. Spatial assessment of wolf-dog hybridization in a single breeding period. Sci. Rep. 7:42475.

Pasaniuc B, Sankararaman S, Kimmel G, Halperin E. 2009. Inference of locus-specific ancestry in closely related populations. Bioinformatics 25:i213–i221.

Pendleton AL, Shen F, Taravella AM, Emery S, Veeramah KR, Boyko AR, Kidd JM. 2018. Comparison of village dog and wolf genomes highlights the role of the neural crest in dog domestication. BMC Biol.16.

Pickrell JK, Pritchard JK. 2012. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. PLoS Genet. 8:e1002967.

Pilot M, Greco C, VonHoldt BM, Jędrzejewska B, Randi E, Jędrzejewski W, Sidorovich VE, Ostrander EA, Wayne RK. 2014. Genome-wide signatures of population bottlenecks and diversifying selection in European wolves. Heredity 112:428–442.

Pilot M, Greco C, VonHoldt BM, Randi E, Jędrzejewski W, Sidorovich VE, Konopiński MK, Ostrander EA, Wayne RK. 2018. Widespread, long-term admixture between grey wolves and domestic dogs across Eurasia and its implications for the conservation status of hybrids. Evol. Appl. 11:662–680.

Pilot M, Moura AE, Okhlopkov IM, Mamaev NV., Manaseryan NH, Hayrapetyan V, Kopaliani N, Tsingarska E, Alagaili AN, Mohammed OB, et al. 2021. Human-modified canids in human-modified landscapes: The evolutionary consequences of hybridization for grey wolves and free-ranging domestic dogs. Evol. Appl. 14:2433–2456.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. 81:559–575.

R Development Core Team. 2017. R: A language and environment for statistical computing. Vienna, Austria.

Randi E. 2011. Genetics and conservation of wolves Canis lupus in Europe. Mamm. Rev. 41:99–111.

Randi E, Hulva P, Fabbri E, Galaverni M, Galov A, Kusak J, Bigi D, Bolfíková BČ, Smetanová M, Caniglia R. 2014. Multilocus detection of wolf x dog hybridization in Italy, and guidelines for marker selection. PLoS One 9:e86409.

Randi E, Lucchini V. 2002. Detecting rare introgression of domestic dog genes into wild wolf (Canis lupus) populations by Bayesian admixture analyses of microsatellite variation. Conserv. Genet. 3:31–45.

Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, Vilo J. 2019. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 47:W191–W198.

Ripple WJ, Estes JA, Beschta RL, Wilmers CC, Ritchie EG, Hebblewhite M, Berger J, Elmhagen B, Letnic M, Nelson MP, et al. 2014. Status and Ecological Effects of the World's Largest Carnivores. Science 343.

Rocha RG, Magalhães V, López-Bao JV, van der Loo W, Llaneza L, Alvares F, Esteves PJ, Godinho R. 2019. Alternated selection mechanisms maintain adaptive diversity in different demographic scenarios of a large carnivore. BMC Evol. Biol. 19:90.

Runemark A, Trier CN, Eroukhmanoff F, Hermansen JS, Matschiner M, Ravinet M, Elgvin TO, Sætre G-P. 2018. Variation and constraints in hybrid genome formation. Nat. Ecol. Evol. 2:549–556.

Salvatori V, Donfrancesco V, Trouwborst A, Boitani L, Linnell JDC, Alvares F, Åkesson M, Balys V, Blanco JC, Chiriac S, et al. 2020. European agreements for nature conservation need to explicitly address wolf-dog hybridisation. Biol. Conserv. 248:108525.

Sankararaman S, Sridhar S, Kimmel G, Halperin E. 2008. Estimating Local Ancestry in Admixed Populations. Am. J. Hum. Genet. 82:290–303.

Sastre N, Vilà C, Salinas M, Bologov V V., Urios V, Sánchez A, Francino O, Ramírez O. 2011. Signatures of demographic bottlenecks in European wolf populations. Conserv. Genet. 12:701–712.

Schumer M, Xu C, Powell DL, Durvasula A, Skov L, Holland C, Blazier JC, Sankararaman S, Andolfatto P, Rosenthal GG, et al. 2018. Natural selection interacts with recombination to shape the evolution of hybrid genomes. Science 360:656–660.

Shama LN, Kubow KB, Jokela J, Robinson CT. 2011. Bottlenecks drive temporal and spatial genetic changes in alpine caddisfly metapopulations. BMC Evol. Biol. 11:278.

Silva P, Galaverni M, Ortega-Del Vecchyo D, Fan Z, Caniglia R, Fabbri E, Randi E, Wayne R, Godinho R. 2020. Genomic evidence for the Old divergence of Southern European wolf populations. Proc. R. Soc. B Biol. Sci. 287:20201206.

Smeds L, Aspi J, Berglund J, Kojola I, Tirronen K, Ellegren H. 2021. Whole-genome analyses provide no evidence for dog introgression in Fennoscandian wolf populations. Evol. Appl. 14:721–734.

Torres RT, Ferreira E, Rocha RG, Fonseca C. 2017. Hybridization between wolf and domestic dog: First evidence from an endangered population in central Portugal. Mamm. Biol.

86:70–74.

Turner LM, Harr B. 2014. Genome-wide mapping in a house mouse hybrid zone reveals hybrid sterility loci and Dobzhansky-Muller interactions. Elife 3:1–25.

Valverde J. 1971. El lobo español. Montes 159:229–241.

Verardi A, Lucchini V, Randi E. 2006. Detecting introgressive hybridization between free-ranging domestic dogs and wild wolves (Canis lupus) by admixture linkage disequilibrium analysis. Mol. Ecol.15:2845–2855.

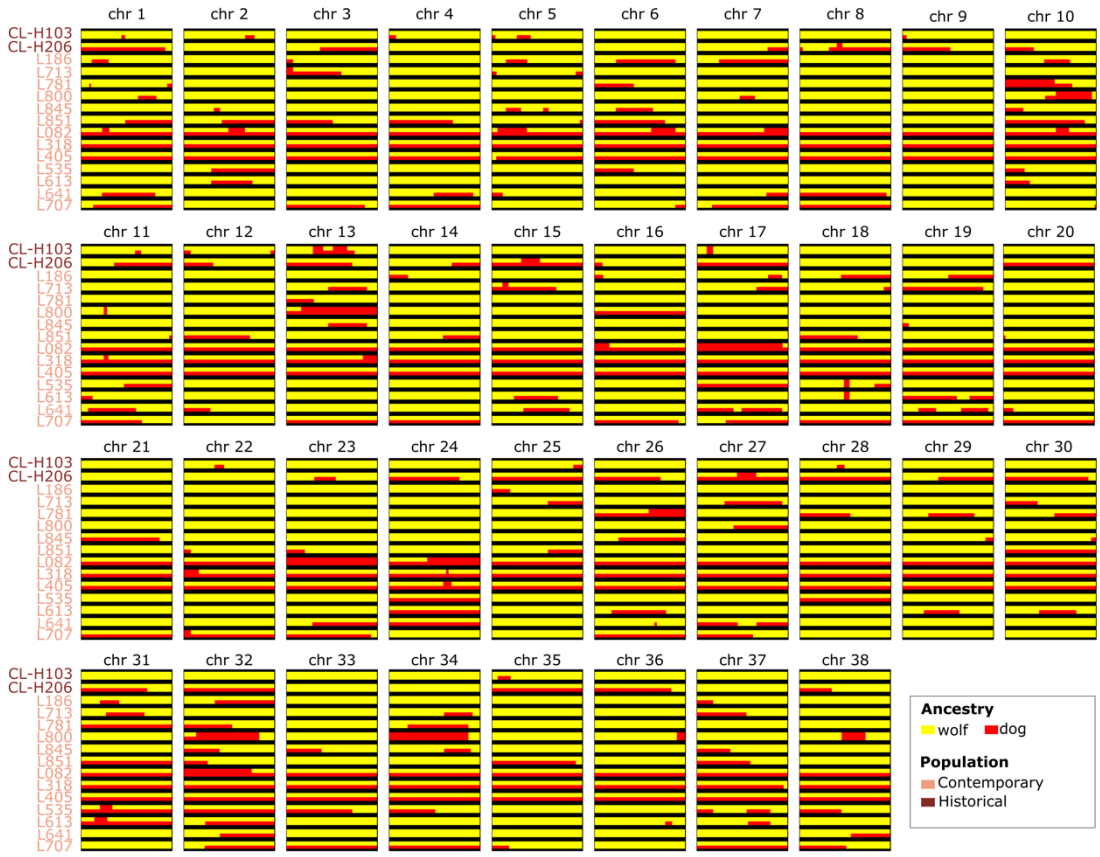
Vilà C, Walker C, Sundqvist A-K, Flagstad Ø, Andersone Z, Casulli A, Kojola I, Valdmann H, Halverson J, Ellegren H. 2003. Combined use of maternal, paternal and bi-parental genetic markers for the identification of wolf–dog hybrids. Heredity 90:17–24.

Weir BS, Cockerham CC. 1984. Estimating F-Statistics for the Analysis of Population Structure. Evolution 38:1358.

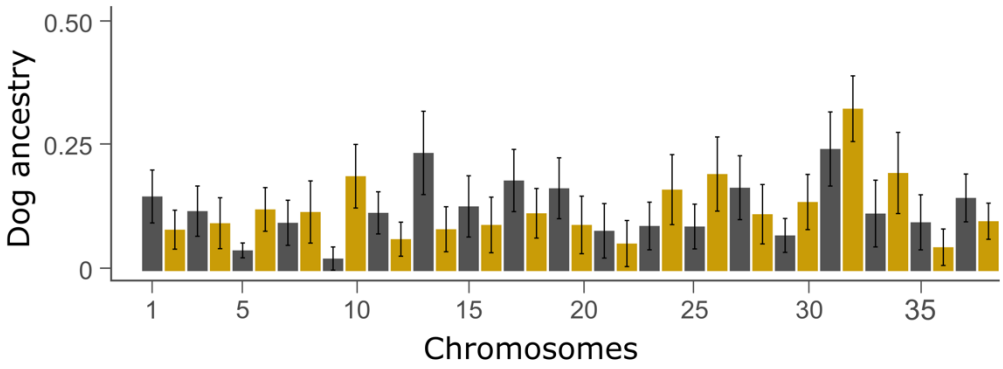
Wong AK, Ruhe AL, Dumont BL, Robertson KR, Guerrero G, Shull SM, Ziegle JS, Millon L V., Broman KW, Payseur BA, et al. 2010. A Comprehensive Linkage Map of the Dog Genome. Genetics 184:595–605.

Wood JLA, Yates MC, Fraser DJ. 2016. Are heritability and selection related to population size in nature? Meta-analysis and conservation implications. Evol. Appl. 9:640–657.

5.7. Supplementary Materials



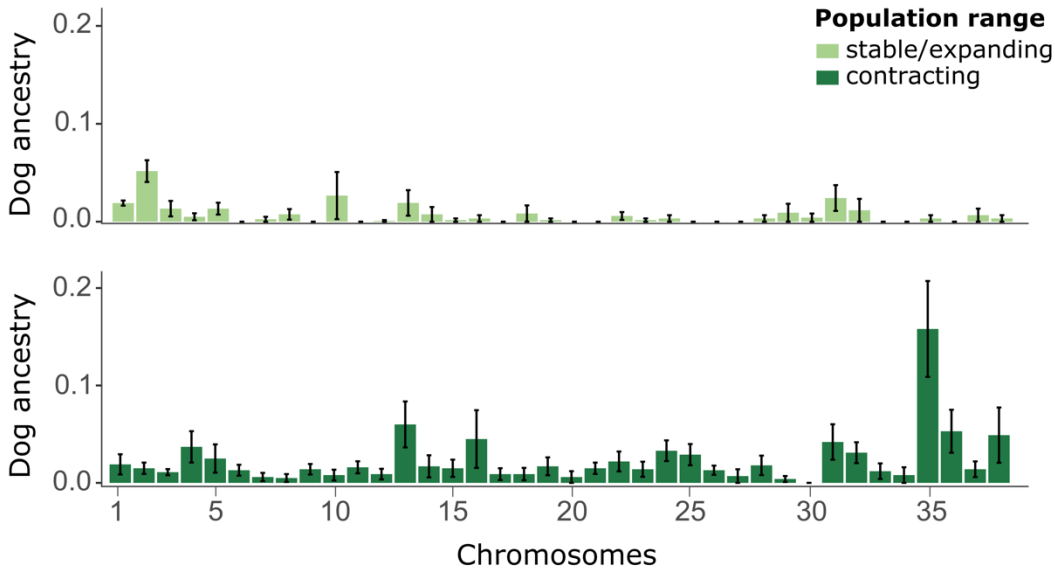
Supplementary Figure 5.1. Local ancestry analysis in hybrids. Results of local ancestry analysis across the 15 wolf-dog hybrids identified in the historical (red) and contemporary (orange) populations of Iberian wolf, using LAMP-ANC on the 20K SNP dataset. Blocks were colored based on the attributed local ancestry: yellow for wolf (both chromosomes), red for dog (both chromosomes), and half yellow/red for wolf/dog (one chromosome of each). Each box represents a chromosome, and every row identifies an individual (sample IDs provided).



Supplementary Figure 5.2. Dog ancestry estimates per chromosome in wolf-dog hybrid individuals. Average proportion of dog ancestry per chromosome measured across backcross hybrid individuals (N=12), determined by LAMP-ANC using the 20K dataset. Each bar displays the average value with the corresponding standard error (SE) interval.



Supplementary Figure 5.3. Local ancestry analysis in historical wolf samples. Results of local ancestry analysis across 22 historical samples from the contracting (N = 10; dark green) and stable/expanding (N = 12; light green) population ranges, using LAMP-ANC on the 60K SNP dataset. Block were colored based on the attributed local ancestry: yellow for wolf (both chromosomes), red for dog (both chromosomes), and half yellow/red for wolf/dog (one chromosome of each). Each box represents a chromosome, and every row identifies an individual (sample IDs provided).



Supplementary Figure 5.4. Dog ancestry estimates per chromosome in historical "pure" wolf samples. Average proportion of dog ancestry per chromosome measured across historical "pure" wolf samples from the contracting (N = 10; dark green) and stable/expanding (N = 12; light green) population ranges. Dog ancestry estimates were determined by LAMP-ANC using the 60K SNP dataset. Each bar displays the average value with the corresponding standard error (SE) interval.

Supplementary Table 5.1. Description of the full dataset used in this study. Information about sample ID, population, collection date (year of the original collection for historical samples), locality within the Iberian Peninsula (for wolves only), dataset where each sample was included, and the study that generate each sample, are provided.

Sample ID	Population/Dog Breed	Period (date)	Locality	Datasets	Reference
CGT33	Dog (Cão de Gado Transmontano)	C	PT	5K, 20K, 60K	[1]
CGT34	Dog (Cão de Gado Transmontano)	C	PT	5K, 20K, 60K	[1]
CGT36	Dog (Cão de Gado Transmontano)	C	PT	5K, 20K, 60K	[1]
CGT38	Dog (Cão de Gado Transmontano)	C	PT	5K, 20K, 60K	[1]
CGT39	Dog (Cão de Gado Transmontano)	C	PT	5K, 20K, 60K	[1]
CCL22	Dog (Castro Laboreiro)	C	PT	5K, 20K, 60K	[1]
CCL32	Dog (Castro Laboreiro)	C	PT	5K, 20K, 60K	[1]
CL04	Dog (Castro Laboreiro)	C	PT	5K, 20K, 60K	[1]
CL5	Dog (Castro Laboreiro)	C	PT	5K, 20K, 60K	[1]
CL11	Dog (Castro Laboreiro)	C	PT	5K, 20K, 60K	[1]
RA01	Dog (Rafeiro do Alentejo)	C	PT	5K, 20K, 60K	[1]
RA05	Dog (Rafeiro do Alentejo)	C	PT	5K, 20K, 60K	[1]
RA16	Dog (Rafeiro do Alentejo)	C	PT	5K, 20K, 60K	[1]
RA22	Dog (Rafeiro do Alentejo)	C	PT	5K, 20K, 60K	[1]
RA23	Dog (Rafeiro do Alentejo)	C	PT	5K, 20K, 60K	[1]
SE_EMD004	Dog (Serra da Estrela)	C	PT	5K, 20K, 60K	[1]
SE_EMD009	Dog (Serra da Estrela)	C	PT	5K, 20K, 60K	[1]
SE_EMD016	Dog (Serra da Estrela)	C	PT	5K, 20K, 60K	[1]
SE_EMD021	Dog (Serra da Estrela)	C	PT	5K, 20K, 60K	[1]
SE_EMD030	Dog (Serra da Estrela)	C	PT	5K, 20K, 60K	[1]
CR2	Dog (Village dog)	C	PT	5K, 20K, 60K	[1]
CR9	Dog (Village dog)	C	PT	5K, 20K, 60K	[1]
CR11	Dog (Village dog)	C	PT	5K, 20K, 60K	[1]
CR14	Dog (Village dog)	C	PT	5K, 20K, 60K	[1]
CR35	Dog (Village dog)	C	Spain	5K, 20K, 60K	[1]
CR39	Dog (Village dog)	C	Spain	5K, 20K, 60K	[1]
CR44	Dog (Village dog)	C	PT	5K, 20K, 60K	[1]
CR63	Dog (Village dog)	C	PT	5K, 20K, 60K	[1]
CR64	Dog (Village dog)	C	PT	5K, 20K, 60K	[1]
CR65	Dog (Village dog)	C	PT	5K, 20K, 60K	[1]
CR67	Dog (Village dog)	C	PT	5K, 20K, 60K	[1]
CR68	Dog (Village dog)	C	PT	5K, 20K, 60K	[1]
CR77	Dog (Village dog)	C	PT	5K, 20K, 60K	[1]
CR80	Dog (Village dog)	C	PT	5K, 20K, 60K	[1]
CR97	Dog (Village dog)	C	PT	5K, 20K, 60K	[1]
CR98	Dog (Village dog)	C	PT	5K, 20K, 60K	[1]
CR104	Dog (Village dog)	C	PT	5K, 20K, 60K	[1]
L400	Dog (Village dog)	C	PT	5K, 20K, 60K	[1]
L501	Dog (Village dog)	C	PT	5K, 20K, 60K	[1]
L624	Dog (Village dog)	C	Spain	5K, 20K, 60K	[1]
L652	Dog (Village dog)	C	Spain	5K, 20K, 60K	[1]
L744	Dog (Village dog)	C	Spain	5K, 20K, 60K	[1]
L791	Dog (Village dog)	C	Spain	5K, 20K, 60K	[1]
L005	Iberian Wolf	C	Bragança, PT	5K, 20K, 60K	[1]
L010	Iberian Wolf	C	Bragança, PT	5K, 20K, 60K	[1]
L029	Iberian Wolf	C	Vila Real, PT	5K, 20K, 60K	[1]
L063	Iberian Wolf	C	Asturias, Spain	5K, 20K, 60K	[1]
L082	Iberian Wolf	C	Asturias, Spain	5K, 20K, 60K	[1]
L086	Iberian Wolf	C	Viseu, PT	5K, 20K, 60K	[1]
L102	Iberian Wolf	C	Bragança, PT	5K, 20K, 60K	[1]
L116	Iberian Wolf	C	Galicia, Spain	5K, 20K, 60K	[1]
L125	Iberian Wolf	C	Viana do Castelo, PT	5K, 20K, 60K	[1]
L180	Iberian Wolf	C	Asturias, Spain	5K, 20K, 60K	[1]
L186	Iberian Wolf	C	Asturias, Spain	5K, 20K, 60K	[1]
L188	Iberian Wolf	C	Asturias, Spain	5K, 20K, 60K	[1]
L190	Iberian Wolf	C	Asturias, Spain	5K, 20K, 60K	[1]
L212	Iberian Wolf	C	Asturias, Spain	5K, 20K, 60K	[1]
L253	Iberian Wolf	C	Lugo, Galicia, Spain	5K, 20K, 60K	[1]
L262	Iberian Wolf	C	Pontevedra, Galicia, Spain	5K, 20K, 60K	[1]
L316	Iberian Wolf	C	Viana do Castelo, PT	5K, 20K, 60K	[1]
L318	Iberian Wolf	C	Salamanca, C-L, Spain	5K, 20K, 60K	[1]
L319	Iberian Wolf	C	Zamora, C-L Spain	5K, 20K, 60K	[1]
L320	Iberian Wolf	C	Ávila, C-L, Spain	5K, 20K, 60K	[1]

L337	Iberian Wolf	C	Viana do Castelo, PT	5K, 20K, 60K	[1]
L338	Iberian Wolf	C	Lugo, Galicia, Spain	5K, 20K, 60K	[1]
L351	Iberian Wolf	C	Asturias, Spain	5K, 20K, 60K	[1]
L357	Iberian Wolf	C	Cantabria, Spain	5K, 20K, 60K	[1]
L360	Iberian Wolf	C	Coruña, Galicia, Spain	5K, 20K, 60K	[1]
L364	Iberian Wolf	C	Pontevedra, Galicia, Spain	5K, 20K, 60K	[1]
L382	Iberian Wolf	C	Asturias, Spain	5K, 20K, 60K	[1]
L387	Iberian Wolf	C	Asturias, Spain	5K, 20K, 60K	[1]
L392	Iberian Wolf	C	Viana do Castelo, PT	5K, 20K, 60K	[1]
L401	Iberian Wolf	C	Lugo, Galicia, Spain	5K, 20K, 60K	[1]
L405	Iberian Wolf	C	Viana do Castelo, PT	5K, 20K, 60K	[1]
L422	Iberian Wolf	C	Viseu, PT	5K, 20K, 60K	[1]
L431	Iberian Wolf	C	Asturias, Spain	5K, 20K, 60K	[1]
L446	Iberian Wolf	C	Asturias, Spain	5K, 20K, 60K	[1]
L474	Iberian Wolf	C	Bragança, PT	5K, 20K, 60K	[2]
L502	Iberian Wolf	C	Coruña, Galicia, Spain	5K, 20K, 60K	[1]
L509	Iberian Wolf	C	Pontevedra, Galicia, Spain	5K, 20K, 60K	[1]
L514	Iberian Wolf	C	Lugo, Galicia, Spain	5K, 20K, 60K	[1]
L535	Iberian Wolf	C	Coruña, Galicia, Spain	5K, 20K, 60K	[1]
L546	Iberian Wolf	C	Lugo, Galicia, Spain	5K, 20K, 60K	[1]
L547	Iberian Wolf	C	Pontevedra, Galicia, Spain	5K, 20K, 60K	[1]
L577	Iberian Wolf	C	León, C-L, Spain	5K, 20K, 60K	[1]
L580	Iberian Wolf	C	Palencia, C-L, Spain	5K, 20K, 60K	[1]
L585	Iberian Wolf	C	Valladolid, C-L, Spain	5K, 20K, 60K	[2]
L596	Iberian Wolf	C	Vila Real, PT	5K, 20K, 60K	[1]
L613	Iberian Wolf	C	Coruña, Galicia, Spain	5K, 20K, 60K	[1]
L625	Iberian Wolf	C	Coruña, Galicia, Spain	5K, 20K, 60K	[1]
L641	Iberian Wolf	C	Asturias, Spain	5K, 20K, 60K	[1]
L649	Iberian Wolf	C	Asturias, Spain	5K, 20K, 60K	[1]
L707	Iberian Wolf	C	Burgos, C-L, Spain	5K, 20K, 60K	[1]
L713	Iberian Wolf	C	Burgos, C-L, Spain	5K, 20K, 60K	[1]
L768	Iberian Wolf	C	Coruña, Galicia, Spain	5K, 20K, 60K	[1]
L772	Iberian Wolf	C	Viseu, PT	5K, 20K, 60K	[1]
L780	Iberian Wolf	C	Viana do Castelo, PT	5K, 20K, 60K	[1]
L781	Iberian Wolf	C	Coruña, Galicia, Spain	5K, 20K, 60K	[1]
L783	Iberian Wolf	C	Lugo, Galicia, Spain	5K, 20K, 60K	[1]
L788	Iberian Wolf	C	Orense, Galicia, Spain	5K, 20K, 60K	[1]
L790	Iberian Wolf	C	Lugo, Galicia, Spain	5K, 20K, 60K	[1]
L795	Iberian Wolf	C	Viana do Castelo, PT	5K, 20K, 60K	[1]
L798	Iberian Wolf	C	Coruña, Galicia, Spain	5K, 20K, 60K	[1]
L800	Iberian Wolf	C	Coruña, Galicia, Spain	5K, 20K, 60K	[1]
L801	Iberian Wolf	C	Coruña, Galicia, Spain	5K, 20K, 60K	[1]
L805	Iberian Wolf	C	Lugo, Galicia, Spain	5K, 20K, 60K	[1]
L806	Iberian Wolf	C	Asturias, Spain	5K, 20K, 60K	[1]
L819	Iberian Wolf	C	León, C-L, Spain	5K, 20K, 60K	[1]
L825	Iberian Wolf	C	Asturias, Spain	5K, 20K, 60K	[1]
L845	Iberian Wolf	C	Guadalajara, C-La Mancha, Spain	5K, 20K, 60K	[1]
L851	Iberian Wolf	C	Guadalajara, C-La Mancha, Spain	5K, 20K, 60K	[1]
CL-H010	Iberian Wolf	H (1975)	Segovia, C-L, Spain	5K, 20K	[2]
CL-H017	Iberian Wolf	H (1972)	León, C-L, Spain	5K, 20K, 60K	[2]
CL-H022	Iberian Wolf	H (1976)	León, C-L, Spain	5K, 20K, 60K	[2]
CL-H027	Iberian Wolf	H (1983)	Burgos, C-L, Spain	5K, 20K	[2]
CL-H029	Iberian Wolf	H (1983)	Valladolid, C-L, Spain	5K, 20K	[2]
CL-H032	Iberian Wolf	H (1983)	León, C-L, Spain	5K	[2]
CL-H034	Iberian Wolf	H (1984)	Zamora, C-L, Spain	5K, 20K	[2]
CL-H035	Iberian Wolf	H (1986)	Zamora, C-L, Spain	5K, 20K, 60K	[2]
CL-H036	Iberian Wolf	H (1984)	Zamora, C-L, Spain	5K, 20K, 60K	[2]
CL-H038	Iberian Wolf	H (1984)	Palencia, C-L, Spain	5K, 20K	[2]
CL-H040	Iberian Wolf	H (1983)	León, C-L, Spain	5K, 20K, 60K	[2]

CL-H042	Iberian Wolf	H (1982)	León, C-L, Spain	5K, 20K	[2]
CL-H052	Iberian Wolf	H (1979)	Guarda, PT	5K, 20K, 60K	[2]
CL-H062	Iberian Wolf	H (1987)	Viseu, PT	5K	[2]
CL-H081	Iberian Wolf	H (1978)	Viseu, PT	5K	[2]
CL-H083	Iberian Wolf	H (1972)	Castelo Branco, PT	5K, 20K, 60K	[2]
CL-H084	Iberian Wolf	H (1972)	Castelo Branco, PT	5K, 20K, 60K	[2]
CL-H086	Iberian Wolf	H (1912)	Castelo Branco, PT	5K, 20K, 60K	[2]
CL-H093	Iberian Wolf	H (1930)	Castelo Branco, PT	5K, 20K	[2]
CL-H100	Iberian Wolf	H (1962)	Beja, PT	5K, 20K, 60K	[2]
CL-H103	Iberian Wolf	H (1959)	Beja, PT	5K, 20K, 60K	[2]
CL-H104	Iberian Wolf	H (1959)	Beja, PT	5K	[2]
CL-H105	Iberian Wolf	H (1959)	Portalegre, PT	5K, 20K	[2]
CL-H116	Iberian Wolf	H (1927)	Jaén, Andalucía, Spain	5K, 20K, 60K	[2]
CL-H117	Iberian Wolf	H (1970)	Ciudad Real, C-La Mancha, SP	5K, 20K, 60K	[2]
CL-H119	Iberian Wolf	H (1944)	Badajoz, Ext, Spain	5K, 20K, 60K	[2]
CL-H126	Iberian Wolf	H (1973)	Setubal, PT	5K, 20K	[2]
CL-H129	Iberian Wolf	H (1991)	Braga, PT	5K, 20K	[2]
CL-H131	Iberian Wolf	H (1987)	Viana do Castelo, PT	5K, 20K, 60K	[2]
CL-H132	Iberian Wolf	H (1955)	Braga, PT	5K, 20K, 60K	[2]
CL-H135	Iberian Wolf	H (1945)	Vila Real, PT	5K, 20K	[2]
CL-H137	Iberian Wolf	NA	Vila Real, PT	5K, 20K, 60K	[2]
CL-H139	Iberian Wolf	H (1980)	Bragança, PT	5K, 20K	[2]
CL-H140	Iberian Wolf	H (1980)	Bragança, PT	5K, 20K	[2]
CL-H141	Iberian Wolf	H (1960)	Bragança, PT	5K	[2]
CL-H142	Iberian Wolf	H (1960)	Bragança, PT	5K, 20K	[2]
CL-H143	Iberian Wolf	H (1970)	Bragança, PT	5K, 20K	[2]
CL-H145	Iberian Wolf	H (1980)	Bragança, PT	5K, 20K	[2]
CL-H146	Iberian Wolf	H (1980)	Bragança, PT	5K, 20K, 60K	[2]
CL-H148	Iberian Wolf	H (1980)	Bragança, PT	5K, 20K	[2]
CL-H149	Iberian Wolf	H (1970)	Braga, PT	5K, 20K, 60K	[2]
CL-H177	Iberian Wolf	H (1984)	Palencia, C-L, Spain	5K, 20K, 60K	[2]
CL-H182	Iberian Wolf	H (1964)	Jaén, Andalucía, Spain	5K, 20K, 60K	[2]
CL-H188	Iberian Wolf	H (1970)	Asturias, Spain	5K, 20K	[2]
CL-H197	Iberian Wolf	H (1969)	León, C-L, Spain	5K, 20K	[2]
CL-H200	Iberian Wolf	H (1974)	Orense, Galicia, Spain	5K, 20K	[2]
CL-H206	Iberian Wolf	H (1972)	Badajoz, Ext, Spain	5K, 20K	[2]
CL-H220	Iberian Wolf	H (1976)	Salamanca, C-L, Spain	5K, 20K, 60K	[2]
CL-H222	Iberian Wolf	H (1971)	Cordoba, Andalucía, Spain	5K	[2]
CL-H234	Iberian Wolf	H (1969)	Pontevedra, Galicia, Spain	5K, 20K	[2]
CL-H240	Iberian Wolf	H (1975)	Palencia, C-L, Spain	5K	[2]
CL-H241	Iberian Wolf	H (1967)	León, C-L, Spain	5K, 20K	[2]
CL-H261	Iberian Wolf	H (1991)	Sevilla, Andalucía, Spain	5K, 20K, 60K	[2]
CL-H268	Iberian Wolf	H (1989)	Zamora, C-L, Spain	5K, 20K	[2]

[1] – Lobo et al. Chapter III; [2] – Pacheco, Lobo et al. 2022; Chapter II.

Abbreviation: C, contemporary; C-L, Castilla y León; H, historical; PT, Portugal; NA, not available

Supplementary Table 5.2. Proportion of dog ancestry in hybrids. Proportion of the global dog ancestry (q) in the genome of the 15 hybrids identified in the historical and contemporary populations, using ADMIXTURE (5K dataset) and LAMP-ANC (20K dataset).

Sample ID	Period	q dog (ADMIXTURE)	q dog (LAMP)
CL-H103	Historical	0.03	0.03
CL-H206	Historical	0.24	0.25
L845	Contemporary	0.05	0.06
L800	Contemporary	0.05	0.12
L781	Contemporary	0.06	0.10
L186	Contemporary	0.06	0.07
L613	Contemporary	0.06	0.09
L713	Contemporary	0.07	0.08
L535	Contemporary	0.09	0.13
L641	Contemporary	0.13	0.11
L851	Contemporary	0.15	0.14
L707	Contemporary	0.28	0.28
L405	Contemporary	0.48	0.50
L318	Contemporary	0.51	0.50
L082	Contemporary	0.58	0.56

Supplementary Table 5.3. Proportion of dog ancestry in "pure" wolf historical samples. Proportion of the global dog ancestry (q) in the genome of historical samples from the contracting and stable/expanding population ranges, using ADMIXTURE (5K dataset) and LAMP-ANC (60K dataset).

Sample ID	Status distribution range	q dog (ADMIXTURE)	q dog (LAMP)
CL-H017	Stable/expanding	0	0.01
CL-H022	Stable/expanding	0	0
CL-H035	Stable/expanding	0	0
CL-H036	Stable/expanding	0	0
CL-H040	Stable/expanding	0	0.01
CL-H052	Contracting	0	0.01
CL-H083	Contracting	0	0.01
CL-H084	Contracting	0	0
CL-H086	Contracting	0.01	0.02
CL-H100	Contracting	0	0.03
CL-H116	Contracting	0	0.04
CL-H117	Contracting	0	0.03
CL-H119	Contracting	0	0.04
CL-H131	Stable/expanding	0	0.01
CL-H132	Stable/expanding	0	0
CL-H146	Stable/expanding	0	0
CL-H149	Stable/expanding	0	0
CL-H177	Stable/expanding	0	0.01
CL-H182	Contracting	0	0.03
CL-H220	Stable/expanding	0	0.03
CL-H261	Contracting	0.01	0.02
CL-H137	Stable/expanding	0	0.01

Supplementary Table 5.4. Differential loss of genetic diversity across geographical regions. Proportion of heterozygous sites (het) present in historical and contemporary wolf samples from the same geographical location. Two examples of different population histories across time: a region where the wolf population is currently threatened and contracting its range (south of Douro River, Central Portugal) and a region where the population is stable/expanding its range (Castilla y León, Northwestern Spain). Average values of heterozygosity for the historical and contemporary populations in each of these two case studies are, respectively: South of Douro (0.195/0.157); Castilla y León (0.193/0.173).

Sample ID	Period (date)	Status current range	Region	Het
CL-H081	H (1978)	Contracting	S-D, PT	0.209
CL-H052	H (1979)	Contracting	S-D, PT	0.196
CL-H062	H (1987)	Contracting	S-D, PT	0.181
L086	C	Contracting	S-D, PT	0.151
L422	C	Contracting	S-D, PT	0.180
L772	C	Contracting	S-D, PT	0.140
CL-H241	H (1967)	Stable/expanding	C-L, Spain	0.209
CL-H197	H (1969)	Stable/expanding	C-L, Spain	0.214
CL-H017	H (1972)	Stable/expanding	C-L, Spain	0.204
CL-H010	H (1975)	Stable/expanding	C-L, Spain	0.174
CL-H240	H (1975)	Stable/expanding	C-L, Spain	0.188
CL-H220	H (1976)	Stable/expanding	C-L, Spain	0.226
CL-H022	H (1976)	Stable/expanding	C-L, Spain	0.182
CL-H042	H (1982)	Stable/expanding	C-L, Spain	0.176
CL-H040	H (1983)	Stable/expanding	C-L, Spain	0.164
CL-H027	H (1983)	Stable/expanding	C-L, Spain	0.171
CL-H029	H (1983)	Stable/expanding	C-L, Spain	0.192
CL-H032	H (1983)	Stable/expanding	C-L, Spain	0.217
CL-H034	H (1984)	Stable/expanding	C-L, Spain	0.214
CL-H036	H (1984)	Stable/expanding	C-L, Spain	0.200
CL-H038	H (1984)	Stable/expanding	C-L, Spain	0.192
CL-H177	H (1984)	Stable/expanding	C-L, Spain	0.159
CL-H035	H (1986)	Stable/expanding	C-L, Spain	0.195
CL-H268	H (1989)	Stable/expanding	C-L, Spain	0.194
L319	C	Stable/expanding	C-L, Spain	0.188
L320	C	Stable/expanding	C-L, Spain	0.125
L580	C	Stable/expanding	C-L, Spain	0.175
L585	C	Stable/expanding	C-L, Spain	0.166
L819	C	Stable/expanding	C-L, Spain	0.209

Abbreviation: C, contemporary; C-L, Castilla y León; H, historical; S-D, South of Douro

Supplementary Table 5.5. ROHs characterization in the historical and contemporary populations. Total amount of homozygosity (in kb) in the genome, average size (kb) of ROH segments, and total number of ROH segments per individual from the historical and contemporary populations. Sample IDs are presented. This analysis was based on the 60K dataset using 22 individuals per population.

Sample ID	Period	ROHs total (kb)	ROHs per segment (kb)	N segments
CL-H017	Historical	315803	4326.07	73
CL-H022	Historical	400337	6159.03	65
CL-H035	Historical	304062	5067.7	60
CL-H036	Historical	284233	5573.19	51
CL-H040	Historical	528805	6009.14	88
CL-H052	Historical	289627	5464.66	53
CL-H083	Historical	792937	12789.3	62
CL-H084	Historical	139357	3871.04	36
CL-H086	Historical	158024	5097.53	31
CL-H100	Historical	184431	3415.39	54
CL-H116	Historical	345029	5750.48	60
CL-H117	Historical	847743	13897.4	61
CL-H119	Historical	172615	4794.87	36
CL-H131	Historical	376234	5452.67	69
CL-H132	Historical	383055	4848.8	79
CL-H146	Historical	254704	5306.34	48
CL-H149	Historical	367555	6563.48	56
CL-H177	Historical	623073	8535.25	73
CL-H182	Historical	595609	9163.22	65
CL-H220	Historical	126941	3340.55	38
CL-H261	Historical	71872	2764.32	26
CL-H137	Historical	169825	4142.08	41
L005	Contemporary	217629	4731.08	46
L029	Contemporary	625302	9474.28	66
L086	Contemporary	721287	8796.19	82
L116	Contemporary	521658	6687.92	78
L125	Contemporary	733560	9780.8	75
L190	Contemporary	312015	5200.25	60
L253	Contemporary	337664	4625.54	73
L316	Contemporary	966511	10621	91
L319	Contemporary	327174	4544.08	72
L320	Contemporary	974026	11325.9	86
L338	Contemporary	692586	8446.17	82
L360	Contemporary	589556	6624.23	89
L382	Contemporary	484122	7225.7	67
L392	Contemporary	843732	10165.5	83
L401	Contemporary	331153	5341.19	62
L422	Contemporary	478380	5833.9	82
L431	Contemporary	831010	10792.3	77
L474	Contemporary	486463	6236.7	78
L502	Contemporary	459279	5467.6	84
L596	Contemporary	237368	4238.72	56
L649	Contemporary	367421	4650.89	79
L772	Contemporary	944412	9350.61	101

Chapter VI.

On taming the effect of transcript level intra-condition count variation during differential expression analysis: a story of dogs, foxes and wolves

Diana Lobo[1,2,3,*], Raquel Linheiro[1], Raquel Godinho[1,2,3] & John Patrick Archer[1,2,*]

* Corresponding authors: diana.lobo@cibio.up.pt (DL); john.archer@cibio.up.pt (JPA)

[1]CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, *InBIO* Laboratório Associado, Campus de Vairão, Universidade do Porto, Vairão, Portugal.
[2]BIOPOLIS, Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, Vairão, Portugal.
[3]Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Porto, Portugal.

Published in *PLOS ONE* (September 22nd, 2022).
DOI: https://doi.org/10.1371/journal.pone.0274591
Published version available in Appendix II.

Abstract

The evolution of RNA-seq technologies has yielded datasets of scientific value that are often generated as condition associated biological replicates within expression studies. With expanding data archives opportunity arises to augment replicate numbers when conditions of interest overlap. Despite correction procedures for estimating transcript abundance, a source of ambiguity is transcript level intra-condition count variation; as indicated by disjointed results between analysis tools. We present TVscript, a tool that removes reference-based transcripts associated with intra-condition count variation above specified thresholds and we explore the effects of such variation on differential expression analysis. Initially iterative differential expression analysis involving simulated counts, where levels of intra-condition variation and sets of over-represented transcripts are explicitly specified, was performed. Then counts derived from inter- and intra-study data representing brain samples of dogs, wolves and foxes (wolves vs. dogs and aggressive vs. tame foxes) were used. For simulations, the sensitivity in detecting differentially expressed transcripts (DETs) increased after removing hyper-variable transcripts, although at levels of intra-condition variation above 5% detection became unreliable. For real data, prior to applying TVscript, ~20% of the transcripts identified as being differentially expressed were associated with high levels of intra-condition variation, an over-representation relative to the reference set. As transcripts harbouring such variation were removed pre-analysis, a discordance from 26 to 40% in the lists of DETs is observed when compared to those obtained using the non-filtered reference. The removal of transcripts possessing intra-condition variation values within (and above) the 97^{th} and 95^{th} percentiles, for wolves vs. dogs and aggressive vs. tame foxes, maximized the sensitivity in detecting DETs as a result of alterations within gene-wise dispersion estimates. Through analysis of our real data the support for seven genes with potential for being involved with selection for tameness is provided. TVscript is available at: https://sourceforge.net/projects/tvscript/.

6.1. Introduction

Developments in RNA-seq technology have revolutionized transcriptomic studies by allowing for a rapid hi-resolution view of transcript expression (Wang et al. 2009). In a typical RNA-seq experiment, transcript expression profiles are estimated for each sample using a metric based upon the number of sequenced reads associated with each transcript within a reference set (Mortazavi et al. 2008; Nagalakshmi et al. 2008; Jiang and Wong 2009; Trapnell et al. 2010; Lee et al. 2011; Kukurba and Montgomery 2015; Conesa et al. 2016). Condition dependent expression profiles can then be used in order to identify which transcripts are differentially expressed (Robinson et al. 2010; Love et al. 2014). A challenge arises due to sources of variation within expression profiles that are independent of, or partially overlapping with, the condition of interest (Hansen et al. 2011; McIntyre et al. 2011; Conesa et al. 2016; Xu and Asakawa 2019). The inclusion of biological replicates reduces the effect of such noise (Liu et al. 2014; Schurch et al. 2016)', and it has been demonstrated that sufficient replicate numbers outweigh sequencing depth in terms of increasing the accuracy within differential expression experiments (Robles et al. 2012; Liu et al. 2014). In studies not involving highly controlled isolated environments, RNA-seq data from the rapidly growing repertoire of published works can be incorporated (Sudmant et al. 2015; Lachmann et al. 2018; Zoabi and Shomron 2021), if data from a matching condition to that being studied is available. This effectively increases the number replicates although variability can be amplified (Rau et al. 2014; Jeng et al. 2021).

Differential expression tools compute a statistical significance for each transcript, based upon the abundance estimates within a condition, that reflect the possibility of that transcript being differentially expressed (Robinson et al. 2010; Love et al. 2014; McDermaid et al. 2019). To reduce the effect of intra-condition variation across biological replicates on the estimation of abundance several methods have been proposed including ALDEx2 (Quinn et al. 2018), EDASeq (Risso et al. 2011) and PEER (Stegle et al. 2010). In addition to these, and more generally applied, are the abundance estimation techniques implemented within established differential expression tools such as DESEQ2 (Love et al. 2014) and EdgeR (Robinson et al. 2010). However, when methods are compared, relative to the final sets of transcripts identified as being differentially expressed, variable results are observed (S. Li et al. 2014; Schurch et al. 2016; Quinn et al. 2018; Wang et al. 2019; Arora et al. 2020; Mou et al. 2020). This is an indication that the problem of intra-condition variation relative to the detection of differentially

expressed transcripts (DETs) using RNA-seq data has not been completely resolved. Furthermore, there is no consensus on the best approach to use (Stupnikov et al. 2021).

Here we explore the effects that individual transcripts associated with high levels of intra-condition count variation have on the end results of differential expression analysis using the tool DESEQ2 (Love et al. 2014); a tool that is well established and that has consistently demonstrated reliability in identifying DETs (Costa-Silva et al. 2017; Wang et al. 2019; Stupnikov et al. 2021). Our aim is to investigate the possibility of whether or not the removal of transcripts, harbouring the highest levels of intra-condition variation, from the reference set used during differential expression analysis can produce sets of DETs that display an increased level of confidence. The latter being achieved through either: (a) the direct removal of transcripts previously identified as being differentially expressed, but whose expression patterns are ambiguous, or (b) the indirect addition, or removal, of transcripts to, or from, those previously identified as being differentially expressed as a consequence of alterations in p-adj values. The latter being associated with shifts in the distribution of intra-condition variation, following the removal of transcripts harbouring the highest levels of such variation. A by-product of this is the explicit quantification of the level of intra-condition abundance variation present within the final lists of DETs.

To aid this exploration we present TVscript, a tool for the identification of transcripts above user-specified levels of intra-condition normalized count variation, the latter being strongly associated with transcript abundance estimation (Mortazavi et al. 2008; Nagalakshmi et al. 2008; Jiang and Wong 2009; Trapnell et al. 2010; Lee et al. 2011). As input TVscript requires one file per condition-associated replicate that contains the per transcript read counts obtained following the mapping of reads from the replicate to a common reference set. As output TVscript produces a set of corresponding count files that are absent of transcripts harbouring normalized intra-condition count variation higher than that associated with a user specified percentile. These updated count files can be subsequently used within the differential expression tool of choice, in our case DESEQ2 (Love et al. 2014). Through multiple iterations of differential expression analysis following filtering at varying thresholds and comparisons back to differential expression analysis performed on non-filtered inputs, the effects of transcripts associated with high intra-condition variation, in relation to quantity and consistency of DETs identified, can be explored.

Using TVscript we first explore the effects of intra-condition per-transcript read count variation through iterative differential expression analysis experiments involving highly controlled simulated count datasets derived from the available dog reference transcriptome (Hoeppner et al. 2014), and where the exact level of background intra-condition count variation could be specified as well as a subset-set of transcripts to be over-represented across replicates (of second conditions used within each iteration). Next, we explored the effects of intra-condition per-transcript read count variation within two distinct case-studies, involving count data obtained following the mapping of intra and inter-study RNA-seq datasets. Within these case-studies differential expression patterns arising from data derived from brain samples of dogs and wolves (inter-study scenario involving frontal cortex, cerebral cortex, prefrontal cortex and frontal lobe) (Albert et al. 2012; Roy et al. 2013; Hoeppner et al. 2014; Fushan et al. 2015), as well as tame and aggressive foxes (intra-study scenario involving prefrontal cortex) (Wang et al. 2018), generated in the scope of domestication experiments are compared at varying thresholds of intra-condition normalized read count variation exclusion.

In addition to exploring the general effects of transcripts harbouring high levels of intra-condition count variation on the outcome of differential expression analysis, we also had an interest in understanding whether or not there were genes commonly up or down-regulated within the brain of both forms of domestic canids (dogs and tame foxes), but simultaneously not so within their “wild/aggressive” counter parts (wolves and wild foxes). Such genes are candidates for being associated with tameness. Domestic dogs present marked behavior differences from wolves, their wild ancestors, due to the evolution of unique social cognitive capabilities (Albert et al. 2012; Li et al. 2013; Y. Li et al. 2014). Tame red foxes resulted from deliberated selection against fear and aggression over several generations of cross-breeding (Lord et al. 2019) and they present several behavioral and phenotypical traits that resemble those found in dogs (Kukekova et al. 2011a; Hekman et al. 2018; Wang et al. 2018).

TVscript is open source and code, a quick start guide and test data, are available (under the GNU General Public License) through the SourceForge project page https://sourceforge.net/projects/tvscript/.

6.2. Materials and Methods

6.2.1. RNA-seq datasets

To explore the effects of intra-condition count variation on the detection of DETs using real data we used both intra and inter-study datasets. At an intra-study level, we combined publicly available RNA-seq data from brain tissue (prefrontal cortex) of 12 tame and 12 aggressive red foxes (Supplementary Table 6.1) generated within the same study (Wang et al. 2018). For the inter-study case, we combined multiple publicly available RNA-seq datasets from several dogs and six wolves (Albert et al. 2012; Roy et al. 2013; Hoeppner et al. 2014; Fushan et al. 2015), also derived from brain tissue (frontal cortex, cerebral cortex, prefrontal cortex and frontal lobe) (Supplementary Table 6.1). In relation to the latter, dogs 1 to 6 and wolves 1 to 6 were derived from Albert et al. (2012), dog 6 from Roy et al. 2013, dog 7 (two replicates) from Fushan et al. 2015 and dogs 8 and 9 (three replicates each) from Hoeppner et al. (2014) as described within the table. All the samples were downloaded from the National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EMBL-EBI), covering a wide range of ages, both sexes as well as multiple replicates and sequencing strategies (Supplementary Table 6.1). We selected these specific case studies because, firstly we were interested in evaluating the effects of intra-condition count variation both at intra and inter-study levels and the domestic dog, being a model organism, has an available high-quality reference transcriptome as well as several high-quality RNA-seq datasets generated across different studies; while secondly, we sought to perform a brief exploratory inter-study scan to investigate if the domestication of both dogs and foxes has resulted in the co-expression of a set of common brain genes, relative to their "wild/aggressive" type, since behavioral modifications are considered to have been the first target in domestication (Hou et al. 2020).

Reads from all samples were mapped to the dog reference transcriptome (Hoeppner et al. 2014), which contained 26,107 annotated transcripts (Ensembl CanFam3.1, release 92) (Yates et al. 2020), using BOWTIE v.2.3.4.1 (Langmead and Salzberg 2012). We did not use a splice-site aware mapper, such as TOPHAT2 (Kim et al. 2013) or HISAT2 (Kim et al. 2019), since introns were not expected to be present within the reference transcriptome. Reads were not mapped to the dog genome as our aim was to explore the effects of intra-condition variation on a predefined set of reference transcripts, and not infer novel transcripts from these previously published datasets. We did not use the

fox reference transcriptome (available on Ensembl) for the mapping the fox datasets within our overall analysis as we were interested in directly comparing DETs from the same underlying reference set identified using the fox, dog and wolf datasets. We did however map all of the fox datasets to the fox reference transcriptome in order to confirm that the proportion of reads mapped was similar to that when using the dog reference transcriptome. The *pileup.sh* script from the BBMAP package (Bushnell and Brian 2014) was used to obtain per transcript abundance estimates (measured by the number of mapped reads to the corresponding transcript) in each sample. Read counts from technical replicates of "Dog_8" and "Dog_9" were averaged and merged into one file (Supplementary Table 6.1), while read counts from the two biological replicates of "Dog_7" were treated separately. Finally, to confirm that the final read count numbers were reliable relative to the dog reference transcriptome, all read datasets from foxes, dogs and wolves were re-mapped using the pseudo-mapper KALLISTO v0.46.1 that implements a rapid and accurate kmer search based strategy for estimating transcript abundance counts (Bray et al. 2016). For each dataset an r^2 correlation value was calculated describing the linear correlation between the per-transcripts counts obtained following BOWTIE2 mapping and the corresponding abundance counts obtained using KALLISTO.

6.2.2. Software

TVscript requires as input: (1) multiple files containing per transcript read counts (one per sample), (2) a file containing the lengths of the transcripts that the reads were mapped to, (3) a percentile threshold value for intra-condition variation and (4) a configuration file that indicates the locations all files as well as the condition allocations of the count files in (1). An example configuration file along with further details is available from the SourceForge project page. The steps that TVscript implements to identify transcripts associated with high levels of disjointed read counts are (see **Figure 6.1** for a workflow): i) each input dataset, containing count values from a particular sample, is allocated to either condition A or B, as indicated within the configuration file; ii) counts are normalized by dividing them by the length of the corresponding reference transcript and by the sum of all counts for that sample; (iii) for each reference transcript (t), the absolute pairwise differences between normalized read counts across all samples within condition A are calculated; (iv) the corresponding variances are calculated; (v) steps (iii) and (iv) are repeated for condition B; (vi) variance scores from each condition are placed in ascending order and associated with corresponding percentiles; (vii) reference transcripts are removed (or filtered) if their variance score is above that associated with

the user specified percentile threshold; (viii) raw read counts associated with the remaining transcripts are outputted into separate files that correspond to each input dataset. As to avoid overwriting the original count files, the names of updated count files are specified within the configuration file. These updated count files can be used as input for differential expression analysis software such as DESEQ2. Note: in relation to step (vi) the final list of variance scores obtained are a representation of those derived from both conditions, and during step (vii) it is the variance score that is associated with the user-defined percentile that is used. The latter means that for separate comparisons (e.g., wolves vs. dogs and aggressive vs. tame foxes) at a given filtering threshold the number of filtered transcripts may not be identical (despite the same reference set being used); as the underlying variance score distribution calculated for each can be different.

TVscript is implemented in Java programming language and runs on all operating systems with installed Java Runtime Environment v.8.0 or higher. It is open source and available under the GNU General Public License v.3.0. Source code, usage instructions and sample data can be found on the SourceForge project page: https://sourceforge.net/projects/tvscript/. Although TVscript is implemented in Java the steps involved can be readily implemented within any language (e.g., R or python), using the detailed description provided above as well as the Java source code that is fully available. There are no dependent packages where code is unavailable. At the time of development, we choose Java mainly due to its platform independence, which can be an advantage within setting up analysis pipelines involving many different tools. That said we are aware that many differential expression analysis tools are R based and future demand may warrant a supported R version.

6.2.3. Controlled intra-condition variation within simulated data

We tested TVscript through a series of iterative differential expression analysis experiments involving highly controlled simulated count data. During each iteration the counts associated with each transcript, within each replicate dataset created, represented simulated transcript expression from the dog reference transcriptome and were obtained using CSReadGen (Linheiro and Archer). Using the latter, the level of random background count variation away from that required for a normalized even coverage across all transcripts could be specified, as well as a subset of transcripts to be over-represented within replicates of a condition. Background count variation refers to varying count levels associated with individual transcripts that are not maintained across replicates of a condition, thus effectively reflecting intra-condition noise, whilst

specifying a subset of transcripts to be over-represented across replicates of a condition reflects identifiable over expressed transcripts. A similar experiment to those described here, but in relation to the effects of chimerism on the results of differential expression, has been described in Linheiro and Archer (2021).

As a preliminary, and to demonstrate the general reliability of DESEQ2 in the absence of random intra-condition count variation, count data was simulated for 22,580 transcripts ranging in length from 300 to 5000 bp from within the dog reference transcriptome. For a single iteration ten replicates of count dataset, each representing three million read pairs (~20X coverage), were generated in accordance with two conditions, A and B (five in each), where within condition B 100 transcripts were selected for count over-representation by a factor of two across replicates. For all other transcripts, counts were generated to represent an even distribution of reads (length 150 bp, insert size 300 bp). These count files were used as input to perform differential expression analysis between condition A and B in DESEQ2 v.1.32.0, considering transcripts with p-adj < 0.05 (corrected by the Benjamini and Hochberg method) to be differentially expressed. The number of transcripts that were detected as being over expressed was recorded. This was repeated 100 times in two different ways: (i) the transcripts initially flagged for count over-representation were kept constant throughout and (ii) during each iteration a new set of random transcripts for over-representation was selected. A brief overview of the R-script we used for differential expression analysis within individual iterations is available on the Zenodo repository (Lobo et al. 2022a).

Next, a similar experiment was performed but where the level of random count variation introduced into the count datasets generated ranged from 1% to 10% in steps of one. Introduced variation was not coupled between replicates, thus reflected intra-condition variation. At each level of variation 100 iterations of the following steps were performed: (i) ten replicates were generated and allocated into two conditions A and B (five in each), where within B 100 selected transcripts had counts over-represented; (ii) at the percent level of variation associated with the iteration, that percent of transcripts from each of the ten replicates were randomly selected for count over-representation; (iii) DESEQ2 was used in a similar manner to before on the count files within conditions A and B to obtain a list of over expressed transcripts; (iv) TVscript was run using a 95th percentile variance threshold to generate ten corresponding modified count files also separated into two conditions (A' and B'); (v) DESEQ2 was again used on these to obtain a list of over expressed transcripts; (vi) the lists of over expressed transcripts obtained in (iv) and (v) were cross compared. Once again this was repeated in two different ways: (i) the 100

transcripts initially flagged for over-representation were kept constant throughout all levels of variation and for each of the associated iterations and (ii) during each level of variation and for each iteration a new set of 100 transcripts were randomly selected.

6.2.4. Exploring the removal of transcripts associated with high intra-condition variation within real data

For each case study (wolves vs. dogs and aggressive vs. tame fox) we ran TVscript using the count datasets described under the section "RNA-seq datasets" and by applying variance filtering thresholds corresponding to variance values associated with the 70th up to the 90th percentiles (in steps of five), and to the 91st up to the 99th (in steps of one). Steps of one were used in the latter as to allow for transcripts associated with the highest levels of intra-condition variation to be explored in more detail. For each threshold value, only transcripts with variance below that value were maintained. During each run, we recorded the number and IDs of all transcripts that were removed so that they could be cross-compared. Following each run, we used the updated count files produced to perform differential expression analysis using DESEQ2 v.1.22.2. Transcripts with p-adj < 0.05 (corrected by the Benjamini and Hochberg method) were considered to be differentially expressed. DESEQ2 identifies DETs by estimating gene-wise dispersions and applying shrinking methods to model counts and thus effectively normalize for individual outliers (Costa-Silva et al. 2017). Distributions of gene-wise dispersions following normalization are conveniently accessible and provide a good metric to visualize the effects of removing transcripts associated differing filter levels. Differential expression analysis using the original non-filtered datasets was also performed. For the aggressive vs. tame fox case study batch effects were not considered as all data came from the same study, tissue and sequencing run, additionally no further information about sample preparation was available. For the wolves vs. dogs case study we tested for effects based on tissue, primarily for quality control of the final transcripts we drew biological-related conclusions about, and compared results obtained to those in the absence of batch information. In our analysis we used differential expression results based solely on the latter, as firstly, effects associated with tissue at an inter-study level are unpredictable as there are many factors involved, such as precision of dissection, time of dissection, time to dissect, state of individual tissue samples as well as individual who prepared sample, and other than publication or information mentioned for the fox case study, no further information on batches was obtainable. Secondly, although DESEQ2 provides an internalized method for accommodating batch effects that

we applied (~batch + condition), the results obtained at an intra-study level, with well-defined batches, between alternative methods of testing are variable (Tran et al. 2020). Lastly, we were primarily exploring the effects of removing hyper variable transcripts on the mechanics of detecting DETs and our simulations and case studies were a means to an end in achieving this. As long as input counts for a given filtering threshold within a given case study or an iteration were consistent with those of the initial input data, the effects of removing hyper variable transcripts could be observed, independent of other factors affecting the data prior to analysis. To visualize the overall effects of covariates broadly affecting the relationships between datasets within each case study, we performed a principal component analysis (PCA) using the *plotPCA* function from DESEQ2 with non-filtered normalized count data.

To evaluate TVscript we used three metrics that when combined quantify the overall impact of intra-condition variation on downstream differential expression analysis. The metrics were: i) number of ambiguous positives within transcripts identified as being differentially expressed in the non-filtered datasets; ii) distributions of dispersion estimates and outliers in differential expression analysis for non-filtered and all filtered datasets; and iii) discordance in the list of DETs between non-filtered and filtered datasets (selected percentiles: 97^{th}, 95^{th}, 90^{th}).

(i) Ambiguous positives

We identified transcripts appearing as being differentially expressed when using the non-filtered datasets as input to DESEQ2 that were associated with the top 10 levels of intra-condition variation (above the 90^{th} percentile threshold value). These we designated as ambiguous positives. Small numbers of these, relative to the overall number of identified DETs would indicate that TVScript is having little direct effect on lists of identified DETs.

(ii) Distributions of dispersion estimates

For the non-filtered and all filtered input datasets (70^{th} up to the 90^{th} percentiles in steps of five and to the 91^{st} up to the 99^{th} in steps of one) we calculated correlation coefficients (r^2) using a linear regression analysis in R (R Development Core Team 2017), between dispersion estimates and the mean of normalized counts, both the latter calculated by DESEQ2 during differential expression analysis. Dispersion is inversely related to the mean, as lower mean counts are affected by variation to a higher degree. If a stronger correlation is seen for the filtered input datasets, then this would suggest that the distribution used to model differential expression could be more reliable in relation to

identifying DETs. In addition to this, we retrieved the number of outliers detected by DESEQ2, expecting a decrease after each filtering step. Outliers are recognized by the DESEQ2 as the points with extremely high dispersion values that cannot by shrunk towards the fit curve. This was performed independently for both case studies.

(iii) Discordance lists of DETs between applied filter levels

We calculated the proportion of discordance between lists of DETs produced when using non-filtered and filtered datasets at the 97th, 95th and 90th percentile threshold values. Two types of observed discordances relative to the non-filtered list were considered: (a) transcripts that were lost directly due to filtering or indirectly due to p-adj values no longer being significant, and (b) transcripts that were added due to alterations in p-adj values. Quantifying the nature of these discordances provides insight into the general consistency of transcripts identified as being differentially expressed across varying filter thresholds. To visualize the overlap between the non-filtered and filtered lists we used the *VennDiagram* v.1.6.20 package in R.

6.2.5. Gene annotation and gene family analysis

For each case study, DETs obtained using the non-filtered and filtered datasets were matched to the correspondent gene ID. This was done with the R package *BioMart* (Durinck et al. 2005) using the Ensembl Gene database (version 94). To begin to identify gene families that displayed similar regulation in both dogs and tame foxes, i.e., relative to the “wild/aggressive” type, we grouped up and down-regulated genes into gene families. Genes within these families were then classified according to whether they were unique to dogs or tame foxes or shared between the two. Within each case study, a gene family was only considered if all the associated genes agreed in relation to their direction of expression (up or down-regulation).

6.3. Results

6.3.1. Mapping success of RNA-seq data

Mapping of the 44 datasets corresponding to dogs and wolves against the dog reference transcriptome revealed an average success of 60% and 58% respectively, in terms of the number of mapped reads (Supplementary Figure 6.1). Similar values between

wolves and dogs were expected, given their recent divergence of ~23,000 years ago (Perri et al. 2021). Comparable proportions of reads failing to map (~40%) have been previously reported for dog brain samples (Roy et al. 2013) and are most likely associated with i) novel genes; ii) regions that are not translated despite being transcribed; iii) contamination with genomic DNA; and iv) uncharacterized chimeras and other artefacts within reference sets resulting from library preparation during sequencing (Tu et al. 2015) and various assembly errors (Rizzi et al. 2019). When a different mapping approach was used for each RNA-seq dataset (i.e., KALLISTO) transcript abundance counts remained consistent with those obtained following BOWTIE2 mapping, as indicated by high r^2 correlation values (Supplementary Figure 6.2). R^2 values ranged between 0.85 and 0.99. All per-transcript mapped read counts, obtained following each mapping approach, have been made available on the Zenodo repository (Lobo et al. 2022b). For the fox datasets, an average of 50% of reads mapped to the dog reference transcriptome using BOWTIE2 (Supplementary Figure 6.1); confirmed by the KALLISTO estimated abundance counts (Supplementary Figure 6.2). This lower percentage of mapped reads, relative to the dog and wolf datasets, could be expected due to an increased genetic divergence from dogs (~10 million years ago) (Wayne et al. 1997) together with the other aforementioned factors. However, when these fox datasets were mapped to the Ensembl available fox reference transcriptome using both BOWTIE2 and KALLISTO an improvement in the overall mapped read counts was not observed (Supplementary Figure 6.3).

6.3.2. Controlled intra-condition variation within simulated data

When not faced with increasing levels of random intra-condition count variation DESEQ2 performed exceptionally well. For 86 of the 100 iterations performed, DESEQ2 recovered all transcripts (N = 100) that were selected for count over-representation (Supplementary Figure 6.4A). Of the other 14 iterations the lowest number recovered was 72. Similarly, for the 100 iterations where the random transcripts selected for count over-representation were re-selected during each, in 87 cases all over-represented transcripts were identified as being over expressed, whilst in the remaining 13 the minimum number identified was 72 (Supplementary Figure 6.4B). However, as levels of introduced intra-condition variation increased, the number of transcripts identified by DESEQ2 fell, both before and after filtering the input counts with TVscript (**Figure 6.2**). At all levels of intra-condition variation, the post-filtered data had an increase in the number of DETs identified. It should be emphasized that the reduction in the number of transcripts identified as being over expressed is not a negative reflection on the performance of

DESEQ2, but instead it is a consequence of purposefully increasing the level of randomness within the count data. The same pattern is true when the 100 selected transcripts for count over-representation are re-selected within each iteration (Supplementary Figure 6.5).

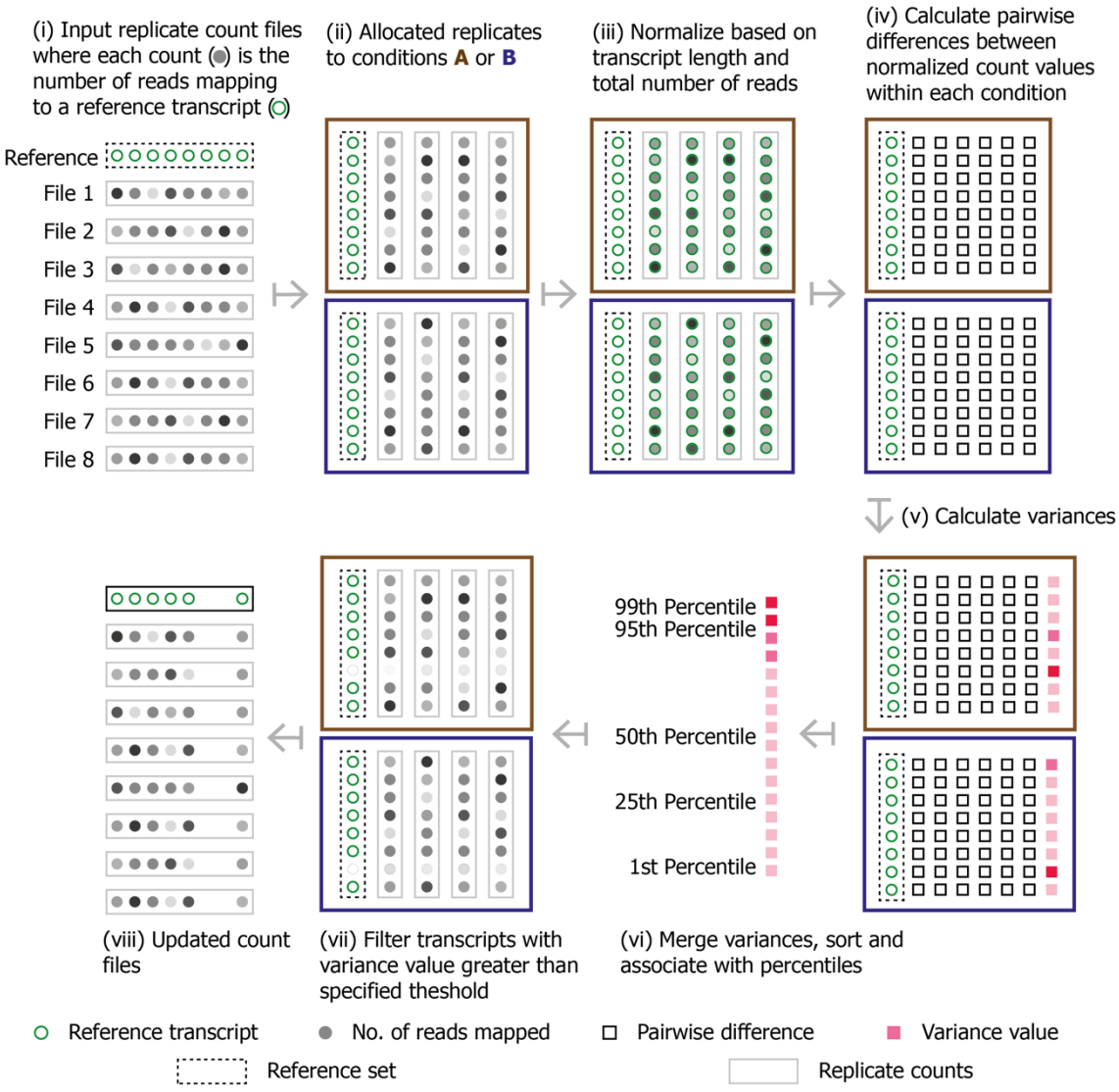


Figure 6.1. Diagramatic overview of how reference based transcripts are removed by TVscript. Steps (i) to (viii) indicate actions taken. Reference transcripts (green circles) are showen for diagramtic purposes in order to highlight how read counts (gray circles) across replicates are treated for each transcript independently. Read counts are grouped into individual files (gray rectangles) in accordance to replicate. These files are grouped into one of two conditions (blue and brown boxes). Remaining keys are indicated at the bottom.

For iterations associated with each increment in random intra-condition variation, the number of transcripts commonly identified as being over expressed both prior-to and post filtering are presented in Supplementary Table 6.2 (over-represented transcripts kept constant across iterations) and Supplementary Table 6.3 (over-represented transcripts re-selected for each iteration). The proportion that these numbers make up relative to the maximum number of transcripts identified as being differentially expressed, pre- and post-filtering, are presented in Supplementary Tables 6.4 (constant) and 6.5 (re-selected). In all cases, below a 5% level of random variation these numbers are high (constant - 1 to 4% averages: 0.96, 0.98, 0.94 and 0.81; re-selected - 1 to 4% averages: 0.96, 0.97, 0.93 and 0.82), indicating that on top of additional transcripts identified post-filtering with TVscript, transcripts identified pre-filtering are still found. Consequently, this suggests that additional transcripts identified as being over expressed post-filtering are not at the expense of previously detected transcripts pre-filtering. Above the 5% level of intra-condition variation the ability to successfully identify the 100 transcripts selected for over-representation within condition B diminishes within iterations (**Figure 6.2**; Supplementary Figure 6.5). This could be indicative of a tentative estimate on the limit of at what level of random intra-condition count variation becomes inhibitory within differential expression analysis studies.

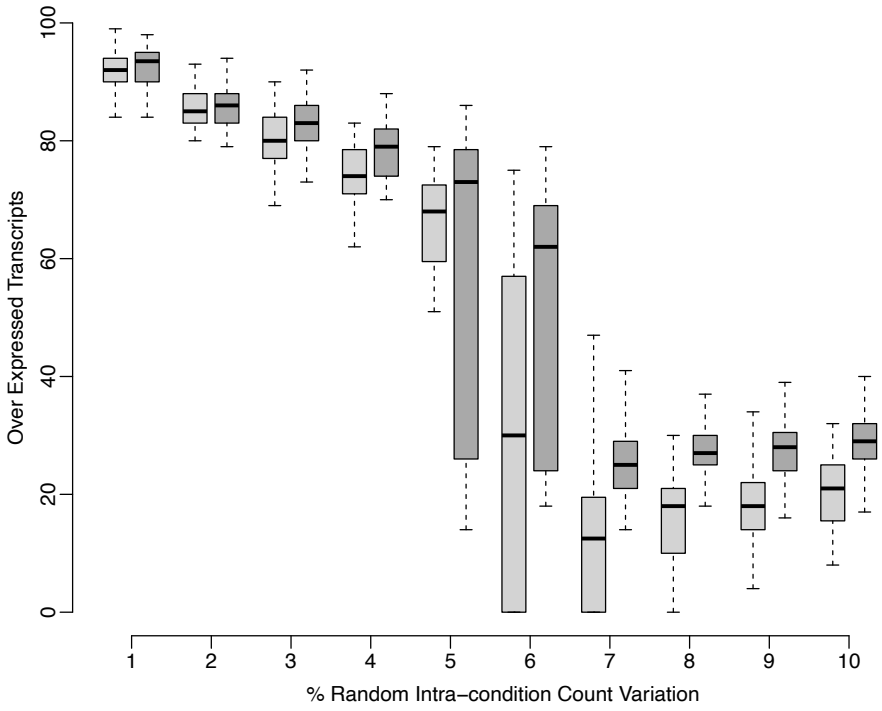


Figure 6.2. Over expressed transcripts pre- and post-filtering using simulated data. The number of transcripts identified by DESEQ2 as being over expressed both prior to (light gray) and post (dark gray) filtering of count datasets

within each of the 100 iterations performed at each level of introduced random intra-condition count variation. Each iteration involved initially simulating ten count datasets divided into conditions A and B following which DESEQ2 was run to attempt to identify the 100 transcripts selected for over-representation as described in the methods. Following this, the ten simulated datasets were filtered using TVscript with a 95th percentile threshold to generate corresponding filtered datasets (divided into corresponding conditions A' and B') on which DESEQ2 was re-run.

6.3.3. Exploring the removal of transcripts associated with high intra-condition variation within real data

No significant difference existed between the overall distributions of the per-transcript intra-condition variation values for wolf and dog samples (Wilcoxon-test, p-value < 0.198, **Figure 6.3A**). The PCA based on the entire set of normalized non-filtered counts, revealed that the wolf samples were more aggregated than dog samples (**Figure 6.3C**). For aggressive and tame fox samples, we observed a significant difference (Wilcoxon-test, p-value < $2.2e^{-16}$, **Figure 6.3B**) between the distributions of the per-transcript intra-condition variation values, most likely resulting from an increased intra-condition variability within tame fox samples. In particular, we found five samples that were differentiated from the remaining seven in the PCA (**Figure 6.3D**), with 80% variance being explained by this clustering in PC1.

Prior to differential expression analysis, for each case study (wolves vs. dogs and aggressive vs. tame foxes), TVscript was used to remove transcripts in accordance with a series of intra-condition variance thresholds (**Figures 6.4A**, **6.4B**; Supplementary Table 6.6). Initially, for wolves and dogs 184 transcripts (out of the 26,107) associated with high intra-condition variation (99th percentile and above) were removed, while for the aggressive and tame fox samples, 235 transcripts were removed. The number of transcripts removed was higher for the fox samples than for those of wolf and dog, reflecting the higher intra-condition variability present. Combined across the top ten levels of intra-condition variation, 12% (N = 3134) and 14.89% (N = 3888) of the reference transcripts were removed in wolf/dog datasets and aggressive/tame fox datasets, respectively (Supplementary Table 6.6).

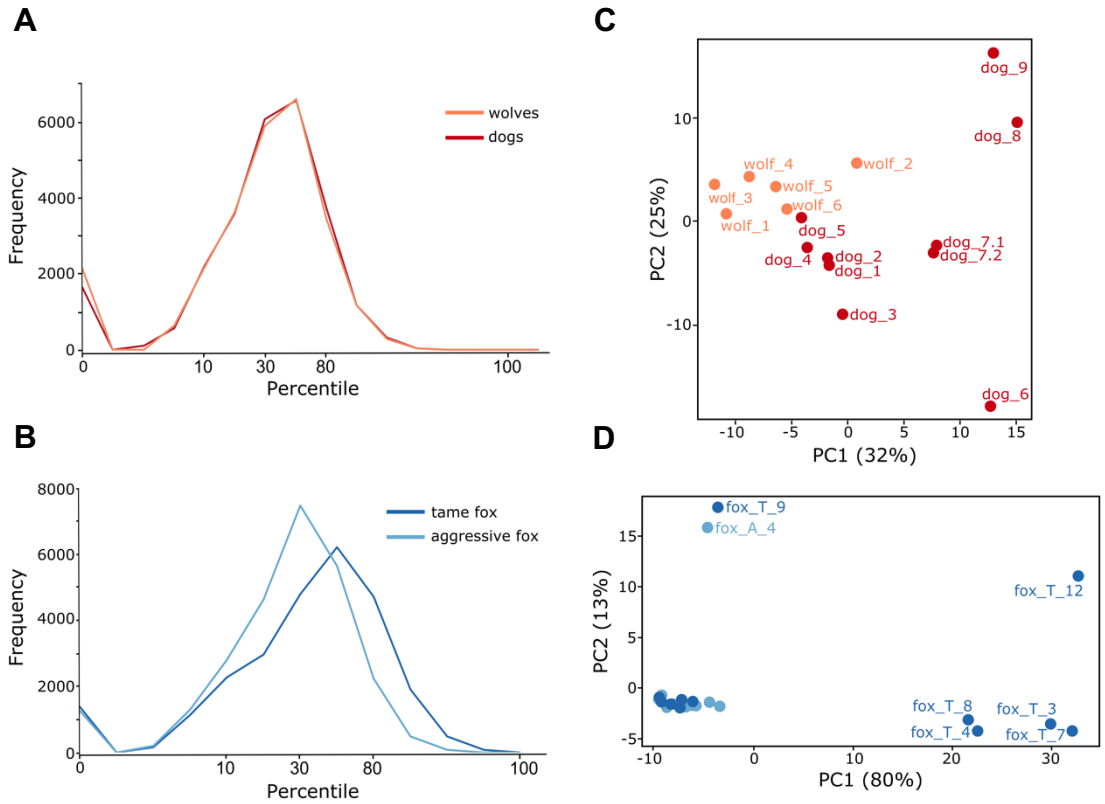


Figure 6.3. Characterization of intra-condition variation. Percentile range of intra-condition variation scores (x-axis) observed prior to filtering, across both case studies, (**A**) wolves (orange) and dogs (red); (**B**) tame (dark blue) and aggressive (light blue) foxes. PCA plots based on normalized non-filtered count data of the individual datasets comparing (**C**) wolf and dog, and (**D**) tame and aggressive fox. In the latter only individual samples that were positioned within a distant cluster are labelled with the sample ID.

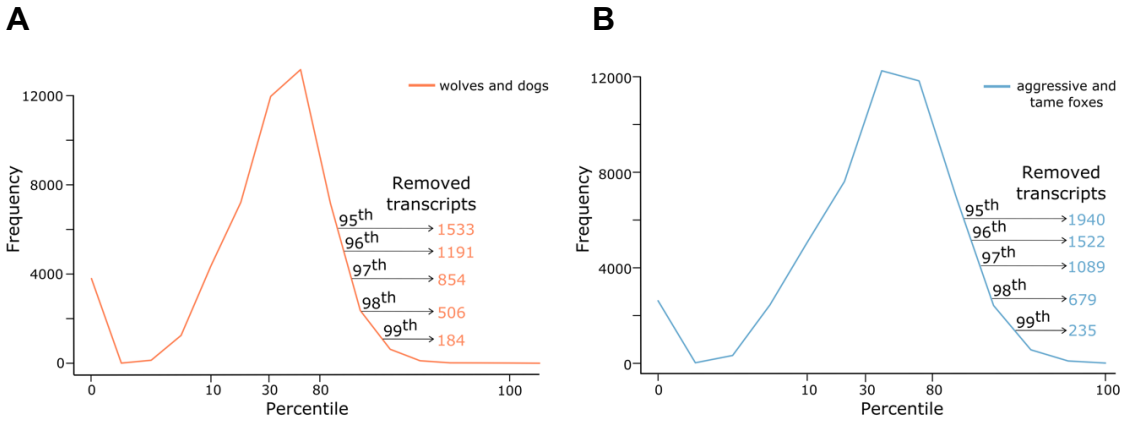


Figure 6.4. Removal of transcripts above specified levels of intra-condition variation. Percentile range of combined intra-condition variation scores (x-axis) present in each case study, (**A**) wolves and dogs; (**B**) tame and aggressive foxes. The number of transcripts removed in the top five percentiles (from the 95th to the 99th) are presented in each panel.

6.3.4. Differential expression analysis

Using non-filtered datasets as input to DESEQ2, 430 DETs were identified between wolves and dogs (**Figure 6.5A**; Supplementary Table 6.7). Of those, 259 were up-regulated, while 171 were down-regulated in dogs. Between aggressive and tame foxes, 651 DETs were identified (**Figure 6.5A**; Supplementary Table 6.8), of which, 532 and 119 were up and down-regulated, respectively, in tame foxes. Post filtering, within the first ten steps of size one from the 99th down to the 90th percentiles, the number of DETs identified peaks at the 97th (N = 430; up = 255, down =175) and the 95th percentiles (N = 730; up = 607, down = 123) in dogs and tame foxes (**Figure 6.5A**), respectively. This indicates that for these data the removal of the 3% (N = 854) and 5% (N = 1940) of transcripts associated with the highest levels of intra-condition variation maximized the detection of DETs.

6.3.5. Evaluation Metrics

(i) Ambiguous positives

Of the transcripts that appeared as being differentially expressed, when using non-filtered datasets as input to DESEQ2, 17.44% (N = 75) and 21.51% (N = 140) were associated with high intra-condition variation (above the 90th percentile threshold) within the wolves vs. dogs and aggressive vs. tame foxes, respectively (**Figure 6.5B**). This was higher than the relative proportion of such transcripts with the reference set in general, where 12.08% and 14.89% of transcripts possessed intra-condition variation above the 90th percentile for the wolf/dog and the aggressive/tame fox, respectively (**Figures 6.4A**, **6.4B**; Supplementary Table 6.6). These were transcripts that we considered as ambiguous positives and the average across both case studies was 19.45%. The number was higher within the fox datasets where elevated variability among samples was observed, suggesting that differences within intra-condition read counts could have influenced the final outcome of identified DETs.

(ii) Distributions of dispersion estimates

Within both case studies, the regression analysis indicates that removing transcripts associated with high levels of intra-condition variation improved correlation coefficients in relation to those from the non-filtered datasets (**Figures 6.6A**, **6.6B**; Supplementary Table 6.9). Associated with the elevated levels of variation observed within the fox datasets, there was a better fit within the wolves vs. dogs comparison ($r^2 > 0.7$) than that

of the aggressive vs. tame fox one ($r^2 > 0.5$). For the latter, there was visible elevation in the number of dispersed points around the line of best fit. With the removal of transcripts associated with the highest levels of intra-condition variation a reduction in the number of outliers within both case studies was also observed (**Figures 6.6A**, **6.6B**; Supplementary Table 6.9).

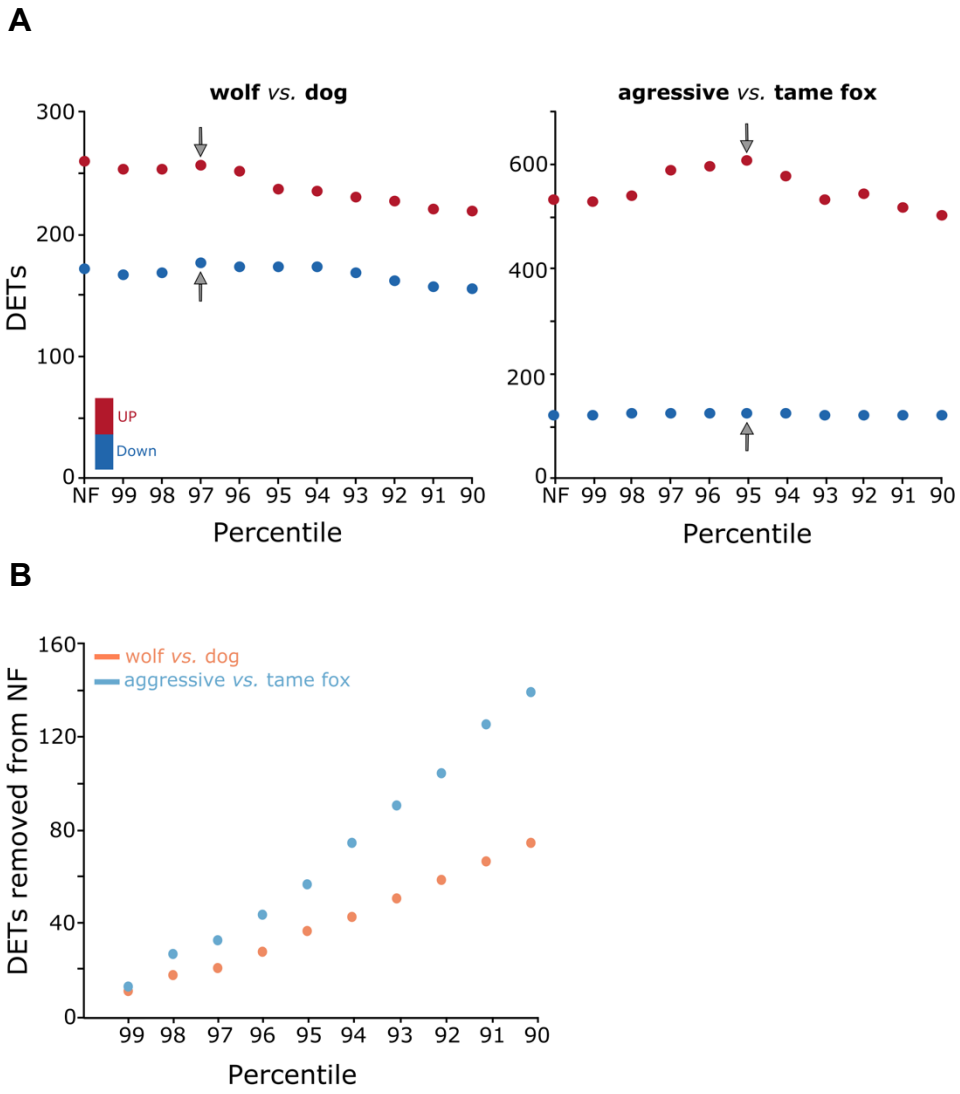


Figure 6.5. Effects of removing transcripts above specified levels of intra-condition variation on differential expression analysis. (**A**) Number of differentially expressed transcripts (DETs) identified using non-filtered (NF) and filtered datasets, based on the top 10 percentiles (99^{th} to the 90^{th}), for both case studies. Up and down-regulated transcripts are represented by red and blue dots, respectively. Gray arrows identify the selected thresholds for which the results of subsequent corresponding differential expression analysis were used for the identification of candidate transcript associated with tameness within each case study. (**B**) Number of transcripts identified as differentially expressed within the non-filtered datasets that were associated the highest levels of intra-condtion variation (99^{th} to 90^{th}) within both case studies, wolves and dogs (orange dots), and tame and aggressive foxes (blue dots).

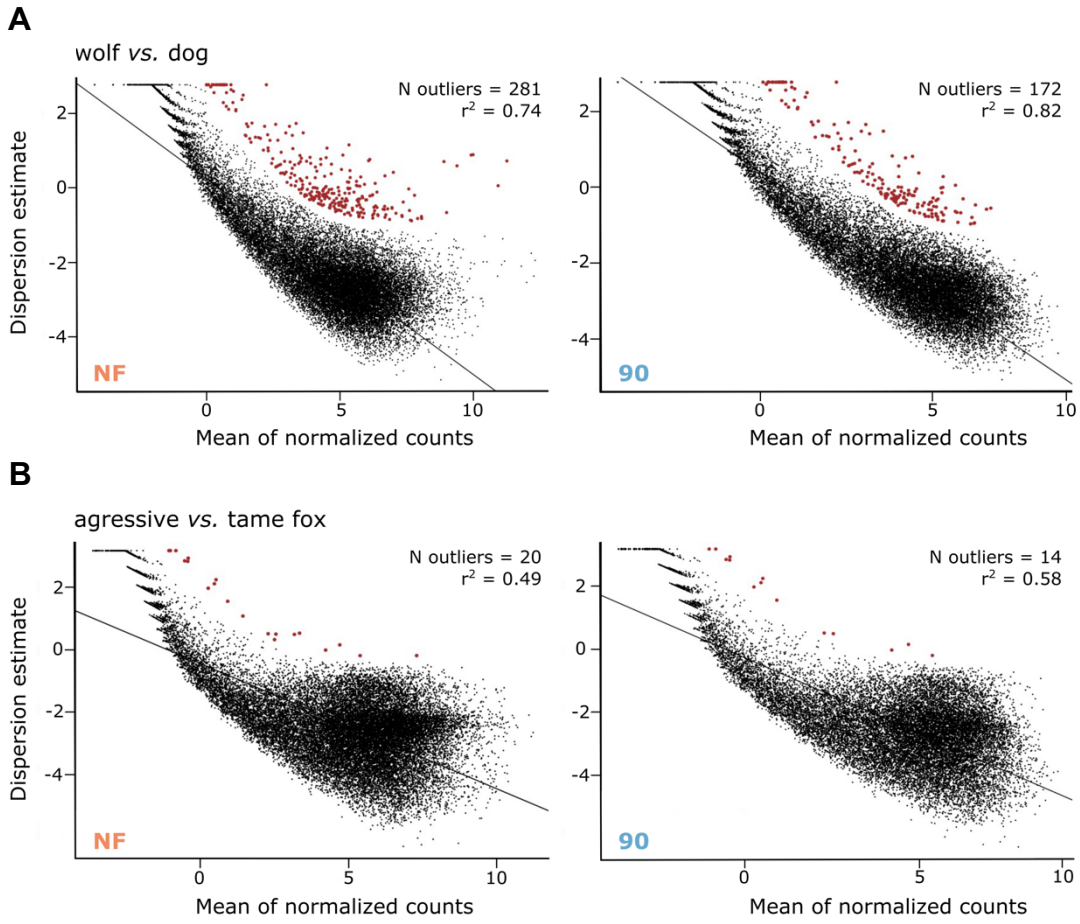
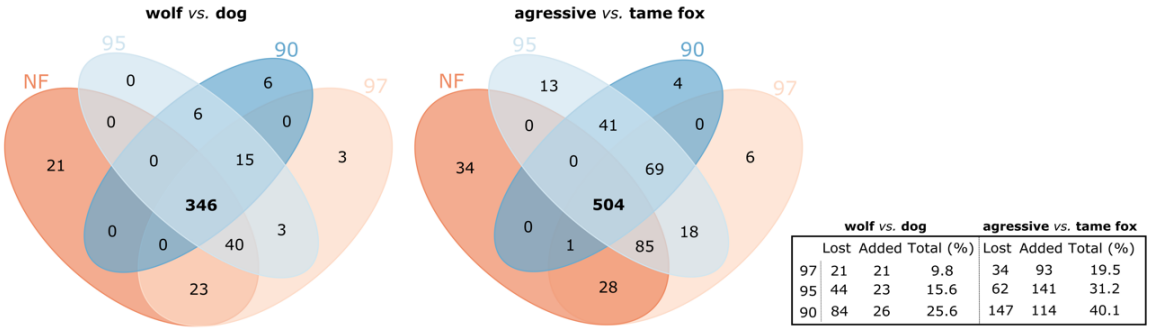


Figure 6.6. Distribution of dispersion estimates. Plots of final dispersion estimates for both case studies, (**A**) wolves and dogs; (**B**) tame and aggressive foxes, calculated using DESEQ2 for the non-filtered (NF; orange) and 10% filtered datasets (90^{th}; blue). Each black dot represents a single transcript, and red dots represent outliers. The number of outliers and correlation index (r^2) are displayed in the top right corner of each pannel. Both x and y-axis are transformed into a logarithm scale. The line in each graph corresponds to the regression analysis between the mean of normalized counts and dispersion estimates.

(iii) Discordance lists of DETs between applied filter levels

Within the wolves vs. dogs case study, from the 430 DETs identified when using non-filtered data as input to DESEQ2, 346 were maintained when using input data filtered at the 97^{th}, 95^{th}, and 90^{th} percentile threshold values (**Figure 6.7** and Supplementary Table 6.7). 26 transcripts were added as differentially expressed following filtering. For this case study, the overall discordance between the DETs identified using filtered and non-filter input data was 25.58% (**Figure 6.7** - inset table). For the second case study, aggressive vs. tame foxes, 504 out of the 651 DETs identified using the non-filtered

inputs were maintained when using filtered input data at the 97th, 95th, and 90th percentile threshold values, with up to 114 being added following filtering (**Figure 6.7** and Supplementary Table 6.8). This time the overall level of discordance was 40.09% (**Figure 6.7** - inset table). Importantly, in both case studies the added transcripts were consistently maintained across the three filter levels. This reflects the general tendency observed within iterative testing using simulated data where the DETs identified using lower filter levels are maintained at higher levels of filtering in addition to any newly identified transcripts (Supplementary Tables 6.4 and 6.5).



	wolf *vs.* dog			agressive *vs.* tame fox		
	Lost	Added	Total (%)	Lost	Added	Total (%)
97	21	21	9.8	34	93	19.5
95	44	23	15.6	62	141	31.2
90	84	26	25.6	147	114	40.1

Figure 6.7. DETs overlapping between non-filtered and filtered datasets. Venn diagrams representing the number of overlapping DETs found following differential expression analysis using non-filtered datasets and filtered datasets (97th, 95th and 90th percentiles), within each case studies. The inset table provides information about the number of DETs lost/added following each filter step in relation to the non-filtered dataset as well as the the percentage of total discordance.

6.3.6. Candidate genes and gene families

By performing annotation using the filtered datasets where the number of DETs was maximized (3% and 5%, in dogs and tame foxes, respectively), we found 21 gene families in common among the up-regulated genes in dogs and tame foxes. These 21 gene families contained 50 genes (**Table 6.1**), of which 19 were exclusive to dogs while 24 were exclusive to tame foxes. The remaining seven genes (RGR, CHRNA5, SQLE, ARHGAP25, ITGA7, MYO7A and TRIB2), were found to be commonly up-regulated in both dogs and tame foxes. When batch effects based on tissue were considered RGR, CHRNA5, MYO7A and TRIB2 were maintained as being commonly up-regulated (Supplementary Figure 6.6). Note: although in relation to the latter SQLE, ARHGAP25 and ITGA7 were lost, batch effects based on tissue across multiple studies where little

other batch information is obtainable could be considered as a very conservative exclusion.

Table 6.1. Shared genes and gene families (Up-regulation). List of the gene families, and shared genes, that were commonly up-regulated in dogs and tame foxes. The number, and name, of the genes within each gene family are provided, with the corresponding log2fold-change values in brackets for each species. Within each family, single genes were charecterized as shared between dogs and tame foxes (bold), or as exclusively to each of the two groups. When more than one transcript for a specific gene was present, all the log2FC values are reported.

Gene Family	Group	N of genes	Gene name and log2FC value
Retinal G protein-coupled receptor	**Shared**	1	**RGR** (2.10 in dogs, 0.78 in tame foxes)
Cholinergic receptor nicotinic alpha	**Shared**	1	**CHRNA5** (1.1 in dogs, 0.4 in tame foxes)
Squalene epoxidase	**Shared**	1	**SQLE** (0.54 in dogs, 0.31 in tame foxes)
Rho GTPase activating protein	**Shared**	1	**ARHGAP25** (0.86 in dogs, 0.72 in tame foxes)
	Tame fox	2	ARHGAP4 (0.64); ARHGAP30 (0.57)
Integrin alpha subunits	Dog	3	ITGA6 (1.25, 1.24); ITGA8 (1.14, 0.90); ITGAX (0.97)
	Tame fox	1	ITGAL (0.73)
	Shared	1	**ITGA7** (0.76 in dogs, 0.46 and 0.49 in tame foxes)
Myosin	Dog	1	MYO3A (1.12)
	Tame fox	3	MYOZ1 (1.53); MYO1F (0.93); MYO1C (0.47)
	Shared	1	**MYO7A** (0.82 in dogs; 0.41 in tame foxes)
Tribbles pseudokinase	Tame fox	2	TRIB1 (0.94); TRIB3 (0.78)
	Shared	1	**TRIB2** (0.61 in dogs; 0.2 in tame foxes)
EF hand calcium binding	Dog	1	EFCAB1 (2.59)
	Tame fox	1	EFCAB2 (0.46)
Transcription factor	Dog	1	TCF23 (2.04)
	Tame fox	1	TCF19 (0.63)
Adhesion G protein-coupled receptors	Dog	1	ADGRG6 (1.45)
	Tame fox	1	ADGRG1 (0.57)
Patatin Like Phospholipase Domain	Dog	1	PNPLA4 (1.41)
	Tame fox	1	PNPLA7 (0.59)
SRY-box	Dog	1	SOX6 (1.26)
	Tame fox	2	SOX17(0.84); SOX10 (0.66)
Hyaluronan and proteoglycan link protein	Dog	1	HAPLN1 (1.15)
	Tame fox	1	HAPLN3 (0.70)
Serine/threonine kinase	Dog	2	STK17A (1.15, 1.14); STK32A (1.10)
	Tame fox	1	STK40 (0.57)
Potassium channels	Dog	1	KCTD16 (0.98)
	Tame fox	1	KCTD15 (0.72)
Podocalyxin like	Dog	1	PODXL (0.95, 0.84)
	Tame fox	1	PODXL2 (0.70, 0.69, 0.67)
ATP binding cassette subfamily B	Dog	1	ABCB1 (0.93)
	Tame fox	1	ABCB9 (0.52)
Zinc finger DHHC-type	Dog	1	ZDHHC15 (0.75)
	Tame fox	1	ZDHHC1 (0.70)
Sushi domain	Dog	1	SUSD1 (0.68)
	Tame fox	2	SUSD3 (0.79); SUSD6 (0.47)
TBC1 domain family	Dog	1	TBC1D5 (0.54)
	Tame fox	1	TBC1D7 (0.27)
Mitogen-activated protein kinase kinase kinases	Dog	1	MAP3K5 (0.51)
	Tame fox	1	MAP3K11 (0.76)

In addition, we also found three gene families, containing four genes, simultaneously down-regulated in both groups (**Table 6.2**). Two of these genes (STMND1 and OASL) were shared between dogs and tame foxes, while the other two were unique to each group. When batch effects based on tissue were taken into account, STMND1 and OASL were maintained as being commonly down-regulated (Supplementary Figure 6.6).

Table 6.2. Shared genes and gene families (Down-regulation). List of the gene families, and shared genes, that were commonly down-regulated in dogs and tame foxes. The number, and name, of the genes within each gene family are provided, with the corresponding log2fold-change values in brackets for each species. Within each family, single genes were characterized as shared between dogs and tame foxes, or as exclusively to each of the two groups. When more than one transcript for a specific gene was present, all the log2FC values are reported.

Gene Family	Group	Number of UE	Gene name and log2FC value
Stathmin domain	**Shared**	1	**STMND1** (-1.18 in dogs, -0.53 in tame foxes)
Oligoadenylate synthetase like	**Shared**	1	**OASL** (-0.41 in dogs, -0.52 in tame foxes)
Heat shock protein family B	Dog	1	HSPB8 (-0.70)
	Tame fox	1	HSPB11 (-0.32)

The same analysis performed using the non-filtered datasets revealed similar results (Supplementary Table 6.10), although the RGR gene family which included a shared gene between dogs and tame foxes, was lost. This gene was not differentially expressed in the non-filtered fox dataset, representing an example of genes added as differentially expressed after filtering.

6.4. Discussion

Studies involving RNA-seq data often rely on the identification of one, few or many DETs in order to draw conclusions about biological pathways or about general transcriptome function and evolution. The explicit quantification of intra-condition count variation associated with such transcripts is important for maintaining the context of ambiguity that may exist following differential expression analysis. This is especially true given the growing ability to base highly informative studies around archived transcriptomics datasets at an inter-study level. Here, we developed a method that quantifies intra-condition variation for each individual transcript within the reference set and that can be used to explore the effects of identifying and removing reference-based transcripts harboring such variation above specified thresholds. By initially applying the method to extensive highly controlled simulated datasets harboring pre-defined levels of intra-

condition count variation we demonstrate the high effectiveness of DESEQ2 in identifying DETs, but also that it can be advantageous to reduce intra-condition variation within the count datasets in relation to identifying additional DETs that could have been overlooked without such filtering (**Figure 6.2**; Supplementary Figure 6.5; Supplementary Tables 6.4 and 6.5). By using highly controlled simulated datasets for initial testing, we also provide a tentative estimate on the limit of random intra-condition count variation above which the ability to reliably detect DETs is diminished (**Figure 6.2**; Supplementary Figure 6.5).

Our real data case study showed that, on average, nearly 20% of the transcripts identified as being differentially expressed prior to filtering contained levels of intra-condition variation equal to or above the 90th percentile value of the total distribution. This was higher than the relative proportion of such transcripts within the reference set and indicates that transcripts associated with higher intra-condition variation have a tendency to being identified as differentially expressed. When transcripts possess large amounts of such variation, some ambiguity in their identification as being differentially expressed is inevitable, since reliable expression patterns for at least one of the two conditions being compared have not been fully established; even if statistical correction is applied. This likely partially explains the level of discordance between various differential expression tools available (S. Li et al. 2014; Schurch et al. 2016; Quinn et al. 2018; Wang et al. 2019; Arora et al. 2020; Mou et al. 2020), for which no consensus on the best approach to apply exists (Stupnikov et al. 2021). However, more importantly, when such transcripts are used for drawing biological conclusion, the context of this uncertainty must be maintained.

We then explored the effects of removing transcripts associated with intra-condition variation, at varying threshold levels, on the gene-wise dispersion estimates used by DESEQ2. Within both case studies, as such transcripts were increasingly removed from input datasets prior to differential expression analysis, the correlation between the mean of normalized counts and dispersion estimates increased, and the number of outliers identified decreased (**Figures 6.6A**, **6.6B**; Supplementary Figure 6.7; Supplementary Table 6.9). This, along with discordances between the lists of DETs identified prior to and post filtering, suggests that transcripts were not simply removed because of physical exclusion from the input data, but that they were also removed, and added, as a result of the effects of removing intra-condition variation from the general gene-wise dispersions applied. The high rates of discordance we found, reaching 40% within the aggressive vs. tame fox case study (**Figure 6.7**; Supplementary Table 6.8), reveal how dependent the identification of DETs is on the accuracy of gene-wise dispersion

estimates used; these in turn being affected by transcripts associated with high intra-condition count variation.

High intra-condition count variation at an inter, and to a lesser extent intra, study level can arise from a range of sources including i) biological differences between samples such as age, sex, diet, and health; ii) *in silica* error involving assembly tools producing poorly understood chimeras within the reference transcriptome (Hsieh et al. 2019; Linheiro and Archer 2021; Linheiro and Archer 2022); iii) ambiguities in read mapping to such references (Reinert et al. 2015); iv) normalization of count data derived from such mapped reads (Evans et al. 2018); and v) including *in vitro* error during library preparation protocols (Brodin et al. 2013; Ma et al. 2019). Although we used DESEQ2 within our study, the results of our exploration on the effects of intra-condition variation in the detection of DETs likely applies to other software used for differential expression analysis that rely on per transcript count information across replicates for the estimation of transcript abundance and dispersion, for example, EdgeR (Robinson et al. 2010), BBSeq (Zhou et al. 2011), DSS (Wu et al. 2013), baySeq (Hardcastle and Kelly 2010) and ShrinkBayes (Van De Wiel et al. 2013).

Following the removal of the 3% and 5% of transcripts associated with the highest levels of variation between wolves and dogs, and aggressive and tame foxes, respectively, we observed an increase in the number of DETs. This pattern is similar to what we observed within our extensive iterative differential expression analysis experiments on simulated data where the levels of intra-condition variation, as well as sets of count over-represented transcripts, were explicitly controlled. Thus, this result suggests that for our case studies the removal of variation at these levels optimized the detection of differentially expressed transcripts whilst maintaining consistency. Using these 3% and 5% cut-offs, amongst the 50 over expressed genes identified, across the 21 shared gene families, seven genes were shared between dogs and tame foxes (**Table 6.1**). Of these seven genes, three main functions related to brain development, neurotransmission, and immune response were identified. These functions have been repeatedly associated with behavior selection during domestication by different approaches, such as QTL analysis (Albert et al. 2009; Kukekova et al. 2011a; Wirén et al. 2013), whole-genome sequencing (Carneiro et al. 2014; Freedman et al. 2016; Kukekova et al. 2018), and RNA data both using microarrays and RNA-seq (Saetre et al. 2004; Kukekova et al. 2011b; Heyne et al. 2014; Y. Li et al. 2014; Wang et al. 2018).

Up until recently, almost no gene overlap had been observed between gene expression profiles involving pairs of domesticated and wild animals (Albert et al. 2012). However, a recently published paper performing population genomic and brain transcriptional comparisons in seven bird and mammal domesticated species has revealed a strong convergent pattern in genes implicated in neurotransmission and neuroplasticity (Hou et al. 2020). These functions are compatible with those found in our analysis. The shared gene ITGA7 belongs to a gene family that is known to play an essential role in the control of neuronal connectivity (Lilja and Ivaska 2018) and the inflammatory response (González-Amaro and Sánchez-Madrid 1999). Other genes from this family, for example, ITGA8, have been previously observed to be over expressed in the brain of tame foxes (Kukekova et al. 2011b), and here we also observed its over expression in dogs providing further evidence of the possible family's role in tameness. Similar functions are associated with the shared genes CHRNA5 (Zhang et al. 2009; Winterer et al. 2010) and TRIB2 (Eyers et al. 2017) from the cholinergic and tribbles family, respectively. Additionally, we found a shared gene involved in sensing local environmental stimuli, the MYO7A, whose mutation results in loss of hearing and vision (Miller et al. 2012). Amongst the three gene families identified as under expressed (**Table 6.2**), we found the shared gene STMND1, which deficiency in the amygdala of mice was connected to a deficiency in innate and learned fear (Martel et al. 2008), a behavior that speculatively could also have an important role in domestication. Although we are aware that this overlap analysis between genes that show the same direction of expression in both dogs and tame foxes is not a formal test for gene convergence, we identified genes involved in several functions previously validated in the scope of domestication.

In this work, we have presented TVscript, a new tool that identifies and removes transcripts associated with high levels of intra-condition variation from RNA-seq count data prior to differential expression analysis. By applying TVscript to simulated data, as well as to real data derived from brain samples of dogs, wolves, tame and aggressive foxes, we demonstrate that as hyper variable transcripts are removed the ability to detect DETs increases in a robust and repeatable manner. Furthermore, we show that above a certain level of random intra-condition count variation, the identification of DETs is no longer viable. We propose that studies using RNA-seq data at an inter, or intra, study level should determine whether or not transcripts identified as being differentially expressed, using non-filtered reference sets, are still identified once filtering based on intra-condition count variation has been performed; regardless of the differential expression software used (or the method of obtaining initial counts). Discussion of such

transcripts can then be presented relative to the context of such filtering, thus taking a step forward in reducing the ambiguity surrounding intra-condition count variation. Such context is likely going to be dataset specific, as indicated between differences between our case studies, as the extent of intra-condition count variation will differ between datasets and will rarely be known as a prior to analysis. The latter is further highlighted by the consistent patterns observed during the iterative simulations that we performed where levels of intra-condition variation were pre-specified. Finally, by comparing the genes that were differentially expressed in the brain of dogs and tame foxes, we provided further tentative support for candidate genes involved with several functions long known for being involved with domestication. These genes, and functions, have potential for being involved with selection for tameness, which appears to have played a crucial role in canine domestication. We use the word tentative to describe our support, as the primary aim of this study was to investigate the effects of intra-condition count variation on the detection of DETs, and the identification of genes involved within an evolutionary process, such as domestication, should be supported by datasets specifically generated for that purpose, and confirmed relative to the different reference transcriptomes involved. The quality of such transcriptomes in turn, in relation to chimeras, missing transcripts and partial redundancies, must also be carefully explored.

Sections containing **data availability statement**, **author contributions**, **funding**, **acknowledgments,** and **conflict of interest** that appear in the published version can be accessed in Appendix II.

6.5. References

Albert FW, Carlborg Ö, Plyusnina I, Besnier F, Hedwig D, Lautenschläger S, Lorenz D, McIntosh J, Neumann C, Richter H, et al. 2009. Genetic architecture of tameness in a rat model of animal domestication. Genetics 182:541–554.

Albert FW, Somel M, Carneiro M, Aximu-Petri A, Halbwax M, Thalmann O, Blanco-Aguiar JA, Plyusnina IZ, Trut L, Villafuerte R, et al. 2012. A Comparison of Brain Gene Expression Levels in Domesticated and Wild Animals. PLoS Genet. 8:e1002962.

Arora S, Pattwell SS, Holland EC, Bolouri H. 2020. Variability in estimated gene expression among commonly used RNA-seq pipelines. Sci. Rep. 10:2734.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 2016 345 34:525–527.

Brodin J, Mild M, Hedskog C, Sherwood E, Leitner T, Andersson B, Albert J. 2013. PCR-Induced Transitions Are the Major Source of Error in Cleaned Ultra-Deep Pyrosequencing Data. PLoS One 8.

Bushnell, Brian. 2014. BBMap: A Fast, Accurate, Splice-Aware Aligner. In: Conference: 9th Annual Genomics of Energy & Environment Meeting.

Carneiro M, Rubin CJ, Palma F Di, Albert FW, Alföldi J, Barrio AM, Pielberg G, Rafati N, Sayyab S, Turner-Maier J, et al. 2014. Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. Science 345:1074–1079.

Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, et al. 2016. A survey of best practices for RNA-seq data analysis. Genome Biol. 13.

Costa-Silva J, Domingues D, Lopes FM. 2017b. RNA-Seq differential expression analysis: An extended review and a software tool. PLoS One 12:e019015.

Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W. 2005. BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. Bioinformatics 21:3439–3440.

Evans C, Hardin J, Stoebel DM. 2018. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Brief. Bioinform. 19:776–792.

Eyers PA, Keeshan K, Kannan N. 2017. Tribbles in the 21st Century: The Evolving Roles of Tribbles Pseudokinases in Biology and Disease. Trends Cell Biol. 27:284–298.

Freedman AH, Schweizer RM, Ortega-Del Vecchyo D, Han E, Davis BW, Gronau I, Silva PM, Galaverni M, Fan Z, Marx P, et al. 2016. Demographically-Based Evaluation of Genomic Regions under Selection in Domestic Dogs. PLoS Genet. 12:e100585.

Fushan AA, Turanov AA, Lee SG, Kim EB, Lobanov A V., Yim SH, Buffenstein R, Lee SR,

Chang KT, Rhee H, et al. 2015. Gene expression defines natural changes in mammalian lifespan. Aging Cell 14:352–365.

González-Amaro R, Sánchez-Madrid F. 1999. Cell adhesion molecules: selectins and integrins. Crit. Rev. Immunol. 19:389–429.

Hansen KD, Wu Z, Irizarry RA, Leek JT. 2011. Sequencing technology does not eliminate biological variability. Nat. Biotechnol. 29:572–573.

Hardcastle TJ, Kelly KA. 2010. BaySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics 11:422.

Hekman J, Johnson J, Edwards W, Vladimirova A, Gulevich R, Ford A, Kharlamova A, Herbeck Y, Acland G, Raetzman L, et al. 2018. Anterior Pituitary Transcriptome Suggests Differences in ACTH Release in Tame and Aggressive Foxes. G3; Genes|Genomes|Genetics 8:859–873.

Heyne HO, Lautenschläger S, Nelson R, Besnier F, Rotival M, Cagan A, Kozhemyakina R, Plyusnina IZ, Trut L, Carlborg Ö, et al. 2014. Genetic influences on brain gene expression in rats selected for tameness and aggression. Genetics 198:1277–1290.

Hoeppner MP, Lundquist A, Pirun M, Meadows JRS, Zamani N, Johnson J, Sundström G, Cook A, FitzGerald MG, Swofford R, et al. 2014. An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. PLoS One 9:91172.

Hou Y, Qi F, Bai X, Ren T, Shen X, Chu Q, Zhang X, Lu X. 2020. Genome-wide analysis reveals molecular convergence underlying domestication in 7 bird and mammals. BMC Genomics 21:1–20.

Hsieh PH, Oyang YJ, Chen CY. 2019. Effect of de novo transcriptome assembly on transcript quantification. Sci. Rep. 9.

Jeng SL, Chi YC, Ma MC, Chan SH, Sun HS. 2021. Gene expression analysis of combined RNA-seq experiments using a receiver operating characteristic calibrated procedure. Comput. Biol. Chem. 93:107515.

Jiang H, Wong WH. 2009. Statistical inferences for isoform expression in RNA-Seq. Bioinformatics 25:1026–1032.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. 2019 378 37:907–915.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14:1–13.

Kukekova A, Johnson J, Teiling C, Li L, Oskina I, Kharlamova A, Shikhevich S, Sun Q, Ponnala L, Temnykh S, et al. 2011b. Sequence comparison of prefrontal cortical brain transcriptome from a tame and an aggressive silver fox (Vulpes vulpes). BMC Genomics 12:482.

Kukekova A, Johnson J, Xiang XY, Feng SH, Liu S, Rando H, Kharlamova A, Herbeck Y, Natalya S, Xiong Z, et al. 2018. The red fox genome assembly identifies genomic regions associated with tame and aggressive behaviors. Nat. Ecol. Evol. 2:1479–1491.

Kukekova A, Trut L, Chase K, Kharlamova A, Johnson J, Temnykh S, Oskina I, Gulevich R, Vladimirova A, Klebanov S, et al. 2011a. Mapping loci for fox domestication: Deconstruction/Reconstruction of a behavioral phenotype. Behav. Genet. 41:593–606.

Kukurba KR, Montgomery SB. 2015. RNA Sequencing and Analysis. Cold Spring Harb. Protoc. 2015:951.

Lachmann A, Torre D, Keenan AB, Jagodnik KM, Lee HJ, Wang L, Silverstein MC, Ma'ayan A. 2018. Massive mining of publicly available RNA-seq data from human and mouse. Nat. Commun. 2018 91 9:1–10.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9:357–359.

Lee Sanghyuk, Li S, Seo CH, Lim B, Yang JO, Oh J, Kim M, Lee Sooncheol, Lee B, Kang C. 2011. Accurate quantification of transcriptome from RNA-Seq data by effective length normalization. Nucleic Acids Res. 39.

Li S, Labaj PP, Zumbo P, Sykacek P, Shi W, Shi L, Phan J, Wu PY, Wang M, Wang C, et al. 2014. Detecting and correcting systematic variation in large-scale RNA sequencing data. Nat. Biotechnol. 32:888–895.

Li Y, vonHoldt BM, Reynolds A, Boyko AR, Wayne RK, Wu DD, Zhang YP. 2013. Artificial selection on brain-expressed genes during the domestication of dog. Mol. Biol. Evol. 30:1867–1876.

Li Y, Wang GD, Wang MS, Irwin DM, Wu DD, Zhang YP. 2014. Domestication of the dog from the Wolf was promoted by enhanced excitatory synaptic plasticity: A hypothesis. Genome Biol. Evol. 6:3115–3121.

Lilja J, Ivaska J. 2018. Integrin activity in neuronal connectivity. J. Cell Sci. 131:jcs212.

Linheiro R, Archer J. CSReadGen website.

Linheiro R, Archer J. 2021. CStone: A de novo transcriptome assembler for short-read data that identifies non-chimeric contigs based on underlying graph structure. PLOS Comput. Biol. 17:e1009631.

Linheiro R, Archer J. 2022. Quantification of the effects of chimerism on read mapping, differential expression and annotation following short-read *de novo* assembly. F1000Research 11120 11:120.

Liu Y, Zhou J, White KP. 2014. RNA-seq differential expression studies: More sequence or more replication? Bioinformatics 30:301–304.

Lobo D, Linheiro R, Godinho R, Archer J. 2022a. On taming the effect of transcript level intra-condition count variation during differential expression analysis: a story of dogs, foxes and wolves: Example R Script for using DESeq2.

Lobo D, Linheiro R, Godinho R, Archer J. 2022b. On taming the effect of transcript level intra-condition count variation during differential expression analysis: a story of dogs, foxes and wolves: Bowtie2 counts and kallisto abundances.

Lord KA, Larson G, Coppinger RP, Karlsson EK. 2019. The History of Farm Foxes Undermines the Animal Domestication Syndrome. Trends Ecol. Evol. 35:125–136.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15:550.

Ma X, Shao Y, Tian L, Flasch DA, Mulder HL, Edmonson MN, Liu Y, Chen X, Newman S, Nakitandwe J, et al. 2019. Analysis of error profiles in deep next-generation sequencing data. Genome Biol. 20.

Martel G, Nishi A, Shumyatsky GP. 2008. Stathmin reveals dissociable roles of the basolateral amygdala in parental and social behaviors. Proc. Natl. Acad. Sci. 105:14620–14625.

McDermaid A, Monier B, Zhao J, Liu B, Ma Q. 2019. Interpretation of differential gene expression results of RNA-seq data: review and integration. Brief. Bioinform. 20:2044.

McIntyre LM, Lopiano KK, Morse AM, Amin V, Oberg AL, Young LJ, Nuzhdin S V. 2011. RNA-seq: Technical variability and sampling. BMC Genomics 12:293.

Miller KA, Williams LH, Rose E, Kuiper M, Dahl HHM, Manji SSM. 2012. Inner Ear Morphology Is Perturbed in Two Novel Mouse Models of Recessive Deafness. PLoS One 7:e512.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods 2008 57 5:621–628.

Mou T, Deng W, Gu F, Pawitan Y, Vu TN. 2020. Reproducibility of Methods to Detect Differentially Expressed Genes from Single-Cell RNA Sequencing. Front. Genet. 10.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320:1344–1349.

Perri AR, Feuerborn TR, Frantz LAF, Larson G, Malhi RS, Meltzer DJ, Witt KE. 2021. Dog domestication and the dual dispersal of people and dogs into the Americas. Proc. Natl. Acad. Sci. 118:e2010083118.

Quinn TP, Crowley TM, Richardson MF. 2018. Benchmarking differential expression analysis tools for RNA-Seq: normalization-based vs. log-ratio transformation-based methods. BMC Bioinformatics 19:274.

R Development Core Team. 2017. R: A language and environment for statistical computing. Vienna, Austria.

Rau A, Marot G, Jaffrézic F. 2014. Differential meta-analysis of RNA-seq data from multiple studies. BMC Bioinformatics 15:91.

Reinert K, Langmead B, Weese D, Evers DJ. 2015. Alignment of Next-Generation Sequencing Reads. Annu. Rev. Genomics Hum. Genet. 16:133–151.

Risso D, Schwartz K, Sherlock G, Dudoit S. 2011. GC-Content Normalization for RNA-Seq Data. BMC Bioinformatics 480.

Rizzi R, Beretta S, Patterson M, Pirola Y, Previtali M, Della Vedova G, Bonizzoni P. 2019. Overlap graphs and de Bruijn graphs: data structures for de novo genome assembly in the big data era. Quant. Biol. 7:278–292.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140.

Robles JA, Qureshi SE, Stephen SJ, Wilson SR, Burden CJ, Taylor JM. 2012. Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. BMC Genomics 13:484.

Roy M, Kim N, Kim K, Chung WH, Achawanantakun R, Sun Y, Wayne R. 2013. Analysis of the canine brain transcriptome with an emphasis on the hypothalamus and cerebral cortex. Mamm. Genome 24:484–499.

Saetre P, Lindberg J, Ellegren H, Vila C, Jazin E, Leonard JA, Olsson K, Pettersson U, Bergstrom TF. 2004. From wild wolf to domestic dog: Gene expression changes in the

brain. Mol. Brain Res. 126:198–206.

Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, Wrobel N, Gharbi K, Simpson GG, Owen-Hughes T, et al. 2016. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA 22:839–851.

Stegle O, Parts L, Durbin R, Winn J. 2010. A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput. Biol. 6:e100.

Stupnikov A, McInerney CE, Savage KI, McIntosh SA, Emmert-Streib F, Kennedy R, Salto-Tellez M, Prise KM, McArt DG. 2021. Robustness of differential gene expression analysis of RNA-seq. Comput. Struct. Biotechnol. J. 19:3470–3481.

Sudmant PH, Alexis MS, Burge CB. 2015. Meta-analysis of RNA-seq expression data across species, tissues and studies. Genome Biol. 16:1–11.

Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, Chen J. 2020. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol. 21:1–32.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28:511–515.

Tu J, Guo J, Li J, Gao S, Yao B, Lu Z. 2015. Systematic Characteristic Exploration of the Chimeras Generated in Multiple Displacement Amplification through Next Generation Sequencing Data Reanalysis. PLoS One 6:e0139857.

Wang T, Li B, Nelson CE, Nabavi S. 2019. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. BMC Bioinformatics 20:40.

Wang X, Pipes L, Trut L, Herbeck Y, Vladimirova A, Gulevich R, Kharlamova A, Johnson J, Acland G, Kukekova A, et al. 2018. Genomic responses to selection for tame/aggressive behaviors in the silver fox (Vulpes vulpes). Proc. Natl. Acad. Sci. 115:10398–10403.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: A revolutionary tool for transcriptomics. Nat. Rev. Genet. 10:57–63.

Wayne RK, Geffen E, Girman DJ, Koepfli K-P, Lau LM, Marshall CR. 1997. Molecular Systematics of the Canidae. Syst. Biol. 46:622–653.

Van De Wiel MA, Leday GGR, Pardo L, Rue H, Van Der Vaart AW, Van Wieringen WN. 2013. Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. Biostatistics 14:113–128.

Winterer G, Mittelstrass K, Giegling I, Lamina C, Fehr C, Brenner H, Breitling LP, Nitz B, Raum E, Müller H, et al. 2010. Risk gene variants for nicotine dependence in the CHRNA5-CHRNA3-CHRNB4 cluster are associated with cognitive performance. Am. J. Med. Genet. Part B Neuropsychiatr. Genet. 153:1448–1458.

Wirén A, Wright D, Jensen P. 2013. Domestication-related variation in social preferences in chickens is affected by genotype on a growth QTL. Genes, Brain Behav. 12:330–337.

Wu H, Wang C, Wu Z. 2013. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. Biostatistics 14:232–243.

Xu Z, Asakawa S. 2019. Physiological RNA dynamics in RNA-Seq analysis. Brief. Bioinform. 20:1725–1733.

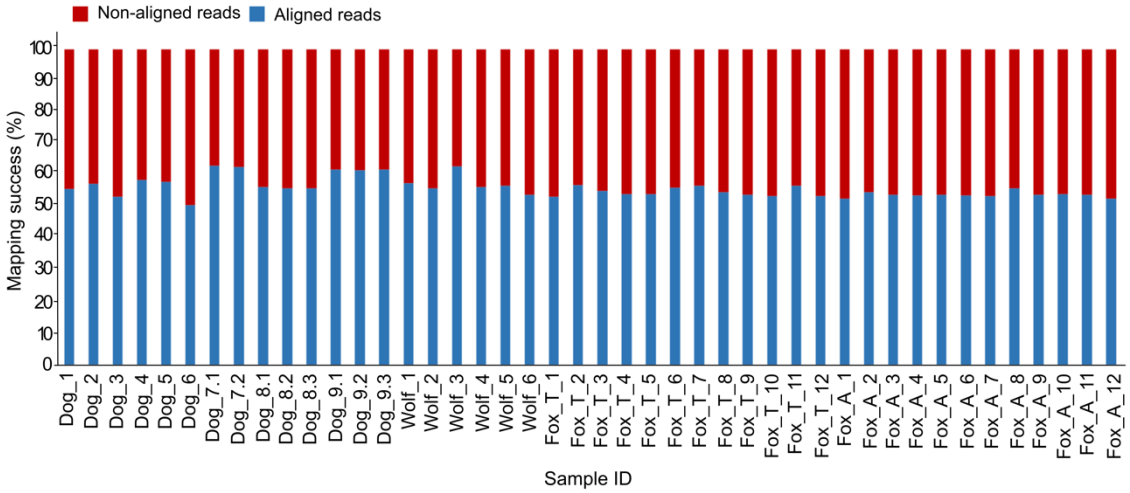
Yates AD, Achuthan P, Akanni W, Allen James, Allen Jamie, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, et al. 2020. Ensembl 2020. Nucleic Acids Res. 48:D682–D688.

Zhang H, Kranzler HR, Poling J, Gruen JR, Gelernter J. 2009. Cognitive flexibility is associated with KIBRA variant and modulated by recent tobacco use. Neuropsychopharmacology 34:2508–2516.

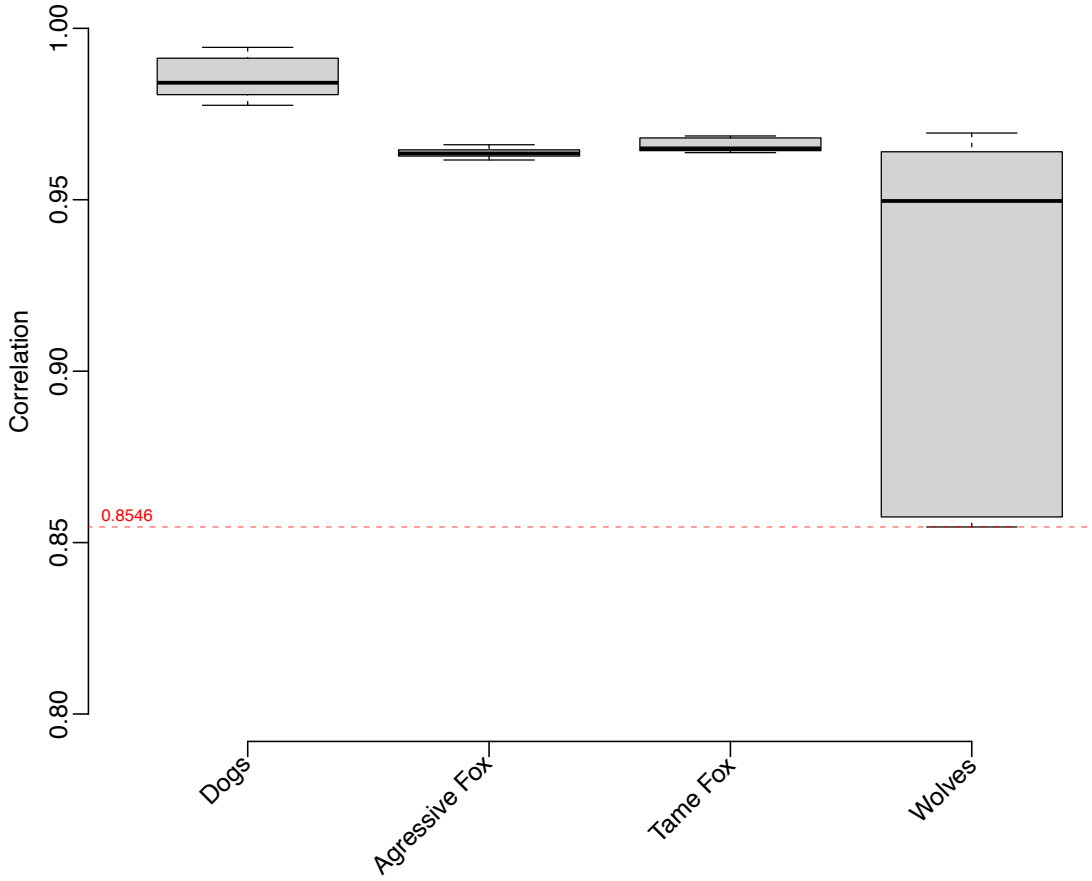
Zhou YH, Xia K, Wright FA. 2011. A powerful and flexible approach to the analysis of RNA sequence count data. Bioinformatics 27:2672–2678.

Zoabi Y, Shomron N. 2021. Processing and Analysis of RNA-seq Data from Public Resources. In: Methods in Molecular Biology. Vol. 2243. Methods Mol Biol. p. 81–94.

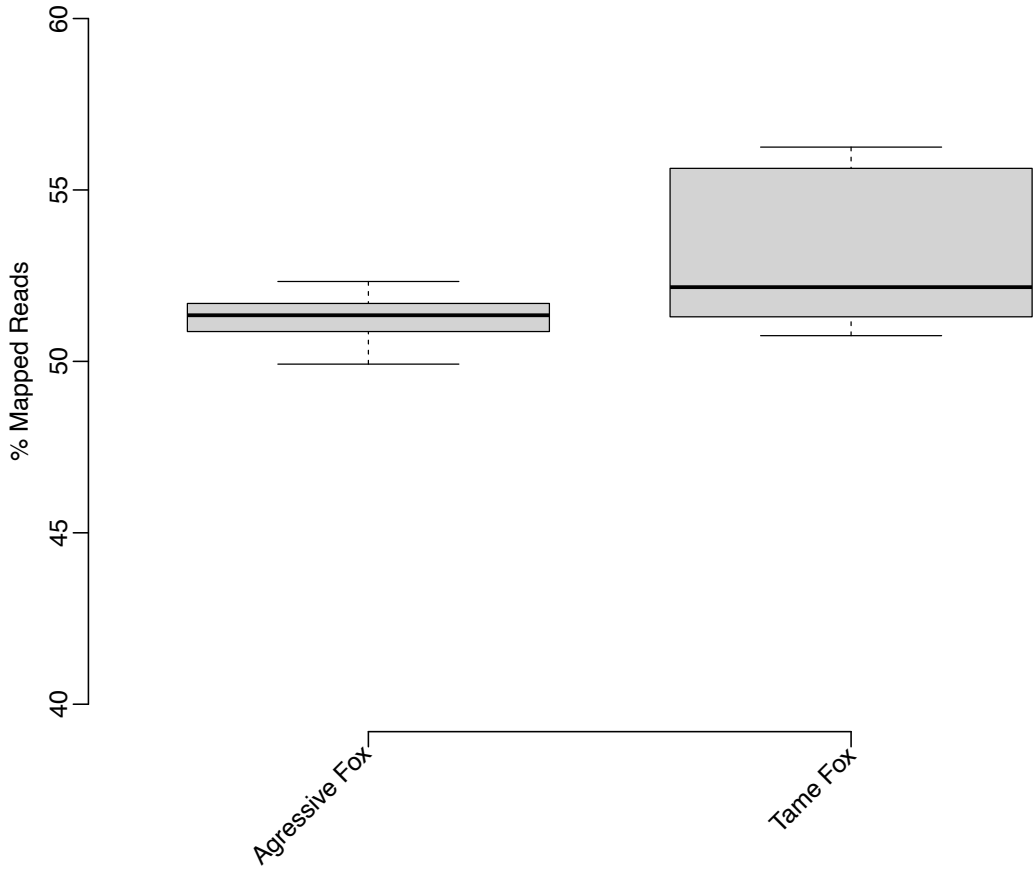
6.6. Supplementary Materials



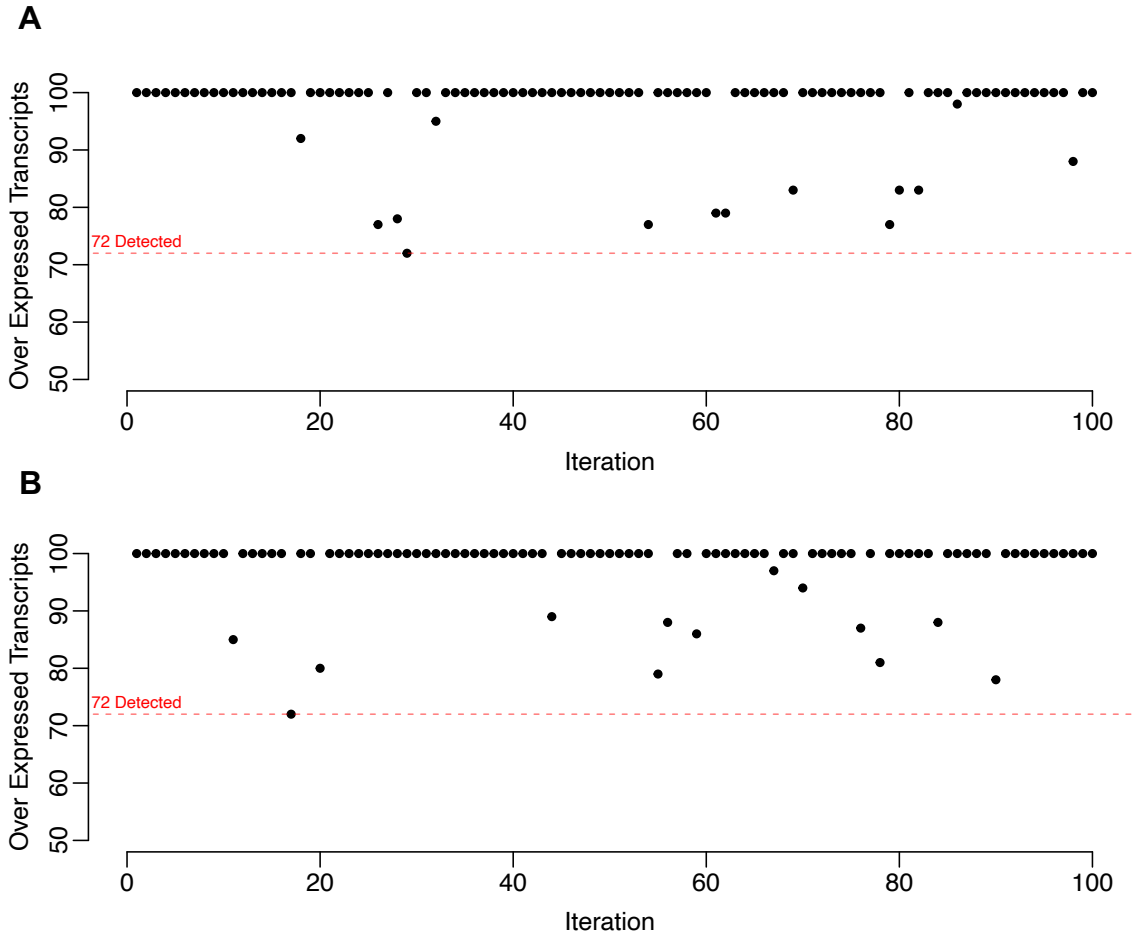
Supplementary Figure 6.1. Alignment rates obtained using BOWTIE2. Mapping success rates (%) resulting from the alignment of the 44 samples used in this study to the complete dog transcriptome. For each sample, the percentage of aligned reads is presented by the blue bars, while the percentage of reads failing to map is represented in red (the number of raw reads is available in Supplementary Table 6.1).



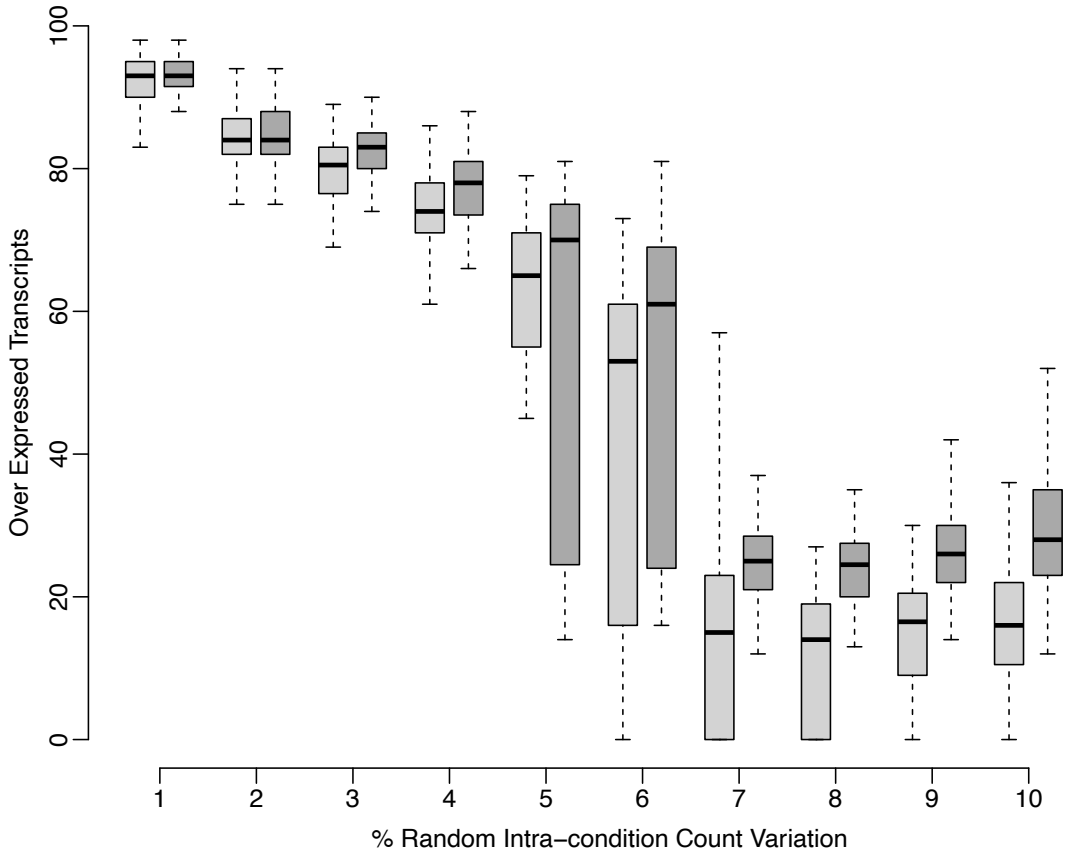
Supplementary Figure 6.2. Correlation between per-transcript counts obtained following BOWTIE2 mapping and count estimates obtained using KALLISTO. R^2 values describing the linear correlation between each count dataset produced from the mapped datasets presented in Supplementary Figure 6.1 and corresponding count estimates produced when pseudo-mapping the same RNA-Seq data to the complete dog transcriptome using KALLISTO.



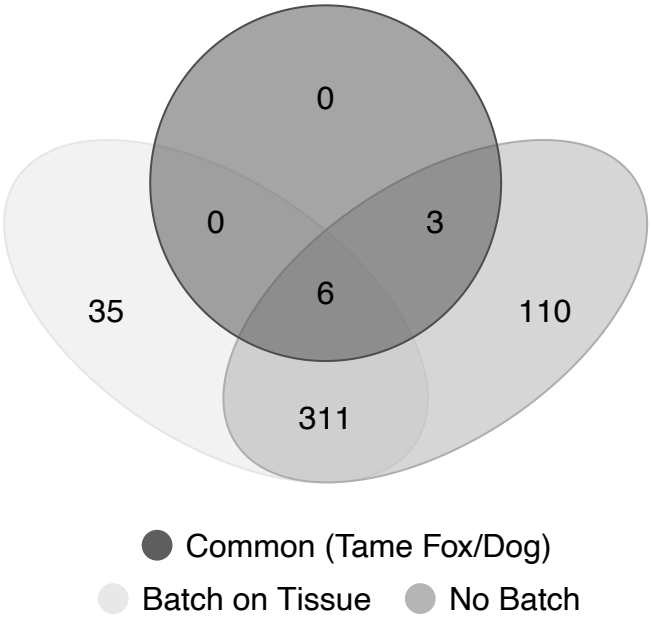
Supplementary Figure 6.3. Re-mapping fox data to the fox reference transcriptome. Read mapping rates achieved when mapping the fox RNA-seq datasets to the fox reference transcriptome.



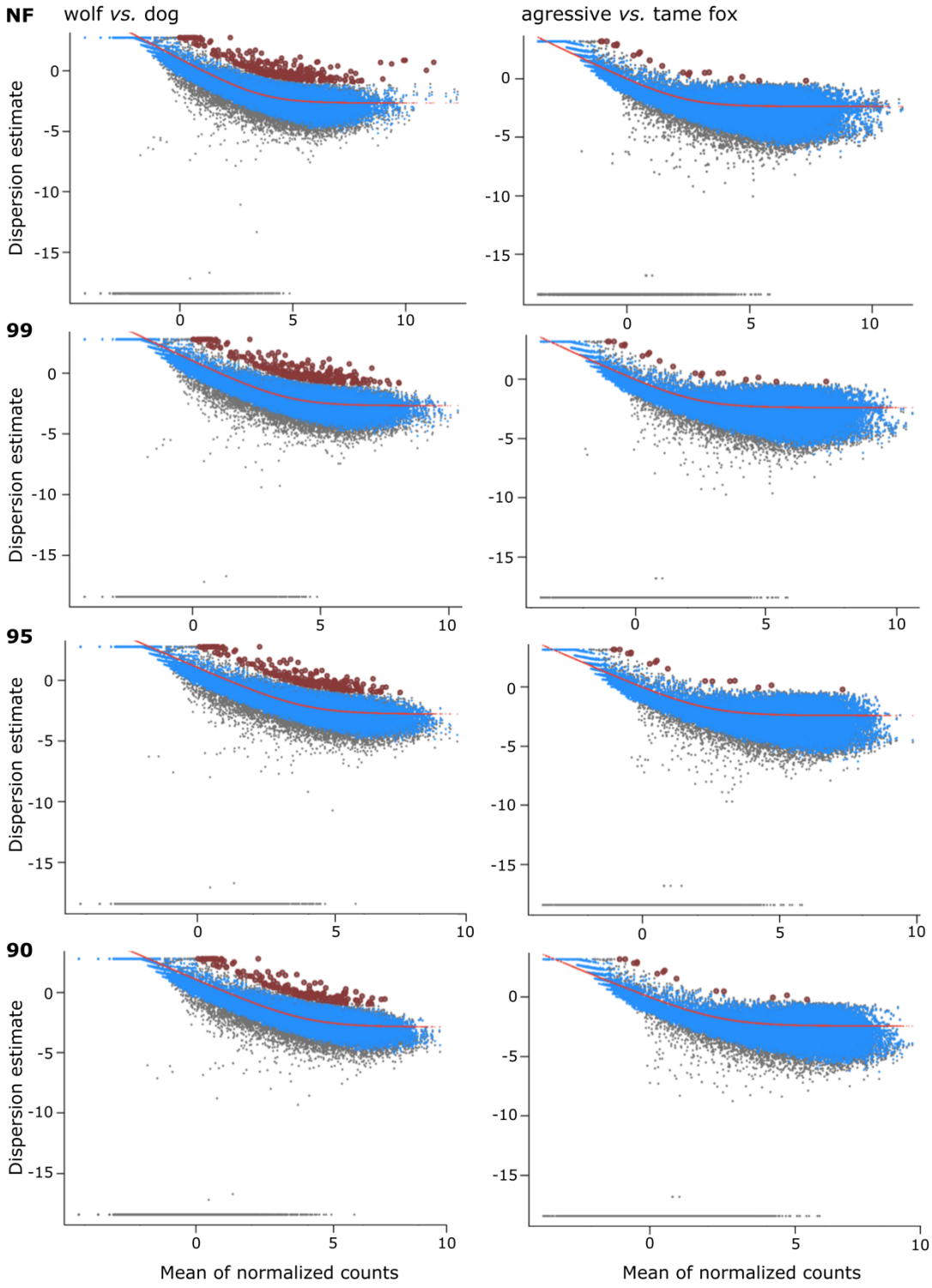
Supplementary Figure 6.4. Transcripts identified by DESEQ2 as being over expressed in the absence of randomly introduced intra-condition variation. Across 100 iterations the dots represent the number of transcripts identified as being over expressed between condition A and B. Each condition contained five replicates. (**A**) The 100 transcripts selected for read over-representation within replicates of condition B were maintained as constant and (**B**) the 100 transcripts selected for read over-representation within replicates of condition B were re-selected during each iteration. During each iteration the ten count datasets that were simulated each reflected even transcript coverage of 3 million read pairs with the exception of the 100 transcripts selected for over-representation in condition B whose count values were increase by a factor of two.



Supplementary Figure 6.5. Over expressed transcripts pre- and post-filtering (transcripts selected for count over-representation were re-selected during each iteration). The number of transcripts identified by DESEQ2 as being over expressed both prior to (light gray) and post (dark gray) filtering within each of the 100 iterations performed at each level of introduced random intra-condition count variation. Each iteration involved initially simulating ten count datasets divided into conditions A and B following which DESEQ2 was run to attempt to identify the 100 transcripts selected for over representation as described in the methods. Following this the ten simulated datasets were filtered using TVScript with a 95th percentile threshold in order to generate corresponding filtered datasets (divided into corresponding conditions A' and B') on which DESEQ2 was re-run.



Supplementary Figure 6.6. Confirmation of shared genes within differential expression analysis taking tissue effects into account. The upper dark gray circle contains the nine genes identified as being either commonly up, or down, regulated within dogs and tame foxes using the filter levels 95th and 97th percentiles whilst only accounting for condition (wolves vs. dogs and aggressive vs. tame fox). Six of these genes (RGR, CHRNA5, MYO7A, TRIB2, STMND1 and OASL) are present when DESEQ2 is run whilst also accounting for differences in tissue (light gray left oval). SQLE, ARHGAP25 and ITGA7 are observed only within the differentially expressed transcript list that is based solely on condition (dark gray right oval).



Supplementary Figure 6.7. Distribution of dispersion estimates. Plots of dispersion estimates in relation to the mean of normalized counts for both case studies, wolves and dogs (left panels), and tame and aggressive foxes (right panels). Estimates were calculated using DESEQ2 for the non-filtered (NF) and all filtered datasets (99^{th}, 95^{th} and 90^{th} are shown as an example). Gray dots represent the gene-wise maximum likelihood estimates (MLE), the red curve shows the fit to the MLEs, and blue dots identify the final maximum *a posteriori* (MAP) estimates of dispersion. Red dots represent the outliers detected by DESEQ2. Both x and y-axis are transformed into a logarithm scale.

Supplementary Table 6.1. **Dataset description.** Full details of all datasets used in this study, including the location of the relative tissue, age, and sex of each individual, replicate information and sequencing details.

Species	Study	Sample ID	No. individuals	Tissue	Compartment	Sex	Age	Condition	Library	Sequencing platform	Insert size (bp)	N of raw reads	Project accession	Run accession
Canis familiaris	Albert et al 2012	Dog_1	single	brain	FC	F	old AD	domesticated	PE	Illumina GA II	51	43 363 388	PRJEB3197	ERR266355
		Dog_2				M						34 735 876		ERR266386
		Dog_3				M						31 956 420		ERR266395
		Dog_4				F						41 264 506		ERR266403
		Dog_5				M						42 044 332		ERR266382
	Roy et al 2013	Dog_6	several	brain	CC	M	1-2 ya	domesticated	PE	Illumina HiSeq 2000	100	178 091 798	PRJEB4668	ERR351173
	Fushan et al 2015	Dog_7.1	single	brain	PFC / FL	M	young AD	domesticated	PE	Illumina HiSeq 2000	51	29 915 644	PRJNA185055	SRR636937
		Dog_7.2										28 668 648		SRR636938
	Hoeppner et al 2014	Dog_8.1	several	brain	NS	NS	NS	domesticated	PE	Illumina HiSeq 2000	580	18 561 120	PRJNA78827	SRR388737
		Dog_8.2										18 534 772		SRR388740
		Dog_8.3										18 488 968		SRR388766
		Dog_9.1	several	brain							450	40 195 676		SRR543733
		Dog_9.2										41 554 898		SRR536881
		Dog_9.3										39 271 576		SRR536883
Canis lupus	Albert et al 2012	Wolf_1	single	brain	FC	M	old AD	wild	PE	Illumina GA II	51	42 800 148	PRJEB3197	ERR266407
		Wolf_2				F						43 721 960		ERR266371
		Wolf_3				F						37 195 758		ERR266359
		Wolf_4				M						40 138 236		ERR266374
		Wolf_5				M						32 500 494		ERR266366
		Wolf_6				M						43 088 096		ERR266400
Vulpes vulpes	Wang et al 2018	Fox_T_1	single	brain	PFC	M	1.5 ya	tame	SE	Illumina HiSeq 2000	51	30 760 396	PRJNA307604	SRR3084300
		Fox_T_2										23 943 747		SRR3084299
		Fox_T_3										38 957 530		SRR3084298
		Fox_T_4										43 980 944		SRR3084297
		Fox_T_5										50 603 352		SRR3084296
		Fox_T_6										35 614 595		SRR3084295
		Fox_T_7										23 317 339		SRR3084294
		Fox_T_8										37 020 077		SRR3084293
		Fox_T_9										20 659 559		SRR3084292
		Fox_T_10										31 077 357		SRR3084291
		Fox_T_11										42 038 051		SRR3084290
		Fox_T_12										25 492 942		SRR3084289
		Fox_A_1						aggressive				33 541 449		SRR3084312

		Fox_A_2										40 582 881		SRR3084311
		Fox_A_3										43 558 186		SRR3084310
		Fox_A_4										32 082 055		SRR3084309
		Fox_A_5										47 274 581		SRR3084308
		Fox_A_6										28 570 574		SRR3084307
		Fox_A_7										35 859 502		SRR3084306
		Fox_A_8										25 896 719		SRR3084305
		Fox_A_9										45 247 750		SRR3084304
		Fox_A_10										45 763 357		SRR3084303
		Fox_A_11										27 253 195		SRR3084302
		Fox_A_12										18 588 067		SRR3084301

Abbreviation: AD, adult; CC, cerebral cortex; F, female; FC, frontal cortex; FL, frontal lobe; M, male; NS, not specified; PE, paired end; PFC, prefrontal cortex; SE, single end; ya, years old.

Supplementary Table 6.2. Common over expressed transcripts pre- and post-filtering (when transcripts selected for count over-representation are kept constant). The number of transcripts from the dog reference set that are commonly identified by DESEQ2 as being over expressed within condition B both prior to and post filtering for each of the 100 iterations performed at each level of introduced random intra-condition count variation. Each iteration involved simulating ten count datasets divided into conditions A and B following which DESEQ2 was run to attempt to identify the 100 transcripts selected for over-representation as described in the methods section. Filtering involved running TVScript with a 95th percentile threshold on the non-filtered datasets to generate corresponding filtered datasets (divided into corresponding conditions A' and B') following which DESEQ2 was re-run and the results compared back to those obtained for the non-filtered data.

	% Randomly Selected for Over-representation									
Iteration	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
rep_0	95	86	75	59	3	0	0	2	6	8
rep_1	97	85	79	66	70	0	0	13	5	6
rep_2	85	86	67	70	41	22	21	0	14	9
rep_3	83	85	78	60	20	32	0	12	9	5
rep_4	94	88	66	69	58	7	7	8	18	0
rep_5	91	84	71	74	6	14	19	9	14	12
rep_6	92	86	77	58	62	13	15	13	0	18
rep_7	91	80	75	19	34	37	0	0	2	0
rep_8	83	78	79	71	71	0	0	6	4	10
rep_9	85	83	79	68	51	0	0	9	3	10
rep_10	88	89	78	66	63	37	0	12	2	0
rep_11	85	87	72	69	46	42	18	4	2	0
rep_12	90	80	71	69	69	23	11	14	16	11
rep_13	94	85	83	17	16	2	0	11	4	11
rep_14	81	84	86	71	14	15	0	11	12	11
rep_15	78	85	74	66	67	0	14	0	8	8
rep_16	92	89	80	67	16	19	17	4	10	0
rep_17	89	77	74	76	12	40	0	3	9	14
rep_18	93	84	80	46	58	16	7	4	15	4
rep_19	90	88	74	19	0	11	0	9	0	16
rep_20	90	84	76	63	14	0	0	0	2	10
rep_21	89	84	78	68	18	33	0	12	10	11
rep_22	81	80	72	71	51	0	0	11	0	14
rep_23	86	85	80	71	68	5	10	2	9	11
rep_24	86	81	84	18	37	1	0	0	0	9
rep_25	91	85	66	66	55	40	0	9	13	0
rep_26	91	89	74	69	61	0	9	11	0	10
rep_27	87	80	72	71	43	54	0	0	11	2
rep_28	89	89	69	67	63	17	6	13	7	4
rep_29	93	86	70	78	58	0	0	0	16	11
rep_30	96	90	69	74	61	12	3	0	16	0
rep_31	87	91	83	75	57	0	10	10	8	9
rep_32	79	79	76	77	13	0	10	0	4	9
rep_33	96	81	73	15	9	13	0	13	8	10
rep_34	93	79	71	79	63	12	8	5	10	7
rep_35	82	82	82	76	46	0	0	0	1	5
rep_36	91	81	79	79	17	0	0	13	11	0
rep_37	83	78	78	22	54	2	0	11	3	11
rep_38	91	88	81	20	51	46	0	0	12	6
rep_39	91	80	73	69	51	11	31	2	9	11
rep_40	86	87	72	54	34	20	0	16	5	0
rep_41	89	89	80	65	9	15	0	7	13	8
rep_42	85	84	81	74	11	11	0	6	16	12
rep_43	97	80	75	60	44	22	14	4	9	10
rep_44	79	87	74	56	54	0	17	12	0	1
rep_45	91	82	71	73	60	0	13	11	11	12
rep_46	84	84	71	63	49	40	4	6	9	11
rep_47	95	76	79	75	14	46	11	17	8	12
rep_48	86	87	67	57	19	0	9	15	13	10
rep_49	92	83	71	20	11	11	0	11	13	8
rep_50	84	79	82	75	9	12	10	3	6	10
rep_51	92	77	74	18	9	20	0	8	9	5
rep_52	93	83	83	63	46	9	8	0	8	3
rep_53	90	83	75	64	4	0	9	0	7	7
rep_54	96	79	72	69	55	0	0	0	1	8
rep_55	91	88	84	71	61	54	13	7	10	11
rep_56	76	87	74	58	17	50	7	14	3	0
rep_57	84	87	71	47	56	0	0	8	11	12

rep_58	84	84	80	61	21	62	0	4	8	9
rep_59	88	87	83	65	11	0	0	0	7	11
rep_60	81	81	80	58	63	28	2	7	5	6
rep_61	80	82	81	64	21	0	3	7	12	0
rep_62	86	90	82	15	0	0	0	9	6	14
rep_63	83	85	79	70	15	4	8	7	9	9
rep_64	92	87	70	60	64	11	16	8	5	7
rep_65	88	80	85	17	39	14	9	7	11	10
rep_66	87	86	70	76	14	14	0	1	8	8
rep_67	91	86	75	65	63	4	9	11	4	12
rep_68	91	86	73	60	52	2	10	16	0	6
rep_69	88	81	78	63	55	7	0	6	10	4
rep_70	92	80	67	61	50	11	6	9	7	4
rep_71	73	84	71	72	49	14	0	13	5	9
rep_72	89	86	79	75	61	0	4	5	9	13
rep_73	94	80	76	78	50	13	0	11	13	14
rep_74	87	75	71	52	45	13	11	6	0	8
rep_75	88	85	72	62	67	48	15	2	6	10
rep_76	95	86	74	67	54	0	8	0	0	13
rep_77	82	81	64	72	16	0	9	12	9	6
rep_78	94	79	77	47	15	0	10	11	7	12
rep_79	86	82	76	17	55	35	0	4	10	0
rep_80	86	76	77	63	11	0	0	8	12	14
rep_81	94	85	85	54	65	31	4	0	10	14
rep_82	87	87	75	55	53	14	0	12	10	6
rep_83	96	81	67	76	4	38	5	5	6	8
rep_84	81	84	77	66	66	12	12	0	0	3
rep_85	77	82	83	72	15	48	0	6	11	15
rep_86	88	89	65	72	57	14	0	8	2	7
rep_87	86	83	84	69	59	16	12	9	11	8
rep_88	91	84	77	60	45	38	0	16	2	13
rep_89	91	86	75	54	11	0	8	10	9	8
rep_90	92	87	73	62	18	49	3	0	12	12
rep_91	89	84	81	19	53	0	20	7	0	11
rep_92	90	83	72	68	15	36	0	15	6	6
rep_93	92	78	74	63	38	55	16	0	5	0
rep_94	77	83	83	56	2	0	16	7	9	2
rep_95	85	87	78	66	16	50	14	10	7	7
rep_96	89	86	74	79	53	45	5	9	0	10
rep_97	90	83	70	14	64	30	0	5	6	6
rep_98	90	80	72	17	47	2	0	13	7	12
rep_99	90	82	85	63	63	33	17	0	9	14

Supplementary Table 6.3. Common over expressed transcripts pre- and post-filtering (when transcripts selected for count over-representation are re-selected during each iteration). Same as Supplementary Table 6.2 but where the 100 transcripts selected for over-representation within condition B are re-selected during each iteration.

	% Randomly Selected for Over-representation									
Iteration	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
rep_0	91	73	78	65	46	0	17	11	11	8
rep_1	85	68	80	69	55	12	0	0	6	0
rep_2	96	83	68	53	14	13	0	7	9	14
rep_3	92	87	77	70	14	13	4	14	4	7
rep_4	90	75	74	72	63	56	13	0	6	0
rep_5	93	81	73	25	55	43	13	0	0	4
rep_6	89	94	74	75	15	11	16	0	3	9
rep_7	92	79	74	13	67	7	0	9	9	8
rep_8	93	79	74	66	67	0	7	11	6	9
rep_9	90	83	80	69	58	2	0	4	10	6
rep_10	89	77	74	59	12	0	0	0	1	14
rep_11	74	79	86	16	48	0	13	3	10	9
rep_12	94	81	72	65	49	14	0	14	9	5
rep_13	83	87	85	19	7	26	0	9	7	3
rep_14	73	70	74	60	56	47	1	8	8	2
rep_15	90	78	82	79	61	16	0	0	8	8
rep_16	92	81	80	63	60	14	0	0	11	10
rep_17	92	81	78	66	26	16	14	6	9	12
rep_18	83	86	72	58	57	15	15	12	0	2
rep_19	92	83	72	65	58	32	0	7	1	0
rep_20	82	83	81	60	62	40	5	0	4	6
rep_21	92	79	67	74	15	10	0	12	2	10
rep_22	92	88	77	64	26	0	13	12	8	0
rep_23	85	87	72	63	14	46	13	15	11	5
rep_24	79	78	78	67	7	50	0	10	7	0
rep_25	77	81	78	14	21	15	10	9	2	8
rep_26	80	83	74	73	52	15	13	5	0	6
rep_27	98	81	81	73	3	24	15	0	16	2
rep_28	89	88	72	76	59	15	5	3	0	0
rep_29	90	86	75	11	45	14	7	11	16	0
rep_30	83	84	73	62	0	12	17	7	8	8
rep_31	82	86	70	58	13	21	8	5	9	4
rep_32	93	88	76	63	18	0	0	4	0	0
rep_33	92	81	74	61	2	54	0	8	0	1
rep_34	90	84	76	69	8	13	14	3	2	5
rep_35	83	70	77	72	53	34	0	1	11	5
rep_36	82	90	88	69	64	16	3	8	8	10
rep_37	84	83	74	59	12	16	0	0	4	2
rep_38	92	86	73	78	11	14	10	6	9	4
rep_39	92	83	80	68	64	41	0	0	13	3
rep_40	91	81	71	77	50	17	10	18	4	5
rep_41	96	83	71	63	14	5	0	0	12	1
rep_42	94	87	69	54	15	17	0	19	8	8
rep_43	95	87	68	53	39	0	10	2	0	4
rep_44	84	85	76	22	39	54	3	0	9	5
rep_45	91	74	62	71	36	0	11	10	2	12
rep_46	94	69	80	61	12	9	7	14	12	5
rep_47	91	86	72	71	14	44	0	1	0	1
rep_48	92	83	77	67	40	37	0	9	8	3
rep_49	79	78	72	79	9	8	11	16	0	5
rep_50	88	87	66	11	12	35	0	6	11	4
rep_51	88	89	82	76	10	9	0	8	5	0
rep_52	91	89	79	67	6	6	11	7	13	7
rep_53	97	85	76	63	0	8	13	10	0	2
rep_54	89	83	61	56	49	17	9	5	0	0
rep_55	90	86	80	62	10	12	17	3	0	0
rep_56	91	83	16	70	3	8	12	0	0	11
rep_57	96	82	78	79	36	0	2	12	15	0
rep_58	95	83	77	64	57	47	0	6	3	3
rep_59	93	76	82	53	45	0	0	15	3	12
rep_60	94	83	78	65	1	37	10	6	7	8
rep_61	81	82	74	18	37	8	0	11	8	0
rep_62	76	80	79	61	73	0	5	4	7	21
rep_63	95	80	55	74	59	17	42	0	8	13
rep_64	91	83	75	69	0	4	47	0	7	5
rep_65	92	78	72	58	17	59	0	5	0	0
rep_66	88	85	72	69	47	14	9	5	0	13

rep_67	96	93	60	63	43	0	7	0	7	5
rep_68	89	82	73	62	59	11	10	0	11	3
rep_69	91	79	77	68	55	0	0	9	7	11
rep_70	93	82	83	69	14	39	9	0	7	11
rep_71	93	82	79	69	16	0	2	0	9	2
rep_72	81	81	69	80	0	54	40	0	1	0
rep_73	87	80	70	75	52	48	0	0	2	4
rep_74	88	81	75	68	65	38	5	0	10	3
rep_75	83	82	80	66	25	0	10	0	5	4
rep_76	92	80	82	63	56	37	0	0	0	2
rep_77	89	84	82	21	1	9	10	4	11	6
rep_78	89	77	78	19	12	0	9	0	8	0
rep_79	74	75	79	72	57	15	1	12	9	5
rep_80	90	89	82	18	53	53	18	15	11	0
rep_81	98	88	88	65	12	51	11	0	5	12
rep_82	91	81	78	67	13	59	0	0	1	0
rep_83	93	88	80	68	17	17	4	12	0	3
rep_84	88	80	69	58	38	46	7	14	4	9
rep_85	94	83	68	64	54	8	13	13	10	14
rep_86	94	85	76	63	66	60	0	0	0	12
rep_87	94	74	80	69	21	5	0	17	10	0
rep_88	94	82	65	51	52	0	0	4	9	7
rep_89	82	83	82	79	64	0	5	7	0	1
rep_90	92	88	82	66	9	50	0	13	3	6
rep_91	89	86	61	67	42	13	7	0	9	11
rep_92	97	79	82	67	57	0	3	0	5	3
rep_93	77	76	68	80	30	26	2	5	0	1
rep_94	93	82	65	13	77	9	0	0	4	9
rep_95	89	76	80	60	60	0	12	11	6	9
rep_96	72	78	69	74	17	0	0	1	7	3
rep_97	91	81	74	65	68	39	14	6	2	3
rep_98	82	85	78	71	53	14	5	6	5	0
rep_99	75	79	78	70	60	46	9	4	6	3

Supplementary Table 6.4. Ratio between the common number of over expressed transcripts pre- and post-filtering and the maximum number detected when transcripts selected for count over-representation are kept constant. Numbers in Supplementary Table 6.2 were divided by the maximum number of over expressed transcripts detected within each correspomding iteration, i.e., the maximum number detected using corresponding non-filtered and filtered datasets.

	% Randomly Selected for Over-representation									
Iteration	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
rep_0	0.99	0.99	0.93	0.88	0.14	NA	NA	0.20	0.38	0.32
rep_1	0.98	0.99	0.95	0.92	0.92	NA	NA	0.59	0.25	0.35
rep_2	0.94	0.99	0.94	0.90	0.69	0.33	0.55	0.00	0.54	0.41
rep_3	0.94	0.99	0.93	0.85	0.27	0.58	0.00	0.63	0.56	0.31
rep_4	0.99	0.99	0.96	0.96	0.78	0.50	0.41	0.42	0.62	NA
rep_5	0.98	0.94	0.93	0.95	0.46	0.22	0.83	0.56	0.74	0.48
rep_6	0.97	0.96	0.96	0.87	0.85	0.20	0.23	0.62	NA	0.67
rep_7	0.98	0.94	0.96	0.26	0.63	0.70	0.00	NA	0.20	NA
rep_8	0.94	0.96	0.92	0.95	0.91	NA	NA	0.40	0.22	0.42
rep_9	0.93	0.99	0.95	0.88	0.81	0.00	NA	0.56	0.38	0.53
rep_10	0.97	0.99	0.92	0.89	0.94	0.71	NA	0.50	0.13	NA
rep_11	0.94	0.99	0.90	0.83	0.78	0.67	0.69	0.27	0.15	NA
rep_12	0.98	0.98	0.93	0.87	0.91	0.50	0.73	0.47	0.46	0.48
rep_13	0.99	0.97	0.94	0.21	0.24	0.29	NA	0.55	0.33	0.61
rep_14	0.89	1.00	0.99	0.90	0.19	0.29	NA	0.55	0.52	0.44
rep_15	0.90	0.98	0.91	0.89	0.87	NA	0.64	NA	0.40	0.35
rep_16	0.97	0.97	0.96	0.88	0.25	0.28	0.85	0.29	0.63	NA
rep_17	0.96	0.94	0.95	0.94	0.19	0.69	NA	0.20	0.45	0.54
rep_18	0.96	0.99	0.94	0.78	0.85	0.70	0.44	0.27	0.44	0.27
rep_19	0.96	0.99	0.93	0.25	NA	0.17	NA	0.50	NA	0.62
rep_20	0.97	0.99	0.97	0.82	0.24	NA	NA	0.00	0.15	0.45
rep_21	0.93	0.97	0.95	0.91	0.25	0.65	NA	0.57	0.50	0.50
rep_22	0.91	0.95	0.90	0.91	0.76	NA	NA	0.48	NA	0.44
rep_23	0.95	0.97	0.98	0.90	0.87	0.38	0.23	0.20	0.53	0.52
rep_24	0.93	0.98	0.94	0.95	0.67	0.08	NA	NA	NA	0.39
rep_25	0.98	0.99	0.89	0.90	0.77	0.74	NA	0.50	0.48	0.00
rep_26	0.98	0.99	0.95	0.93	0.84	NA	0.16	0.58	0.00	0.48
rep_27	0.95	0.98	0.90	0.89	0.69	0.82	NA	NA	0.48	0.13
rep_28	0.98	1.00	0.90	0.94	0.89	0.24	0.40	0.65	0.54	0.29
rep_29	0.99	0.99	0.91	0.95	0.79	NA	0.00	NA	0.57	0.52
rep_30	0.99	1.00	0.95	0.95	0.87	0.24	0.20	NA	0.62	NA
rep_31	0.95	0.98	0.98	0.96	0.81	NA	0.63	0.48	0.44	0.43
rep_32	0.91	0.99	0.95	0.95	0.24	NA	0.53	NA	0.21	0.60
rep_33	1.00	0.99	0.94	0.20	0.60	0.25	NA	0.62	0.42	0.37
rep_34	0.99	0.96	0.92	0.96	0.84	0.75	0.44	0.19	0.36	0.33
rep_35	0.92	0.95	0.98	0.95	0.77	NA	NA	NA	0.17	0.24
rep_36	0.99	0.93	0.96	0.96	0.24	NA	NA	0.57	0.46	NA
rep_37	0.92	0.98	0.94	0.27	0.86	0.09	NA	0.61	0.33	0.38
rep_38	0.97	1.00	0.95	0.29	0.75	0.78	NA	0.00	0.55	0.30
rep_39	0.98	0.99	0.89	0.90	0.71	0.65	0.47	0.20	0.33	0.52
rep_40	0.97	1.00	0.96	0.87	0.61	0.43	NA	0.64	0.28	NA
rep_41	0.98	0.98	0.96	0.88	0.56	0.63	NA	0.50	0.57	0.44
rep_42	0.93	1.00	0.93	0.93	0.20	0.19	NA	0.35	0.59	0.44
rep_43	0.99	0.96	0.96	0.85	0.69	0.37	0.30	0.33	0.39	0.48
rep_44	0.89	0.99	0.95	0.78	0.81	NA	0.68	0.60	NA	0.07
rep_45	0.98	0.99	0.96	0.90	0.83	NA	0.57	0.52	0.42	0.57
rep_46	0.98	0.99	0.95	0.91	0.78	0.70	0.33	0.46	0.33	0.48
rep_47	0.99	0.94	0.96	0.95	0.21	0.75	0.73	0.74	0.50	0.40
rep_48	0.97	0.97	0.89	0.79	0.33	NA	0.45	0.71	0.54	0.48
rep_49	0.98	0.98	0.91	0.26	0.16	0.73	NA	0.55	0.50	0.38
rep_50	0.92	0.99	0.96	0.91	0.13	0.23	0.67	0.27	0.43	0.45
rep_51	0.99	0.92	0.93	0.22	0.18	0.37	NA	0.32	0.43	0.36
rep_52	0.98	0.95	0.97	0.93	0.75	0.53	0.44	NA	0.40	0.23
rep_53	0.98	0.99	0.97	0.86	0.24	NA	0.45	NA	0.54	0.29
rep_54	1.00	0.95	0.94	0.92	0.83	NA	0.00	NA	0.20	0.40
rep_55	0.97	0.99	0.97	0.91	0.87	0.79	0.62	0.33	0.53	0.41
rep_56	0.90	0.99	0.93	0.83	0.22	0.77	0.54	0.58	0.33	NA
rep_57	0.92	1.00	0.91	0.73	0.81	NA	NA	0.38	0.58	0.52
rep_58	0.92	0.99	0.91	0.81	0.28	0.83	0.00	0.31	0.57	0.41
rep_59	0.95	0.99	0.97	0.88	0.17	NA	NA	NA	0.35	0.41
rep_60	0.93	0.99	0.96	0.83	0.84	0.58	0.13	0.35	0.31	0.43
rep_61	0.89	0.96	0.93	0.88	0.29	0.00	0.33	0.33	0.63	NA
rep_62	0.89	0.99	1.00	0.21	NA	NA	NA	0.56	0.33	0.54
rep_63	0.97	0.98	1.00	0.93	0.21	0.21	0.47	0.39	0.45	0.50

rep_64	0.97	0.99	0.92	0.87	0.82	0.65	0.73	0.44	0.23	0.47
rep_65	0.98	0.99	0.94	0.81	0.71	0.30	0.60	0.37	0.50	0.45
rep_66	0.98	0.98	0.89	0.93	0.61	0.67	0.00	0.17	0.42	0.33
rep_67	0.98	0.99	0.94	0.89	0.91	0.20	0.45	0.65	0.25	0.46
rep_68	0.99	1.00	0.92	0.83	0.73	0.10	0.56	0.64	NA	0.33
rep_69	0.95	0.98	0.91	0.91	0.81	0.35	NA	0.33	0.50	0.25
rep_70	0.98	0.98	0.91	0.86	0.70	0.22	0.40	0.43	0.44	0.19
rep_71	0.90	0.99	0.92	0.97	0.84	0.74	NA	0.68	0.29	0.35
rep_72	0.96	0.98	0.94	0.95	0.77	NA	0.36	0.31	0.56	0.45
rep_73	0.98	0.99	0.94	0.95	0.76	0.25	NA	0.50	0.57	0.58
rep_74	0.92	0.93	0.96	0.81	0.79	0.36	0.61	0.40	0.00	0.32
rep_75	0.98	1.00	0.94	0.89	0.92	0.84	0.25	0.20	0.33	0.38
rep_76	0.99	0.96	0.94	0.92	0.78	NA	0.47	NA	NA	0.52
rep_77	0.89	0.98	0.86	0.90	0.25	0.00	0.53	0.60	0.60	0.46
rep_78	1.00	0.99	0.96	0.75	0.22	0.00	0.48	0.52	0.39	0.43
rep_79	0.89	0.96	0.96	0.23	0.75	0.66	NA	0.27	0.50	NA
rep_80	0.98	0.95	0.92	0.84	0.20	0.00	NA	0.53	0.48	0.48
rep_81	0.99	0.97	0.97	0.83	0.87	0.66	0.18	0.00	0.42	0.52
rep_82	0.98	0.99	0.94	0.86	0.84	0.22	NA	0.52	0.59	0.32
rep_83	0.99	0.99	0.94	0.93	0.21	0.66	0.14	0.21	0.38	0.35
rep_84	0.92	0.99	0.96	0.88	0.86	0.19	0.67	NA	NA	0.25
rep_85	0.91	0.98	0.98	0.95	0.23	0.83	NA	0.33	0.58	0.58
rep_86	0.97	1.00	0.89	0.91	0.84	0.24	NA	0.42	0.14	0.50
rep_87	0.98	0.98	0.94	0.90	0.80	0.84	0.75	0.50	0.69	0.36
rep_88	0.99	0.98	0.96	0.86	0.73	0.70	NA	0.73	0.17	0.59
rep_89	0.97	0.98	0.95	0.86	0.17	NA	0.53	0.48	0.45	0.44
rep_90	0.98	0.96	0.95	0.87	0.25	0.86	0.30	0.00	0.50	0.41
rep_91	0.98	1.00	0.99	0.28	0.79	NA	0.74	0.44	NA	0.39
rep_92	0.98	0.99	0.97	0.88	0.21	0.67	NA	0.68	0.32	0.33
rep_93	0.97	0.95	0.91	0.85	0.67	0.82	0.73	NA	0.33	NA
rep_94	0.91	0.97	0.99	0.84	0.12	NA	0.70	0.39	0.56	0.25
rep_95	0.93	1.00	0.98	0.92	0.24	0.75	0.24	0.48	0.35	0.37
rep_96	0.94	0.99	0.97	0.98	0.78	0.78	0.36	0.45	0.00	0.53
rep_97	0.97	0.99	0.91	0.20	0.91	0.57	NA	0.38	0.30	0.32
rep_98	0.96	0.99	0.95	0.23	0.75	0.17	NA	0.62	0.37	0.40
rep_99	0.97	0.96	0.98	0.88	0.83	0.69	0.77	NA	0.50	0.52

Supplementary Table 6.5. Ratio between the common number of over expressed transcripts pre- and post-filtering and the maximum number detected when transcripts selected for count over-representation are re-selected during each iteration. Numbers in Supplementary Table 6.3 were divided by the maximum number of over expressed transcripts detected within each correspomding iteration, i.e., the maximum number detected using corresponding non-filtered and filtered datasets.

	% Randomly Selected for Over-representation									
Iteration	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
rep_0	0.99	0.90	0.92	0.90	0.75	0.00	0.68	0.58	0.65	0.53
rep_1	0.96	0.89	0.98	0.92	0.75	0.20	0.00	NA	0.38	NA
rep_2	0.99	1.00	0.93	0.84	0.74	0.19	NA	0.37	0.41	0.61
rep_3	0.99	1.00	0.96	0.89	0.23	0.23	0.40	0.64	0.29	0.58
rep_4	0.97	0.95	0.93	0.90	0.83	0.82	0.62	NA	0.50	0.00
rep_5	0.99	0.99	0.94	0.86	0.85	0.72	0.65	NA	NA	0.40
rep_6	1.00	1.00	0.97	0.93	0.20	0.44	0.64	0.00	0.27	0.41
rep_7	0.99	0.99	0.97	0.17	0.93	0.30	NA	0.36	0.47	0.42
rep_8	0.98	0.99	0.91	0.93	0.92	NA	0.39	0.58	0.40	0.50
rep_9	0.99	0.97	0.95	0.88	0.82	0.09	NA	0.33	0.43	0.22
rep_10	0.98	0.99	0.91	0.88	0.18	NA	0.00	0.00	0.33	0.58
rep_11	0.88	0.93	1.00	0.21	0.80	NA	0.62	0.27	0.50	0.41
rep_12	0.98	0.99	0.99	0.92	0.79	0.20	NA	0.74	0.45	0.33
rep_13	0.92	0.99	0.98	0.27	0.11	0.59	NA	0.41	0.33	0.21
rep_14	0.88	0.85	0.97	0.90	0.89	0.78	0.08	0.50	0.44	0.15
rep_15	0.96	1.00	0.99	0.96	0.86	0.29	NA	NA	0.40	0.47
rep_16	1.00	1.00	0.98	0.88	0.83	0.25	0.00	NA	0.50	0.53
rep_17	0.96	0.99	0.96	0.88	0.57	0.62	0.61	0.38	0.43	0.44
rep_18	0.91	0.98	0.90	0.91	0.85	0.25	0.68	0.52	0.00	0.13
rep_19	0.97	0.98	0.88	0.84	0.87	0.65	NA	0.54	0.13	0.00
rep_20	0.91	0.99	0.94	0.88	0.93	0.70	0.31	NA	0.31	0.33
rep_21	0.97	0.92	0.93	0.96	0.21	0.17	0.00	0.67	0.18	0.40
rep_22	0.96	1.00	0.92	0.90	0.37	NA	0.22	0.50	0.40	NA
rep_23	0.94	0.99	0.96	0.86	0.20	0.75	0.59	0.79	0.65	0.23
rep_24	0.91	0.95	0.94	0.94	0.58	0.79	NA	0.56	0.44	0.00
rep_25	0.90	0.96	0.94	0.20	0.33	0.22	0.45	0.64	0.18	0.36
rep_26	0.93	0.98	0.93	0.92	0.81	0.29	0.22	0.38	NA	0.40
rep_27	1.00	0.98	0.94	0.95	0.15	0.86	0.75	NA	0.62	0.25
rep_28	0.97	1.00	0.92	0.95	0.81	0.28	0.38	0.30	NA	0.00
rep_29	0.97	1.00	0.96	0.17	0.76	0.21	0.35	0.50	0.59	0.00
rep_30	0.93	1.00	0.92	0.86	NA	0.22	0.65	0.58	0.36	0.42
rep_31	0.92	0.98	0.93	0.85	0.18	0.47	0.47	0.50	0.35	0.21
rep_32	1.00	0.98	0.95	0.88	0.25	NA	NA	0.31	0.00	0.00
rep_33	0.98	0.95	0.96	0.91	0.08	0.82	NA	0.50	0.00	0.08
rep_34	0.96	1.00	0.97	0.91	0.26	0.21	0.58	0.43	0.17	0.29
rep_35	0.92	0.95	0.94	0.92	0.79	0.71	NA	0.13	0.52	0.36
rep_36	0.87	1.00	0.99	0.93	0.91	0.62	0.27	0.33	0.47	0.56
rep_37	0.93	0.98	0.95	0.86	0.18	0.24	NA	NA	0.29	0.15
rep_38	0.96	0.99	0.89	0.95	0.18	0.22	0.40	0.32	0.45	0.25
rep_39	0.97	0.94	0.96	0.92	0.85	0.71	NA	NA	0.43	0.23
rep_40	0.98	0.95	0.93	0.96	0.79	0.27	0.17	0.72	0.50	0.31
rep_41	0.99	0.99	0.95	0.82	0.19	0.22	NA	NA	0.52	0.17
rep_42	0.99	0.99	0.95	0.78	0.21	0.68	NA	0.73	0.33	0.32
rep_43	0.97	0.98	0.89	0.85	0.71	NA	0.53	0.22	NA	0.22
rep_44	0.93	0.98	0.95	0.28	0.70	0.86	0.19	NA	0.36	0.26
rep_45	0.98	0.96	0.89	0.89	0.65	NA	0.58	0.59	0.17	0.55
rep_46	1.00	0.87	0.94	0.87	0.15	0.29	0.44	0.58	0.48	0.31
rep_47	0.99	0.99	0.90	0.95	0.67	0.69	NA	0.06	NA	0.08
rep_48	0.98	0.99	0.92	0.92	0.74	0.66	NA	0.50	0.44	0.27
rep_49	0.92	0.98	0.95	0.94	0.38	0.53	0.20	0.59	NA	0.31
rep_50	0.98	0.99	0.89	0.20	0.17	0.64	NA	0.43	0.48	0.14
rep_51	0.99	0.99	0.98	0.94	0.20	0.17	NA	0.40	0.29	0.00
rep_52	0.98	0.98	0.95	0.94	0.11	0.16	0.19	0.41	0.76	0.39
rep_53	1.00	0.98	0.97	0.89	NA	0.50	0.57	0.48	0.00	0.17
rep_54	0.97	0.98	0.88	0.82	0.75	0.71	0.47	0.33	0.00	NA
rep_55	0.98	0.97	0.95	0.86	0.15	0.23	0.28	0.20	NA	NA
rep_56	0.98	0.98	0.25	0.91	0.14	0.15	0.46	0.00	NA	0.50
rep_57	0.99	0.93	0.95	0.95	0.65	NA	0.29	0.60	0.60	0.00
rep_58	0.98	0.99	0.92	0.89	0.81	0.73	NA	0.40	0.18	0.20
rep_59	0.98	0.97	0.96	0.87	0.74	NA	NA	0.71	0.21	0.39
rep_60	0.98	0.98	0.94	0.88	0.05	0.67	0.19	0.29	0.30	0.44
rep_61	0.91	0.94	0.93	0.23	0.67	0.15	NA	0.48	0.73	0.00
rep_62	0.90	0.98	0.95	0.86	0.92	NA	0.36	0.31	0.35	0.58
rep_63	0.99	0.99	0.89	0.93	0.86	0.35	0.70	NA	0.47	0.48

rep_64	0.99	0.99	0.95	0.93	NA	0.33	0.77	NA	0.35	0.36
rep_65	0.99	0.98	0.89	0.82	0.26	0.82	NA	0.38	NA	NA
rep_66	0.95	0.96	0.91	0.93	0.78	0.24	0.53	0.42	NA	0.57
rep_67	1.00	0.99	0.85	0.90	0.73	0.00	0.18	0.00	0.44	0.45
rep_68	0.98	0.96	0.91	0.86	0.81	0.50	0.27	NA	0.50	0.14
rep_69	0.97	0.93	0.95	0.91	0.77	NA	NA	0.50	0.47	0.44
rep_70	0.98	0.96	0.93	0.90	0.22	0.72	0.56	NA	0.64	0.39
rep_71	0.99	0.99	0.96	0.93	0.25	0.00	0.15	NA	0.43	0.17
rep_72	0.87	0.99	0.88	0.99	NA	0.81	0.70	NA	0.11	NA
rep_73	0.94	0.98	0.92	0.93	0.83	0.79	NA	NA	0.29	0.22
rep_74	0.98	0.99	0.96	0.94	0.89	0.72	0.50	NA	0.38	0.20
rep_75	0.92	0.98	0.98	0.88	0.56	NA	0.53	NA	0.56	0.36
rep_76	0.97	0.99	0.99	0.85	0.88	0.73	0.00	NA	NA	0.08
rep_77	0.98	0.99	0.96	0.26	0.05	0.20	0.17	0.31	0.65	0.26
rep_78	0.99	1.00	0.94	0.23	0.23	NA	0.56	NA	0.38	NA
rep_79	0.88	0.99	0.94	0.92	0.88	0.27	0.08	0.75	0.38	0.29
rep_80	0.98	0.98	0.96	0.24	0.79	0.82	0.69	0.60	0.46	NA
rep_81	1.00	0.97	1.00	0.89	0.16	0.80	0.17	NA	0.31	0.48
rep_82	0.96	0.98	0.96	0.87	0.21	0.81	NA	0.00	0.17	NA
rep_83	0.96	1.00	0.98	0.91	0.23	0.25	0.27	0.57	0.00	0.38
rep_84	0.97	0.98	0.90	0.87	0.72	0.79	0.27	0.67	0.36	0.47
rep_85	0.99	0.97	0.94	0.82	0.78	0.14	0.54	0.57	0.36	0.58
rep_86	0.98	0.98	0.90	0.88	0.88	0.88	NA	NA	0.00	0.52
rep_87	1.00	0.97	0.94	0.88	0.36	0.25	NA	0.74	0.53	0.00
rep_88	0.98	0.99	0.87	0.85	0.79	NA	NA	0.31	0.53	0.29
rep_89	0.91	0.95	0.94	0.92	0.94	NA	0.28	0.37	NA	0.10
rep_90	0.98	0.99	0.98	0.90	0.16	0.79	NA	0.72	0.21	0.46
rep_91	0.98	0.98	0.85	0.89	0.75	0.24	0.54	NA	0.53	0.61
rep_92	1.00	0.96	0.98	0.92	0.85	0.00	0.25	NA	0.26	0.21
rep_93	0.89	0.90	0.96	0.94	0.64	0.58	0.13	0.29	NA	0.08
rep_94	0.97	1.00	0.90	0.19	0.99	0.64	NA	NA	0.36	0.35
rep_95	0.96	0.95	0.99	0.83	0.88	NA	0.20	0.52	0.50	0.39
rep_96	0.87	0.92	0.90	0.91	0.25	NA	0.00	0.10	0.47	0.20
rep_97	0.99	0.98	0.94	0.94	0.94	0.72	0.31	0.32	0.13	0.23
rep_98	0.92	0.99	0.99	0.92	0.83	0.22	0.42	0.43	0.28	NA
rep_99	0.86	0.98	0.95	0.92	0.86	0.75	0.53	0.24	0.35	0.19

Supplementary Table 6.6. Removal of intra-condition variation. Number of transcripts kept and removed from the reference in each case study, wolves and dogs, and aggressive and tame foxes, across the filtered levels used (from the 99th to the 70th percentile). The first ten percentiles were explored in greater detail in steps of one, while the remaining were performed in steps of 5.

	Wolves and dogs		**Aggressive and tame foxes**	
Percentile	Transcripts kept	Transcripts removed	Transcripts kept	Transcripts removed
99	25923	184	25872	235
98	25601	506	25428	679
97	25253	854	25018	1089
96	24916	1191	24585	1522
95	24574	1533	24167	1940
94	24237	1870	23767	2340
93	23921	2186	23372	2735
92	23611	2496	22978	3129
91	23293	2814	22578	3529
90	22973	3134	22219	3888
85	21399	4708	20407	5700
80	19867	6240	18723	7384
75	18361	7746	17128	8979
70	16958	9149	15661	10446

Supplementary Table 6.7. DETs in wolf vs. dog. Complete list of DETs in dogs when compared to wolves, identified using non-filtered datasets, and those that got removed (red) within the highest 10% of intra-condition variation, as well as those added (green) as differentially expressed across selected filtered datasets (97th, 95th, and 90th percentiles). The correspondent annotated gene ID (when available), log2FC values and p-values are provided.

Ranking	Transcript ID	Gene ID	log2 FC	p-adj value	Filter level
1	ENSCAFT00000049574	BCO1	3.68	1.51E-16	Not filtered
2	ENSCAFT00000030934	TKTL1	5.12	4.77E-10	Not filtered
3	ENSCAFT00000049123	CCDC190	2.23	4.90E-10	Not filtered
4	ENSCAFT00000029783	PIEZO2	2.12	7.21E-10	Not filtered
5	ENSCAFT00000010686	EFCAB1	2.60	2.47E-09	Not filtered
6	ENSCAFT00000047954	SCN11A	1.61	3.72E-08	Not filtered
7	ENSCAFT00000044099	PIEZO2	1.90	1.58E-07	Not filtered
8	ENSCAFT00000043213	SCN11A	1.46	4.19E-07	Not filtered
9	ENSCAFT00000022535	ERBB4	1.36	6.12E-07	Not filtered
10	ENSCAFT00000038874		-3.39	1.17E-06	Not filtered
11	ENSCAFT00000044516		-2.98	1.74E-06	Not filtered
12	ENSCAFT00000048021		-0.96	2.03E-06	Not filtered
13	**ENSCAFT00000036702**	**CRYZ**	**1.49**	**6.14E-06**	**95th**
14	**ENSCAFT00000035561**		**-8.10**	**6.14E-06**	**98th**
15	ENSCAFT00000036942		-3.76	8.95E-06	Not filtered
16	ENSCAFT00000048139	LYPD6	1.36	1.17E-05	Not filtered
17	ENSCAFT00000001452	COL14A1	1.79	1.19E-05	Not filtered
18	ENSCAFT00000025553	AP5Z1	-1.17	1.51E-05	Not filtered
19	ENSCAFT00000005798	BVES	1.28	2.34E-05	Not filtered
20	ENSCAFT00000027675	NSMCE1	-0.62	2.34E-05	Not filtered
21	ENSCAFT00000012020	TDRD12	1.38	2.95E-05	Not filtered
22	ENSCAFT00000021631		-3.27	3.40E-05	Not filtered
23	**ENSCAFT00000013788**	**PTK2B**	**-0.78**	**4.58E-05**	**97th**
24	ENSCAFT00000019357	CTSZ	-1.29	4.62E-05	Not filtered
25	ENSCAFT00000000233	CCDC68	1.98	5.43E-05	Not filtered
26	ENSCAFT00000010803	NLRP14	1.68	8.97E-05	Not filtered
27	**ENSCAFT00000001241**		**-0.98**	**1.04E-04**	**94th**
28	ENSCAFT00000048789	ACCS	0.69	1.05E-04	Not filtered
29	**ENSCAFT00000028679**	**TTR**	**5.73**	**1.16E-04**	**99th**
30	ENSCAFT00000010637	STK32A	1.11	1.39E-04	Not filtered
31	ENSCAFT00000049816	PODXL	0.96	1.39E-04	Not filtered
32	ENSCAFT00000048591	HAPLN1	1.16	1.57E-04	Not filtered
33	ENSCAFT00000022447		-2.96	1.77E-04	Not filtered
34	ENSCAFT00000012985	PLSCR4	1.27	2.00E-04	Not filtered
35	ENSCAFT00000032487	SLC44A5	1.45	2.33E-04	Not filtered
36	ENSCAFT00000045631		-2.92	2.33E-04	Not filtered
37	ENSCAFT00000007509	ITGA8	1.14	2.49E-04	Not filtered
38	ENSCAFT00000015081	SEMA3G	1.18	2.56E-04	Not filtered
39	ENSCAFT00000012064	ADM	-1.50	3.15E-04	Not filtered
40	ENSCAFT00000004810	NT5E	1.45	3.45E-04	Not filtered
41	ENSCAFT00000017267	SYNM	-0.41	3.68E-04	Not filtered
42	ENSCAFT00000043196		-0.83	3.68E-04	Not filtered
43	ENSCAFT00000004066	PLPPR1	1.01	3.79E-04	Not filtered
44	ENSCAFT00000006882	MYO3A	1.12	3.81E-04	Not filtered
45	ENSCAFT00000005420	FIGNL1	0.92	4.24E-04	Not filtered
46	ENSCAFT00000022445		-4.61	4.69E-04	Not filtered
47	ENSCAFT00000013736	SOX6	1.27	5.12E-04	Not filtered
48	ENSCAFT00000001407	COL11A2	1.23	5.12E-04	Not filtered
49	ENSCAFT00000031837	VCAM1	1.37	6.01E-04	Not filtered
50	ENSCAFT00000039500	POSTN	1.83	6.36E-04	Not filtered
51	ENSCAFT00000026335	MRI1	-1.33	7.30E-04	Not filtered
52	ENSCAFT00000022067	SERPINE1	-2.03	7.84E-04	Not filtered
53	ENSCAFT00000027824	MYH3	1.92	9.48E-04	Not filtered
54	ENSCAFT00000047995	PITPNC1	1.15	9.79E-04	Not filtered
55	ENSCAFT00000045799	TDRD12	1.12	9.79E-04	Not filtered
56	ENSCAFT00000020699	ITGA6	1.25	1.02E-03	Not filtered
57	**ENSCAFT00000015150**	**SPARCL1**	**0.81**	**1.08E-03**	**99th**
58	**ENSCAFT00000039007**	**CLU**	**1.00**	**1.21E-03**	**99th**
59	ENSCAFT00000046959		4.49	1.33E-03	Not filtered
60	ENSCAFT00000044997	SLC25A48	1.29	1.46E-03	Not filtered
61	ENSCAFT00000022158	SYTL5	1.11	1.46E-03	Not filtered
62	ENSCAFT00000006393	SPATA5	0.88	1.46E-03	Not filtered
63	ENSCAFT00000026672		-1.43	1.46E-03	Not filtered
64	**ENSCAFT00000011873**		**-1.66**	**1.46E-03**	**92nd**

65	ENSCAFT00000021042		-3.49	1.46E-03	Not filtered
66	**ENSCAFT00000038737**	**SPARCL1**	**0.94**	**1.48E-03**	**99th**
67	**ENSCAFT00000010134**	**CHL1**	**0.99**	**1.56E-03**	**91st**
68	ENSCAFT00000035851		-0.98	1.64E-03	Not filtered
69	ENSCAFT00000026736	ITGAX	0.97	1.76E-03	Not filtered
70	ENSCAFT00000022797	VSIR	1.29	1.85E-03	Not filtered
71	**ENSCAFT00000013347**	**CLU**	**0.86**	**1.85E-03**	**99th**
72	**ENSCAFT00000049861**	**HSPB8**	**-0.69**	**1.87E-03**	**90th**
73	ENSCAFT00000014739	CRHBP	1.13	2.23E-03	Not filtered
74	**ENSCAFT00000047619**	**CHL1**	**1.02**	**2.28E-03**	**92nd**
75	ENSCAFT00000003015	ATP10D	0.63	2.52E-03	Not filtered
76	ENSCAFT00000003270	PDGFRA	1.46	2.59E-03	Not filtered
77	ENSCAFT00000011166	RGS20	0.98	2.61E-03	Not filtered
78	ENSCAFT00000010726	SUCLG2	0.84	2.80E-03	Not filtered
79	ENSCAFT00000021625		-5.08	2.85E-03	Not filtered
80	ENSCAFT00000010945	FZD10	1.82	2.86E-03	Not filtered
81	ENSCAFT00000043504		-1.37	2.89E-03	Not filtered
82	ENSCAFT00000013975	GRIK1	0.87	2.97E-03	Not filtered
83	ENSCAFT00000046707		-1.10	3.19E-03	Not filtered
84	**ENSCAFT00000044854**	**CHL1**	**1.01**	**3.21E-03**	**92nd**
85	ENSCAFT00000037196	TLR1	1.23	3.23E-03	Not filtered
86	ENSCAFT00000015510	ACAD8	-0.68	3.33E-03	Not filtered
87	ENSCAFT00000049563	THEM4	1.10	3.34E-03	Not filtered
88	**ENSCAFT00000037353**	**SMIM3**	**-1.36**	**3.34E-03**	**92nd**
89	ENSCAFT00000002896	ABCB1	0.93	3.55E-03	Not filtered
90	ENSCAFT00000029252	RHBDL3	-1.33	3.55E-03	Not filtered
91	ENSCAFT00000019970	CD84	1.56	3.73E-03	Not filtered
92	ENSCAFT00000002169	PODXL	0.85	3.73E-03	Not filtered
93	ENSCAFT00000025758	FUT8	0.71	3.73E-03	Not filtered
94	ENSCAFT00000014585	ARSB	0.59	3.73E-03	Not filtered
95	ENSCAFT00000000458	ADGRG6	1.46	3.82E-03	Not filtered
96	ENSCAFT00000043004	GBA2	-0.53	3.85E-03	Not filtered
97	ENSCAFT00000049725		1.23	3.90E-03	Not filtered
98	ENSCAFT00000028525	DTNA	0.53	3.94E-03	Not filtered
99	ENSCAFT00000035852		-1.00	4.27E-03	Not filtered
100	ENSCAFT00000032396	ADGRL4	1.21	4.62E-03	Not filtered
101	ENSCAFT00000046656	CX3CR1	1.07	4.62E-03	Not filtered
102	ENSCAFT00000001409	COL11A2	1.12	4.64E-03	Not filtered
103	ENSCAFT00000036013	MST1R	-0.87	4.74E-03	Not filtered
104	ENSCAFT00000029594	TENM1	0.90	4.74E-03	Not filtered
105	ENSCAFT00000006096	RAB3C	0.90	4.78E-03	Not filtered
106	**ENSCAFT00000027126**	**NUPR1**	**-1.35**	**4.94E-03**	**93rd**
107	**ENSCAFT00000009052**	**P2RY2**	**-1.67**	**5.05E-03**	**92nd**
108	ENSCAFT00000048687	GRIK1	0.88	5.56E-03	Not filtered
109	ENSCAFT00000050122		-1.02	5.56E-03	Not filtered
110	**ENSCAFT00000019347**		**-1.03**	**5.56E-03**	**91st**
111	**ENSCAFT00000039881**		**-1.39**	**5.56E-03**	**98th**
112	ENSCAFT00000029596	TENM1	0.98	5.61E-03	Not filtered
113	ENSCAFT00000047013		1.54	5.64E-03	Not filtered
114	ENSCAFT00000020574	STPG1	1.33	5.64E-03	Not filtered
115	ENSCAFT00000005594	STK17A	1.16	5.64E-03	Not filtered
116	ENSCAFT00000021648		4.82	5.69E-03	Not filtered
117	ENSCAFT00000008732		-3.07	5.69E-03	Not filtered
118	ENSCAFT00000005257	CDCA8	-0.81	5.79E-03	Not filtered
119	ENSCAFT00000023962	KCNN1	-1.05	5.79E-03	Not filtered
120	ENSCAFT00000021514		-3.48	5.79E-03	Not filtered
121	ENSCAFT00000020206	EPS8	0.96	6.06E-03	Not filtered
122	ENSCAFT00000043005	CCDC3	0.83	6.06E-03	Not filtered
123	ENSCAFT00000031740	TLDC1	-0.83	6.06E-03	Not filtered
124	ENSCAFT00000048547	MRPL34	-1.02	6.06E-03	Not filtered
125	ENSCAFT00000012229	MPEG1	1.43	6.49E-03	Not filtered
126	ENSCAFT00000043964	JADE2	-0.56	6.62E-03	Not filtered
127	ENSCAFT00000026974	EME1	-0.95	6.62E-03	Not filtered
128	ENSCAFT00000015805	STMND1	-1.18	6.62E-03	Not filtered
129	ENSCAFT00000038122	OOEP	4.44	6.84E-03	Not filtered
130	ENSCAFT00000019796	CASQ1	2.04	7.05E-03	Not filtered
131	ENSCAFT00000021851		-0.68	7.05E-03	Not filtered
132	**ENSCAFT00000024572**	**SCG3**	**0.92**	**7.47E-03**	**98th**
133	ENSCAFT00000027202	SHE	0.69	7.53E-03	Not filtered
134	ENSCAFT00000049656	ABCG2	1.01	7.58E-03	Not filtered
135	ENSCAFT00000005878	KEL	-1.05	7.66E-03	Not filtered
136	ENSCAFT00000031659	CA5A	1.38	7.89E-03	Not filtered
137	ENSCAFT00000016783	MTTP	1.20	7.89E-03	Not filtered
138	ENSCAFT00000022236	AIFM2	-0.80	7.89E-03	Not filtered
139	ENSCAFT00000018222	PITPNC1	1.04	8.07E-03	Not filtered

140	**ENSCAFT00000018127**		**-0.70**	**8.07E-03**	**96th**
141	ENSCAFT00000000397	MAP3K5	0.51	8.08E-03	Not filtered
142	ENSCAFT00000013222	TM4SF18	1.08	8.34E-03	Not filtered
143	ENSCAFT00000000298	SLC18B1	0.83	8.38E-03	Not filtered
144	ENSCAFT00000049200	TRIB2	0.62	8.40E-03	Not filtered
145	ENSCAFT00000012150	CD180	1.26	9.10E-03	Not filtered
146	ENSCAFT00000049322	FEN1	-0.70	9.10E-03	Not filtered
147	ENSCAFT00000015494	ELOVL2	1.11	9.37E-03	Not filtered
148	ENSCAFT00000039489	SFRP4	-1.14	9.45E-03	Not filtered
149	ENSCAFT00000046208	SNAI3	-1.72	9.47E-03	Not filtered
150	ENSCAFT00000046049	ITGA6	1.26	9.59E-03	Not filtered
151	ENSCAFT00000001083	ANGPT1	1.07	9.64E-03	Not filtered
152	ENSCAFT00000008784	CLDN10	0.73	9.64E-03	Not filtered
153	ENSCAFT00000032339	CDH3	-0.90	9.88E-03	Not filtered
154	ENSCAFT00000008856	SUPT5H	-0.45	1.02E-02	Not filtered
155	ENSCAFT00000006034	SLC7A11	1.32	1.04E-02	Not filtered
156	ENSCAFT00000011237	RAB3C	0.80	1.08E-02	Not filtered
157	ENSCAFT00000014333	CALB1	2.48	1.10E-02	Not filtered
158	ENSCAFT00000010417	ELK3	0.68	1.10E-02	Not filtered
159	ENSCAFT00000008294	VHL	0.56	1.10E-02	Not filtered
160	ENSCAFT00000043101	EMP1	0.74	1.11E-02	Not filtered
161	ENSCAFT00000001042	ZBTB12	-0.70	1.11E-02	Not filtered
162	ENSCAFT00000006826	ATP6V0A4	1.48	1.12E-02	Not filtered
163	ENSCAFT00000046164		-3.07	1.14E-02	Not filtered
164	**ENSCAFT00000014467**	**BHMT**	**1.10**	**1.17E-02**	**95th**
165	ENSCAFT00000022939	MAOA	0.77	1.18E-02	Not filtered
166	ENSCAFT00000012133	DPY19L3	0.57	1.21E-02	Not filtered
167	ENSCAFT00000030449	PNMA6A	-0.99	1.21E-02	Not filtered
168	ENSCAFT00000010017	CCNA1	-0.67	1.21E-02	Not filtered
169	ENSCAFT00000045654	PPP2R1B	0.62	1.24E-02	Not filtered
170	ENSCAFT00000026017	SEL1L3	0.54	1.24E-02	Not filtered
171	ENSCAFT00000047154	CYYR1	1.02	1.25E-02	Not filtered
172	**ENSCAFT00000008624**	**EHD3**	**-0.52**	**1.27E-02**	**93rd**
173	ENSCAFT00000000346	TRIM36	0.69	1.27E-02	Not filtered
174	ENSCAFT00000032167	LRRC8C	0.68	1.34E-02	Not filtered
175	ENSCAFT00000016073	BMPR1B	1.31	1.35E-02	Not filtered
176	ENSCAFT00000028869	LAMA3	-1.60	1.38E-02	Not filtered
177	ENSCAFT00000024392	TNFSF10	1.07	1.39E-02	Not filtered
178	ENSCAFT00000027898	GDPD1	0.60	1.39E-02	Not filtered
179	**ENSCAFT00000013168**	**GPD1**	**-0.89**	**1.43E-02**	**94th**
180	ENSCAFT00000014475	BHMT2	1.11	1.43E-02	Not filtered
181	ENSCAFT00000043020	FNDC1	1.03	1.43E-02	Not filtered
182	ENSCAFT00000044792	DIO2	1.74	1.46E-02	Not filtered
183	**ENSCAFT00000031879**	**VAT1L**	**0.86**	**1.46E-02**	**95th**
184	ENSCAFT00000030320		-1.00	1.46E-02	Not filtered
185	ENSCAFT00000017406	NPNT	1.16	1.46E-02	Not filtered
186	ENSCAFT00000044761	BDKRB2	-2.62	1.46E-02	Not filtered
187	ENSCAFT00000039527	C3AR1	1.13	1.51E-02	Not filtered
188	ENSCAFT00000044109		-0.75	1.51E-02	Not filtered
189	ENSCAFT00000027222	GSTZ1	-0.85	1.51E-02	Not filtered
190	**ENSCAFT00000023993**	**QDPR**	**-1.16**	**1.51E-02**	**99th**
191	ENSCAFT00000049832	STK17A	1.14	1.51E-02	Not filtered
192	ENSCAFT00000010805	NLRP14	1.49	1.53E-02	Not filtered
193	**ENSCAFT00000021914**	**TMEM47**	**1.07**	**1.58E-02**	**95th**
194	ENSCAFT00000014470	ADAM28	0.95	1.58E-02	Not filtered
195	**ENSCAFT00000017013**	**ADD3**	**0.78**	**1.58E-02**	**94th**
196	ENSCAFT00000026818	SLC35B1	-0.43	1.58E-02	Not filtered
197	ENSCAFT00000032475	CES2	-1.15	1.58E-02	Not filtered
198	ENSCAFT00000045856	VSX2	-1.25	1.58E-02	Not filtered
199	ENSCAFT00000021879	CCR5	1.48	1.59E-02	Not filtered
200	**ENSCAFT00000008101**	**SLC6A1**	**0.58**	**1.61E-02**	**95th**
201	ENSCAFT00000046233	LTBP2	-2.26	1.67E-02	Not filtered
202	ENSCAFT00000006384	PIK3CG	0.99	1.68E-02	Not filtered
203	**ENSCAFT00000027435**		**-0.53**	**1.70E-02**	**91st**
204	ENSCAFT00000015321	ABCG2	0.86	1.70E-02	Not filtered
205	**ENSCAFT00000017616**		**-0.87**	**1.72E-02**	**93rd**
206	**ENSCAFT00000018912**	**PCBP3**	**-1.50**	**1.74E-02**	**98th**
207	ENSCAFT00000049390	MOV10L1	1.12	1.86E-02	Not filtered
208	ENSCAFT00000048622	SUN2	-0.48	1.87E-02	Not filtered
209	ENSCAFT00000012270	SMAD1	0.58	1.87E-02	Not filtered
210	ENSCAFT00000002736	CHRNA5	1.11	1.88E-02	Not filtered
211	ENSCAFT00000004046	COL19A1	0.90	1.88E-02	Not filtered
212	ENSCAFT00000008744	SLCO2B1	0.72	1.88E-02	Not filtered
213	ENSCAFT00000016720	PKNOX1	-0.69	1.88E-02	Not filtered

214	ENSCAFT00000048536	POLN	-0.86	1.88E-02	Not filtered
215	ENSCAFT00000036890		-3.58	1.88E-02	Not filtered
216	ENSCAFT00000048492	CENPW	1.21	1.90E-02	Not filtered
217	**ENSCAFT00000000986**		**-1.86**	**1.90E-02**	**99th**
218	ENSCAFT00000026830	LTBP2	-2.12	1.90E-02	Not filtered
219	ENSCAFT00000017178	SLC9B2	0.98	1.95E-02	Not filtered
220	ENSCAFT00000032225	DHX38	-0.35	1.95E-02	Not filtered
221	**ENSCAFT00000043488**	**CSPG5**	**0.72**	**1.97E-02**	**92nd**
222	ENSCAFT00000009440	TBC1D5	0.55	1.97E-02	Not filtered
223	ENSCAFT00000023694	PDE4A	-0.62	1.97E-02	Not filtered
224	ENSCAFT00000007481	FADS6	-0.74	2.01E-02	Not filtered
225	ENSCAFT00000001085	RSPO2	0.90	2.04E-02	Not filtered
226	**ENSCAFT00000000984**		**-1.80**	**2.12E-02**	**99th**
227	ENSCAFT00000043813	PGAP1	1.18	2.13E-02	Not filtered
228	ENSCAFT00000029073	PPP1R13B	-0.51	2.23E-02	Not filtered
229	ENSCAFT00000036549		0.71	2.24E-02	Not filtered
230	ENSCAFT00000045214	NHSL1	0.79	2.28E-02	Not filtered
231	ENSCAFT00000005411	CSF3R	-1.93	2.35E-02	Not filtered
232	ENSCAFT00000038735		0.67	2.36E-02	Not filtered
233	ENSCAFT00000012461	OCLN	1.32	2.37E-02	Not filtered
234	**ENSCAFT00000043311**		**-1.01**	**2.37E-02**	**99th**
235	**ENSCAFT00000000831**	**VIP**	**0.76**	**2.37E-02**	**91st**
236	ENSCAFT00000045921		1.02	2.38E-02	Not filtered
237	ENSCAFT00000038891	JAM2	0.74	2.38E-02	Not filtered
238	ENSCAFT00000000665	CPM	-1.15	2.38E-02	Not filtered
239	**ENSCAFT00000002268**	**MEST**	**0.96**	**2.39E-02**	**93rd**
240	ENSCAFT00000008331	MYO7A	0.82	2.40E-02	Not filtered
241	ENSCAFT00000015657	PPM1M	-0.62	2.46E-02	Not filtered
242	ENSCAFT00000048836	ZNF471	1.22	2.48E-02	Not filtered
243	ENSCAFT00000003615	ECM2	1.09	2.48E-02	Not filtered
244	ENSCAFT00000031573	ACSF3	-0.49	2.48E-02	Not filtered
245	ENSCAFT00000045753	PDZRN4	0.85	2.50E-02	Not filtered
246	ENSCAFT00000029091	ALKBH5	-0.83	2.51E-02	Not filtered
247	ENSCAFT00000020583	PCYT1A	0.48	2.52E-02	Not filtered
248	ENSCAFT00000012231	PREX2	1.11	2.56E-02	Not filtered
249	ENSCAFT00000045600	ZDBF2	0.87	2.56E-02	Not filtered
250	ENSCAFT00000047906		-1.56	2.56E-02	Not filtered
251	ENSCAFT00000023165	PYGL	-0.59	2.57E-02	Not filtered
252	ENSCAFT00000044729	GPR179	0.67	2.57E-02	Not filtered
253	ENSCAFT00000049677	PNPLA4	1.42	2.58E-02	Not filtered
254	ENSCAFT00000044927		0.73	2.58E-02	Not filtered
255	ENSCAFT00000013389	PAG1	0.65	2.58E-02	Not filtered
256	ENSCAFT00000024467	KIAA0753	-0.55	2.58E-02	Not filtered
257	ENSCAFT00000036074		-0.67	2.64E-02	Not filtered
258	ENSCAFT00000016681	SLC19A3	0.98	2.67E-02	Not filtered
259	ENSCAFT00000049634		3.88	2.67E-02	Not filtered
260	ENSCAFT00000021268		1.64	2.67E-02	Not filtered
261	ENSCAFT00000024816	GNB5	-0.51	2.68E-02	Not filtered
262	ENSCAFT00000031839	CDC14A	1.02	2.70E-02	Not filtered
263	ENSCAFT00000002695	C11H9orf72	0.71	2.72E-02	Not filtered
264	ENSCAFT00000015985	LZTS1	-0.46	2.72E-02	Not filtered
265	ENSCAFT00000002677	TEK	0.86	2.75E-02	Not filtered
266	ENSCAFT00000008545	SOCS3	-1.50	2.76E-02	Not filtered
267	ENSCAFT00000047656	RGR	2.11	2.79E-02	Not filtered
268	ENSCAFT00000049541	GRAMD2B	0.65	2.79E-02	Not filtered
269	ENSCAFT00000009396	KCNH8	-0.90	2.82E-02	Not filtered
270	ENSCAFT00000046228	COL25A1	1.14	2.84E-02	Not filtered
271	**ENSCAFT00000013466**		**1.10**	**2.84E-02**	**93rd**
272	ENSCAFT00000000481	LRIG3	-0.76	2.84E-02	Not filtered
273	ENSCAFT00000031177	RPUSD1	-0.78	2.84E-02	Not filtered
274	ENSCAFT00000050051		4.64	2.86E-02	Not filtered
275	**ENSCAFT00000001481**	**GJA1**	**1.01**	**2.88E-02**	**98th**
276	ENSCAFT00000031814	COL11A2	0.80	2.88E-02	Not filtered
277	ENSCAFT00000018339	ABLIM1	0.62	2.88E-02	Not filtered
278	ENSCAFT00000048682		-0.58	2.88E-02	Not filtered
279	ENSCAFT00000031391	KCNT1	-0.64	2.88E-02	Not filtered
280	ENSCAFT00000001158	CERK	-0.68	2.88E-02	Not filtered
281	ENSCAFT00000044626	MYH13	-1.31	2.89E-02	Not filtered
282	ENSCAFT00000016709		-3.71	2.90E-02	Not filtered
283	**ENSCAFT00000022271**	**LANCL1**	**0.58**	**2.92E-02**	**96th**
284	ENSCAFT00000011905	SLC7A10	1.22	2.95E-02	Not filtered
285	ENSCAFT00000047021	NIPAL1	0.96	2.95E-02	Not filtered
286	ENSCAFT00000027279	ZDHHC15	0.76	2.95E-02	Not filtered
287	ENSCAFT00000010231	KCTD16	0.98	2.98E-02	Not filtered
288	ENSCAFT00000007517	ITGA8	0.91	2.98E-02	Not filtered

289	**ENSCAFT00000011978**	**RILPL1**	**-0.71**	**2.98E-02**	**98th**
290	ENSCAFT00000043178	PTPDC1	-0.99	3.02E-02	Not filtered
291	ENSCAFT00000012618	RHOBTB3	0.93	3.03E-02	Not filtered
292	ENSCAFT00000020084	DERA	0.62	3.03E-02	Not filtered
293	ENSCAFT00000023491	GULP1	0.92	3.11E-02	Not filtered
294	ENSCAFT00000011507	PTPRG	0.66	3.11E-02	Not filtered
295	**ENSCAFT00000018895**	**BZW1**	**0.66**	**3.11E-02**	**96th**
296	ENSCAFT00000044975	JADE2	-0.65	3.11E-02	Not filtered
297	ENSCAFT00000012478	RETSAT	0.67	3.14E-02	Not filtered
298	ENSCAFT00000009653	MGAT4C	0.82	3.18E-02	Not filtered
299	ENSCAFT00000043408	GTF3C3	0.72	3.20E-02	Not filtered
300	ENSCAFT00000004517	NFYC	-0.65	3.30E-02	Not filtered
301	ENSCAFT00000050102		1.52	3.39E-02	Not filtered
302	**ENSCAFT00000046855**	**CERK**	**-0.59**	**3.39E-02**	**90th**
303	ENSCAFT00000049206		-1.37	3.41E-02	Not filtered
304	ENSCAFT00000003755	ANKRD39	-0.73	3.41E-02	Not filtered
305	ENSCAFT00000049378		-0.86	3.53E-02	Not filtered
306	ENSCAFT00000027762	PAQR5	1.00	3.53E-02	Not filtered
307	ENSCAFT00000002737	CHRNA5	0.89	3.53E-02	Not filtered
308	ENSCAFT00000026321	MSN	0.52	3.53E-02	Not filtered
309	ENSCAFT00000003267	ADGRF5	0.87	3.56E-02	Not filtered
310	ENSCAFT00000007981	PAK3	0.60	3.56E-02	Not filtered
311	ENSCAFT00000037089	DPY19L3	0.60	3.59E-02	Not filtered
312	ENSCAFT00000003998	P3H1	-0.72	3.62E-02	Not filtered
313	ENSCAFT00000049967	SERPINB9	0.69	3.63E-02	Not filtered
314	ENSCAFT00000000818	SNCAIP	0.63	3.63E-02	Not filtered
315	ENSCAFT00000042869	TFAP4	-0.82	3.63E-02	Not filtered
316	ENSCAFT00000015365	WNT8B	-2.98	3.63E-02	Not filtered
317	**ENSCAFT00000046391**	**RANBP3L**	**1.37**	**3.63E-02**	**91st**
318	**ENSCAFT00000029168**	**TMEM98**	**-0.79**	**3.66E-02**	**95th**
319	ENSCAFT00000028199		-4.63	3.70E-02	Not filtered
320	ENSCAFT00000000638	PLEKHG1	0.59	3.71E-02	Not filtered
321	ENSCAFT00000037219		0.75	3.72E-02	Not filtered
322	ENSCAFT00000024144	SNAP29	0.50	3.72E-02	Not filtered
323	ENSCAFT00000019616	HMBS	-0.62	3.75E-02	Not filtered
324	ENSCAFT00000010343	AMDHD1	-0.71	3.79E-02	Not filtered
325	ENSCAFT00000046491	RHBDL3	-1.08	3.79E-02	Not filtered
326	ENSCAFT00000035378		-2.02	3.79E-02	Not filtered
327	ENSCAFT00000025789	PGM2	0.52	3.80E-02	Not filtered
328	ENSCAFT00000045572	GHR	0.72	3.81E-02	Not filtered
329	ENSCAFT00000046022	TCF23	2.05	3.82E-02	Not filtered
330	**ENSCAFT00000024665**	**TMOD2**	**0.55**	**3.83E-02**	**95th**
331	ENSCAFT00000002700	TOM1	-0.67	3.88E-02	Not filtered
332	ENSCAFT00000046313	CCR6	2.57	3.88E-02	Not filtered
333	**ENSCAFT00000009471**	**ANAPC11**	**-0.98**	**3.89E-02**	**97th**
334	**ENSCAFT00000023016**	**IGFBP2**	**-1.12**	**3.89E-02**	**92nd**
335	**ENSCAFT00000045335**	**PTPRZ1**	**1.32**	**3.91E-02**	**93rd**
336	ENSCAFT00000049097	PPA2	0.73	3.98E-02	Not filtered
337	ENSCAFT00000043409		1.64	3.98E-02	Not filtered
338	**ENSCAFT00000045219**	**ABLIM1**	**0.49**	**3.98E-02**	**90th**
339	ENSCAFT00000044179	NHS	0.95	4.00E-02	Not filtered
340	**ENSCAFT00000030037**	**DEXI**	**-1.27**	**4.01E-02**	**96th**
341	ENSCAFT00000023864	KIF17	-0.68	4.01E-02	Not filtered
342	ENSCAFT00000045914		-1.60	4.01E-02	Not filtered
343	**ENSCAFT00000025184**	**MED11**	**-0.88**	**4.04E-02**	**90th**
344	**ENSCAFT00000019956**	**SH3BGRL3**	**-1.12**	**4.04E-02**	**99th**
345	ENSCAFT00000019633	SLCO1A2	0.95	4.06E-02	Not filtered
346	ENSCAFT00000010087	DHX57	0.63	4.06E-02	Not filtered
347	**ENSCAFT00000003721**	**GRHPR**	**-0.76**	**4.06E-02**	**91st**
348	ENSCAFT00000009843	POSTN	1.42	4.07E-02	Not filtered
349	**ENSCAFT00000038326**	**SQLE**	**0.55**	**4.07E-02**	**90th**
350	ENSCAFT00000026389	ADAM21	-0.95	4.12E-02	Not filtered
351	**ENSCAFT00000023712**	**CYP27A1**	**-1.21**	**4.12E-02**	**94th**
352	ENSCAFT00000032397	ADGRL4	1.07	4.13E-02	Not filtered
353	ENSCAFT00000000847	RGS17	0.83	4.13E-02	Not filtered
354	ENSCAFT00000029649	SMARCA1	0.78	4.14E-02	Not filtered
355	ENSCAFT00000017004	SLC39A8	0.66	4.16E-02	Not filtered
356	ENSCAFT00000032451	E2F4	-0.52	4.22E-02	Not filtered
357	ENSCAFT00000003269	ADGRF5	0.72	4.23E-02	Not filtered
358	ENSCAFT00000045054	EPS8	0.94	4.24E-02	Not filtered
359	**ENSCAFT00000047366**	**SEC23A**	**0.73**	**4.24E-02**	**90th**
360	ENSCAFT00000049670		-0.78	4.24E-02	Not filtered
361	ENSCAFT00000043666		-1.85	4.25E-02	Not filtered
362	ENSCAFT00000044299	KDM1A	-0.31	4.26E-02	Not filtered

363	ENSCAFT00000031183	KLHL21	-0.72	4.26E-02	Not filtered
364	ENSCAFT00000004714		0.87	4.26E-02	Not filtered
365	ENSCAFT00000008283	FNDC4	-0.63	4.26E-02	Not filtered
366	**ENSCAFT00000006737**		**-0.62**	**4.26E-02**	**96th**
367	ENSCAFT00000043067	KIF17	-0.74	4.26E-02	Not filtered
368	**ENSCAFT00000011343**		**-1.39**	**4.26E-02**	**95th**
369	ENSCAFT00000022161	SYTL5	1.36	4.27E-02	Not filtered
370	ENSCAFT00000000089	ITGA7	0.76	4.27E-02	Not filtered
371	**ENSCAFT00000007160**	**FOSB**	**-1.08**	**4.27E-02**	**93rd**
372	ENSCAFT00000024296	CD38	1.00	4.28E-02	Not filtered
373	ENSCAFT00000023745	ZSCAN25	-0.61	4.28E-02	Not filtered
374	**ENSCAFT00000018282**	**AIP**	**-0.72**	**4.28E-02**	**94th**
375	ENSCAFT00000017419	ACSL5	1.05	4.29E-02	Not filtered
376	ENSCAFT00000025658	FAM114A1	0.76	4.29E-02	Not filtered
377	ENSCAFT00000016750	OASL	-0.40	4.30E-02	Not filtered
378	**ENSCAFT00000008137**	**RNF157**	**-0.39**	**4.32E-02**	**94th**
379	ENSCAFT00000029520	LAMA1	0.95	4.34E-02	Not filtered
380	ENSCAFT00000047858	PCSK5	0.66	4.35E-02	Not filtered
381	**ENSCAFT00000029484**		**0.64**	**4.35E-02**	**96th**
382	ENSCAFT00000020806	SART1	-0.46	4.35E-02	Not filtered
383	ENSCAFT00000004933	NOD1	-0.70	4.35E-02	Not filtered
384	ENSCAFT00000003500	PTPDC1	-1.01	4.35E-02	Not filtered
385	**ENSCAFT00000029354**		**-2.51**	**4.37E-02**	**92nd**
386	ENSCAFT00000038139	SYCE1	3.29	4.42E-02	Not filtered
387	ENSCAFT00000046875	PIK3C2A	0.78	4.42E-02	Not filtered
388	**ENSCAFT00000009006**	**NCCRP1**	**-3.50**	**4.42E-02**	**90th**
389	ENSCAFT00000014595	GPR15	-2.61	4.43E-02	Not filtered
390	ENSCAFT00000044965	EGLN2	-0.77	4.43E-02	Not filtered
391	**ENSCAFT00000045554**		**-2.60**	**4.46E-02**	**91st**
392	**ENSCAFT00000022594**	**NCAN**	**0.87**	**4.47E-02**	**93rd**
393	ENSCAFT00000007000	GPR17	1.35	4.49E-02	Not filtered
394	**ENSCAFT00000020636**	**ACE**	**1.31**	**4.49E-02**	**97th**
395	**ENSCAFT00000025424**	**AKAP5**	**0.76**	**4.49E-02**	**95th**
396	ENSCAFT00000027662	MYH13	-0.98	4.49E-02	Not filtered
397	ENSCAFT00000023000		-1.26	4.50E-02	Not filtered
398	ENSCAFT00000022292	CDCP1	0.83	4.56E-02	Not filtered
399	ENSCAFT00000023561	DNM3	0.49	4.56E-02	Not filtered
400	ENSCAFT00000001623	TXNDC15	0.44	4.56E-02	Not filtered
401	ENSCAFT00000028088		-0.36	4.56E-02	Not filtered
402	ENSCAFT00000019444	MLPH	-2.01	4.56E-02	Not filtered
403	ENSCAFT00000002939	TRPM3	0.94	4.64E-02	Not filtered
404	ENSCAFT00000015989	ANO5	0.74	4.64E-02	Not filtered
405	ENSCAFT00000012249	C29H8orf34	0.90	4.66E-02	Not filtered
406	ENSCAFT00000008063	CCDC97	-0.56	4.66E-02	Not filtered
407	**ENSCAFT00000010234**	**LAMP1**	**-0.32**	**4.68E-02**	**96th**
408	ENSCAFT00000023805	MAPK1IP1L	0.51	4.72E-02	Not filtered
409	ENSCAFT00000013457	BLNK	1.08	4.72E-02	Not filtered
410	ENSCAFT00000039684		0.83	4.72E-02	Not filtered
411	ENSCAFT00000022836	ARHGEF10	-0.75	4.72E-02	Not filtered
412	ENSCAFT00000029569	CACHD1	0.54	4.75E-02	Not filtered
413	ENSCAFT00000012751	ZNF366	0.93	4.75E-02	Not filtered
414	ENSCAFT00000018076	ACTL6A	0.73	4.76E-02	Not filtered
415	ENSCAFT00000004830	SUSD1	0.69	4.76E-02	Not filtered
416	ENSCAFT00000011640		0.73	4.78E-02	Not filtered
417	ENSCAFT00000045071	TIPARP	0.96	4.80E-02	Not filtered
418	**ENSCAFT00000024554**	**DES**	**-1.26**	**4.81E-02**	**91st**
419	ENSCAFT00000021954	SEC23A	0.53	4.83E-02	Not filtered
420	ENSCAFT00000035362	RCCD1	-0.78	4.83E-02	Not filtered
421	ENSCAFT00000005238	ARHGAP25	0.86	4.83E-02	Not filtered
422	ENSCAFT00000049648		-1.05	4.85E-02	Not filtered
423	ENSCAFT00000013538	P2RY12	0.93	4.87E-02	Not filtered
424	ENSCAFT00000043524	COL19A1	0.75	4.87E-02	Not filtered
425	ENSCAFT00000000563	GNS	0.51	4.87E-02	Not filtered
426	ENSCAFT00000020997	PISD	-0.55	4.87E-02	Not filtered
427	**ENSCAFT00000024280**	**KCNA1**	**-0.66**	**4.88E-02**	**90th**
428	**ENSCAFT00000017551**	**SLC25A6**	**-0.65**	**4.89E-02**	**98th**
429	**ENSCAFT00000046388**		**-1.37**	**4.91E-02**	**99th**
430	ENSCAFT00000044007	ELN	1.19	4.95E-02	Not filtered
	ENSCAFT00000023394	**FBN1**			**Added 97th.95th.90th**
	ENSCAFT00000009676	**LHFPL6**			**Added 97th.95th.90th**
	ENSCAFT00000008870	**PFKP**			**Added 97th.95th**
	ENSCAFT00000047460	**ANKRD31**			**Added 97th.95th.90th**
	ENSCAFT00000012552				**Added 97th.95th.90th**
	ENSCAFT00000048320	**ZNHIT1**			**Added 97th.95th.90th**

ENSCAFT00000029069	PCYOX1L	Added 97th.95th.90th
ENSCAFT00000007457		Added 97th.95th
ENSCAFT00000030988	HAGH	Added 97th
ENSCAFT00000031390	KCNT1	Added 97th.95th.90th
ENSCAFT00000039399	SLC25A28	Added 97th.95th.90th
ENSCAFT00000029752	RANBP3	Added 97th.95th.90th
ENSCAFT00000045978	TIMP3	Added 97th.95th
ENSCAFT00000026940	AURKB	Added 97th.95th.90th
ENSCAFT00000029071	XAB2	Added 97th.95th.90th
ENSCAFT00000049561	MRI1	Added 97th.95th.90th
ENSCAFT00000007204	XPC	Added 97th.95th.90th
ENSCAFT00000043301	COL11A1	Added 97th
ENSCAFT00000022098	FAM20B	Added 97th
ENSCAFT00000025496	LDB3	Added 97th.95th.90th
ENSCAFT00000007637	PHYH	Added 97th.95th.90th
ENSCAFT00000002119	SUN2	Added 95th.90th
ENSCAFT00000025645	EPHX1	Added 95th.90th
ENSCAFT00000026577	PDE6B	Added 95th.90th
ENSCAFT00000014889	C18H11orf94	Added 95th.90th
ENSCAFT00000036022	PHYH	Added 95th.90th
ENSCAFT00000020738	TPD52L2	Added 90th
ENSCAFT00000018896	POLG	Added 90th
ENSCAFT00000010783	SLCO2A1	Added 90th
ENSCAFT00000018755	CRYBA4	Added 90th
ENSCAFT00000017387	HFE	Added 90th
ENSCAFT00000028420	CLK3	Added 90th

Supplementary Table 6.8. DETs in aggressive vs. tame fox. Complete list of DETs in tame foxes when compared to aggressive foxes, identified using non-filtered datasets, and those that got removed (red) within the highest 10% of intra-condition variation, as also those added (green) as differentially expressed across selected filtered datasets (97th, 95th, and 90th percentiles). The correspondent annotated gene ID (when available), log2FC values and p-values are provided.

Ranking	Transcript ID	Gene ID	log2 FC	p adj value	Filter level
1	ENSCAFT00000001643	TRIB1	0.91	2.82E-07	Not-filtered
2	ENSCAFT00000012381	PRIMPOL	-0.86	2.82E-07	Not-filtered
3	ENSCAFT00000002697	FSCN3	-0.79	2.82E-07	Not-filtered
4	ENSCAFT00000003134	CCDC138	-0.98	9.06E-06	Not-filtered
5	ENSCAFT00000015069	FARS2	0.58	2.20E-05	Not-filtered
6	ENSCAFT00000032259	NQO1	0.62	2.66E-05	Not-filtered
7	ENSCAFT00000046785		1.10	5.56E-05	Not-filtered
8	ENSCAFT00000009874	DUSP6	0.67	5.56E-05	Not-filtered
9	ENSCAFT00000047474		1.07	9.37E-05	Not-filtered
10	**ENSCAFT00000001962**	**EGR1**	**0.86**	**1.72E-04**	**95th**
11	ENSCAFT00000015649		0.92	1.77E-04	Not-filtered
12	ENSCAFT00000048992		1.11	1.85E-04	Not-filtered
13	ENSCAFT00000035279		1.04	1.85E-04	Not-filtered
14	ENSCAFT00000044393	KHK	1.03	1.85E-04	Not-filtered
15	ENSCAFT00000045717		0.99	1.85E-04	Not-filtered
16	ENSCAFT00000015525		0.98	1.85E-04	Not-filtered
17	**ENSCAFT00000048012**	**NRSN1**	**0.77**	**1.85E-04**	**97th**
18	ENSCAFT00000029200	ESM1	-1.45	2.94E-04	Not-filtered
19	ENSCAFT00000039205		0.99	3.11E-04	Not-filtered
20	ENSCAFT00000043849	EGR2	0.72	4.02E-04	Not-filtered
21	**ENSCAFT00000031607**	**APRT**	**0.99**	**4.20E-04**	**96th**
22	ENSCAFT00000031690	IRF8	0.40	4.20E-04	Not-filtered
23	ENSCAFT00000038072	SAPCD1	-1.35	4.32E-04	Not-filtered
24	ENSCAFT00000017242		1.12	4.96E-04	Not-filtered
25	ENSCAFT00000015258		-1.39	5.47E-04	Not-filtered
26	ENSCAFT00000026984	FOS	1.03	5.65E-04	Not-filtered
27	ENSCAFT00000031697	C5H16orf74	1.09	5.67E-04	Not-filtered
28	ENSCAFT00000008522	SLITRK6	-1.50	6.59E-04	Not-filtered
29	ENSCAFT00000020697	EGR2	0.67	8.43E-04	Not-filtered
30	**ENSCAFT00000009043**	**BHLHE40**	**0.54**	**9.46E-04**	**91st**
31	ENSCAFT00000007514	KHK	0.84	1.00E-03	Not-filtered
32	**ENSCAFT00000001517**	**NDUFA6**	**0.35**	**1.02E-03**	**93rd**
33	ENSCAFT00000011062	DDX4	-2.02	1.02E-03	Not-filtered
34	ENSCAFT00000005919	DKKL1	0.63	1.05E-03	Not-filtered
35	ENSCAFT00000003135	CCDC138	-1.03	1.07E-03	Not-filtered
36	**ENSCAFT00000043415**	**JUNB**	**1.11**	**1.23E-03**	**96th**
37	ENSCAFT00000018592	INPP5D	0.68	1.23E-03	Not-filtered
38	ENSCAFT00000004310		-1.91	1.23E-03	Not-filtered
39	ENSCAFT00000011853	TMEM232	-1.08	1.47E-03	Not-filtered
40	**ENSCAFT00000014839**	**BPHL**	**0.41**	**1.48E-03**	**91st**
41	**ENSCAFT00000027182**		**1.07**	**1.56E-03**	**94th**
42	ENSCAFT00000007767	ATRAID	0.62	1.67E-03	Not-filtered
43	ENSCAFT00000009509	RYR1	-0.41	1.67E-03	Not-filtered
44	ENSCAFT00000046725		-1.91	1.69E-03	Not-filtered
45	ENSCAFT00000016864		0.97	1.73E-03	Not-filtered
46	ENSCAFT00000005238	ARHGAP25	0.70	1.73E-03	Not-filtered
47	ENSCAFT00000013583	WDR54	0.55	1.79E-03	Not-filtered
48	ENSCAFT00000007829	ECHDC3	0.80	2.09E-03	Not-filtered
49	ENSCAFT00000009853	PCDHGA1	0.57	2.16E-03	Not-filtered
50	**ENSCAFT00000028075**	**ARMCX3**	**0.32**	**2.19E-03**	**92nd**
51	ENSCAFT00000042883	DUSP4	0.49	2.34E-03	Not-filtered
52	**ENSCAFT00000001053**	**SKIV2L**	**0.87**	**2.62E-03**	**91st**
53	ENSCAFT00000037859	YIPF1	0.39	2.84E-03	Not-filtered
54	ENSCAFT00000028559	PRPS1	0.47	3.05E-03	Not-filtered
55	ENSCAFT00000011460	FBLN7	0.54	3.25E-03	Not-filtered
56	ENSCAFT00000006262	HSD17B14	0.70	3.58E-03	Not-filtered
57	ENSCAFT00000036730	MAP2K3	0.69	3.58E-03	Not-filtered
58	ENSCAFT00000007384		0.65	3.58E-03	Not-filtered
59	ENSCAFT00000018894	PTAFR	0.61	3.58E-03	Not-filtered
60	ENSCAFT00000031437		0.42	4.04E-03	Not-filtered
61	**ENSCAFT00000046412**	**GNG4**	**0.79**	**4.35E-03**	**95th**
62	ENSCAFT00000015473	GLB1L3	-0.59	4.62E-03	Not-filtered
63	ENSCAFT00000027070	MYCBPAP	0.81	4.64E-03	Not-filtered
64	ENSCAFT00000005724	CDK5RAP2	0.52	4.64E-03	Not-filtered
65	ENSCAFT00000037497	C31H21orf62	-1.79	4.69E-03	Not-filtered

66	ENSCAFT00000045655	TDRD15	-1.72	4.93E-03	Not-filtered
67	ENSCAFT00000019265	GPR3	0.77	4.96E-03	Not-filtered
68	ENSCAFT00000048546	PDE4C	-0.53	5.16E-03	Not-filtered
69	ENSCAFT00000030960	TTC34	0.68	5.81E-03	Not-filtered
70	ENSCAFT00000029439	MYO1F	0.90	5.84E-03	Not-filtered
71	ENSCAFT00000045052	IGFBP6	1.00	6.34E-03	Not-filtered
72	ENSCAFT00000048748	CLCF1	0.83	6.73E-03	Not-filtered
73	ENSCAFT00000026360	EIF4E1B	-1.23	6.73E-03	Not-filtered
74	ENSCAFT00000007670	STAC	-0.53	6.86E-03	Not-filtered
75	ENSCAFT00000046113	LDLR	0.52	6.94E-03	Not-filtered
76	ENSCAFT00000011875	THNSL2	-0.41	6.94E-03	Not-filtered
77	ENSCAFT00000013860	SPI1	0.80	7.78E-03	Not-filtered
78	ENSCAFT00000018652		0.48	7.78E-03	Not-filtered
79	ENSCAFT00000038333	DNAJB5	0.41	7.78E-03	Not-filtered
80	ENSCAFT00000000961	CCDC192	-3.77	7.78E-03	Not-filtered
81	ENSCAFT00000037518		-1.73	7.78E-03	Not-filtered
82	ENSCAFT00000000149	CCBE1	-0.77	7.78E-03	Not-filtered
83	ENSCAFT00000048904	RYR1	-0.39	7.78E-03	Not-filtered
84	ENSCAFT00000014735	DCLRE1B	0.29	8.22E-03	Not-filtered
85	**ENSCAFT00000018984**		**0.28**	**8.32E-03**	**90th**
86	ENSCAFT00000011990		-3.51	8.32E-03	Not-filtered
87	ENSCAFT00000018504	SCN7A	-1.23	8.32E-03	Not-filtered
88	ENSCAFT00000031681		0.83	8.37E-03	Not-filtered
89	ENSCAFT00000005356	WHRN	0.88	8.49E-03	Not-filtered
90	ENSCAFT00000027791	LDLR	0.63	8.49E-03	Not-filtered
91	ENSCAFT00000036732	MAP2K3	0.58	8.79E-03	Not-filtered
92	ENSCAFT00000020898	NPL	-0.74	8.99E-03	Not-filtered
93	ENSCAFT00000005230	PLEK	0.81	9.06E-03	Not-filtered
94	ENSCAFT00000042917	TMEM119	0.90	9.15E-03	Not-filtered
95	ENSCAFT00000045786		0.75	9.25E-03	Not-filtered
96	ENSCAFT00000035869		-1.76	9.26E-03	Not-filtered
97	ENSCAFT00000042894	JHY	-1.03	9.26E-03	Not-filtered
98	ENSCAFT00000025337	RASAL3	0.80	1.02E-02	Not-filtered
99	ENSCAFT00000046077	SCN7A	-0.95	1.05E-02	Not-filtered
100	ENSCAFT00000012335	GRK7	-0.96	1.09E-02	Not-filtered
101	ENSCAFT00000007621	PLK3	0.66	1.10E-02	Not-filtered
102	**ENSCAFT00000013864**		**0.35**	**1.11E-02**	**97th**
103	ENSCAFT00000020626	SDHC	0.31	1.11E-02	Not-filtered
104	ENSCAFT00000046956	PTGFRN	0.38	1.12E-02	Not-filtered
105	ENSCAFT00000014439		0.35	1.12E-02	Not-filtered
106	ENSCAFT00000027073	MYCBPAP	0.77	1.13E-02	Not-filtered
107	ENSCAFT00000004532		-2.00	1.16E-02	Not-filtered
108	ENSCAFT00000043505		0.71	1.18E-02	Not-filtered
109	ENSCAFT00000025850	DNAH14	-1.31	1.23E-02	Not-filtered
110	ENSCAFT00000046505	ETV4	0.75	1.23E-02	Not-filtered
111	ENSCAFT00000035664	RARA	0.59	1.27E-02	Not-filtered
112	ENSCAFT00000032475	CES2	0.67	1.27E-02	Not-filtered
113	**ENSCAFT00000043933**	**APOA1**	**0.79**	**1.32E-02**	**99th**
114	**ENSCAFT00000027074**	**CALR**	**0.35**	**1.38E-02**	**98th**
115	ENSCAFT00000045964	COL18A1	0.95	1.38E-02	Not-filtered
116	ENSCAFT00000013671	STBD1	-0.33	1.38E-02	Not-filtered
117	ENSCAFT00000020769	EPHB3	0.71	1.40E-02	Not-filtered
118	ENSCAFT00000015123	COX15	-0.36	1.42E-02	Not-filtered
119	ENSCAFT00000017152	RBM20	-0.58	1.43E-02	Not-filtered
120	**ENSCAFT00000013297**		**0.63**	**1.44E-02**	**91st**
121	ENSCAFT00000049637	SERPINB6	0.72	1.46E-02	Not-filtered
122	ENSCAFT00000031982	SLC25A25	0.48	1.47E-02	Not-filtered
123	ENSCAFT00000005713	CADPS2	0.48	1.48E-02	Not-filtered
124	ENSCAFT00000006366	TP53I3	0.68	1.53E-02	Not-filtered
125	ENSCAFT00000047739	DUSP1	0.65	1.57E-02	Not-filtered
126	ENSCAFT00000045561	FSCN3	-0.85	1.57E-02	Not-filtered
127	ENSCAFT00000007951	DLEC1	-0.61	1.57E-02	Not-filtered
128	ENSCAFT00000030998	CEP104	-0.38	1.57E-02	Not-filtered
129	ENSCAFT00000026053	CAPN8	0.97	1.58E-02	Not-filtered
130	ENSCAFT00000028872	CAPN6	-0.73	1.60E-02	Not-filtered
131	**ENSCAFT00000011743**	**NR4A1**	**0.94**	**1.62E-02**	**93rd**
132	**ENSCAFT00000021280**	**ECHS1**	**0.71**	**1.63E-02**	**95th**
133	ENSCAFT00000031608	CDT1	0.91	1.64E-02	Not-filtered
134	**ENSCAFT00000026906**	**PER1**	**0.77**	**1.66E-02**	**94th**
135	ENSCAFT00000050055	SHCBP1L	-1.44	1.67E-02	Not-filtered
136	**ENSCAFT00000042866**	**RPRML**	**1.18**	**1.69E-02**	**95th**
137	ENSCAFT00000036880	PTGFRN	0.45	1.69E-02	Not-filtered
138	ENSCAFT00000020415	LINGO4	0.91	1.70E-02	Not-filtered
139	ENSCAFT00000003238	SULT1C4	-0.87	1.75E-02	Not-filtered

140	ENSCAFT00000023918		-0.70	1.79E-02	Not-filtered
141	ENSCAFT00000014831	CCDC148	-0.92	1.85E-02	Not-filtered
142	ENSCAFT00000013794	LMBR1L	0.47	1.88E-02	Not-filtered
143	ENSCAFT00000030761		0.38	1.88E-02	Not-filtered
144	ENSCAFT00000049378		-0.49	1.90E-02	Not-filtered
145	ENSCAFT00000032461	B3GNT9	1.07	1.98E-02	Not-filtered
146	ENSCAFT00000037411	PROB1	0.90	1.98E-02	Not-filtered
147	ENSCAFT00000045154	CD82	0.86	1.98E-02	Not-filtered
148	ENSCAFT00000043092	DCXR	0.78	1.98E-02	Not-filtered
149	ENSCAFT00000022898		0.75	1.98E-02	Not-filtered
150	ENSCAFT00000032088	IL34	0.72	1.98E-02	Not-filtered
151	ENSCAFT00000020750	EPHB3	0.68	1.98E-02	Not-filtered
152	ENSCAFT00000014810	SERPINB6	0.67	1.98E-02	Not-filtered
153	ENSCAFT00000019411	COL6A3	-1.31	1.98E-02	Not-filtered
154	ENSCAFT00000000277	MC5R	-0.94	1.98E-02	Not-filtered
155	ENSCAFT00000011022	SLC38A9	-0.58	1.98E-02	Not-filtered
156	ENSCAFT00000021964	IL16	0.63	2.02E-02	Not-filtered
157	ENSCAFT00000011311	RARA	0.59	2.02E-02	Not-filtered
158	ENSCAFT00000045379	TTC34	0.69	2.03E-02	Not-filtered
159	**ENSCAFT00000026939**	**PPP1R9B**	**0.76**	**2.03E-02**	**99th**
160	ENSCAFT00000045878		0.73	2.05E-02	Not-filtered
161	ENSCAFT00000026036	TNFRSF1B	0.84	2.06E-02	Not-filtered
162	ENSCAFT00000036157	C3	0.66	2.16E-02	Not-filtered
163	ENSCAFT00000013878	COQ9	0.34	2.21E-02	Not-filtered
164	ENSCAFT00000048568	CXHXorf36	0.38	2.29E-02	Not-filtered
165	ENSCAFT00000016750	OASL	-0.54	2.33E-02	Not-filtered
166	ENSCAFT00000020372	IL10RA	0.58	2.36E-02	Not-filtered
167	ENSCAFT00000022337	PEX5	0.40	2.36E-02	Not-filtered
168	ENSCAFT00000017458	SLC37A2	0.53	2.37E-02	Not-filtered
169	ENSCAFT00000013687		0.48	2.37E-02	Not-filtered
170	**ENSCAFT00000027951**	**FAXDC2**	**0.50**	**2.39E-02**	**94th**
171	ENSCAFT00000013858	ADRA1A	0.46	2.39E-02	Not-filtered
172	ENSCAFT00000011115	PRR13	0.41	2.41E-02	Not-filtered
173	ENSCAFT00000024439	RCSD1	0.41	2.47E-02	Not-filtered
174	**ENSCAFT00000045771**		**1.31**	**2.48E-02**	**96th**
175	ENSCAFT00000046339	IER5	0.68	2.48E-02	Not-filtered
176	ENSCAFT00000045328	FBXW2	0.21	2.48E-02	Not-filtered
177	ENSCAFT00000019420	COL6A3	-1.32	2.48E-02	Not-filtered
178	**ENSCAFT00000031805**	**IER5L**	**0.83**	**2.48E-02**	**94th**
179	ENSCAFT00000036487	REEP6	0.78	2.48E-02	Not-filtered
180	ENSCAFT00000028543	MPI	0.43	2.48E-02	Not-filtered
181	ENSCAFT00000024605	ZWINT	1.46	2.52E-02	Not-filtered
182	ENSCAFT00000026954	TMEM219	0.41	2.52E-02	Not-filtered
183	**ENSCAFT00000027559**	**SLC39A1**	**0.66**	**2.52E-02**	**90th**
184	ENSCAFT00000048192	LTBR	0.78	2.53E-02	Not-filtered
185	**ENSCAFT00000029765**		**0.71**	**2.53E-02**	**99th**
186	ENSCAFT00000022090	TOR3A	0.42	2.53E-02	Not-filtered
187	ENSCAFT00000023727	RPL22L1	0.40	2.53E-02	Not-filtered
188	ENSCAFT00000003268	CD72	0.47	2.54E-02	Not-filtered
189	ENSCAFT00000032002	ENG	0.71	2.58E-02	Not-filtered
190	ENSCAFT00000042960		1.29	2.59E-02	Not-filtered
191	**ENSCAFT00000014198**	**RAB11FIP5**	**0.72**	**2.59E-02**	**90th**
192	**ENSCAFT00000026249**		**0.84**	**2.61E-02**	**91st**
193	ENSCAFT00000007977	MYD88	0.58	2.61E-02	Not-filtered
194	ENSCAFT00000020742	FUCA1	0.23	2.61E-02	Not-filtered
195	ENSCAFT00000019426	COL6A3	-1.31	2.61E-02	Not-filtered
196	ENSCAFT00000008181	CSRNP1	0.84	2.62E-02	Not-filtered
197	ENSCAFT00000029568	C3	0.63	2.62E-02	Not-filtered
198	ENSCAFT00000004941	PABPC4	0.85	2.64E-02	Not-filtered
199	**ENSCAFT00000000260**	**NAB2**	**0.72**	**2.65E-02**	**92nd**
200	ENSCAFT00000009541	MYADML2	0.92	2.67E-02	Not-filtered
201	ENSCAFT00000017456	RGMA	0.75	2.67E-02	Not-filtered
202	ENSCAFT00000046370		0.65	2.67E-02	Not-filtered
203	ENSCAFT00000031626	CTU2	0.65	2.67E-02	Not-filtered
204	**ENSCAFT00000028433**	**SEMA7A**	**0.75**	**2.69E-02**	**97th**
205	ENSCAFT00000020170	MRGBP	0.66	2.72E-02	Not-filtered
206	ENSCAFT00000011842	PEPD	0.49	2.72E-02	Not-filtered
207	**ENSCAFT00000001475**	**CYB5R3**	**0.77**	**2.74E-02**	**96th**
208	ENSCAFT00000047908	CCDC116	0.77	2.76E-02	Not-filtered
209	ENSCAFT00000048129	SASH3	0.62	2.76E-02	Not-filtered
210	ENSCAFT00000044057		0.59	2.76E-02	Not-filtered
211	ENSCAFT00000045210	TAS1R2	0.75	2.79E-02	Not-filtered
212	ENSCAFT00000045709	WNT1	0.68	2.79E-02	Not-filtered
213	ENSCAFT00000018498	MECR	0.51	2.79E-02	Not-filtered
214	ENSCAFT00000018820	COL18A1	0.75	2.82E-02	Not-filtered

215	ENSCAFT00000029463	TCTEX1D1	-0.68	2.82E-02	Not-filtered
216	ENSCAFT00000031121	HES2	0.96	2.83E-02	Not-filtered
217	ENSCAFT00000047285	TRIM17	-1.02	2.83E-02	Not-filtered
218	ENSCAFT00000038993	TTC6	-1.01	2.83E-02	Not-filtered
219	ENSCAFT00000011362		1.22	2.85E-02	Not-filtered
220	**ENSCAFT00000021138**	**APOA1**	**0.76**	**2.85E-02**	**99th**
221	ENSCAFT00000032517	CMTM3	0.73	2.85E-02	Not-filtered
222	ENSCAFT00000011777	KCTD15	0.69	2.85E-02	Not-filtered
223	ENSCAFT00000037777	SRD5A1	0.55	2.85E-02	Not-filtered
224	**ENSCAFT00000023243**	**TMBIM1**	**0.54**	**2.85E-02**	**93rd**
225	ENSCAFT00000030969	TMEM203	0.52	2.85E-02	Not-filtered
226	ENSCAFT00000009183	SPATA24	0.51	2.85E-02	Not-filtered
227	ENSCAFT00000049973	ACO1	0.32	2.85E-02	Not-filtered
228	ENSCAFT00000008481	SPRY2	0.25	2.85E-02	Not-filtered
229	**ENSCAFT00000028730**		**0.23**	**2.85E-02**	**95th**
230	ENSCAFT00000000209	FECH	0.20	2.85E-02	Not-filtered
231	ENSCAFT00000006192	MGAM	-1.74	2.85E-02	Not-filtered
232	ENSCAFT00000032444		-1.46	2.85E-02	Not-filtered
233	ENSCAFT00000044050	UOX	-0.65	2.85E-02	Not-filtered
234	ENSCAFT00000011576	HTR7	-0.41	2.85E-02	Not-filtered
235	**ENSCAFT00000038047**	**OXLD1**	**0.65**	**2.85E-02**	**92nd**
236	ENSCAFT00000017780	RBM20	-0.53	2.85E-02	Not-filtered
237	ENSCAFT00000018791	RHOD	0.44	2.85E-02	Not-filtered
238	ENSCAFT00000047799	SHD	0.86	2.85E-02	Not-filtered
239	ENSCAFT00000038326	SQLE	0.29	2.85E-02	Not-filtered
240	ENSCAFT00000000090	ITGA7	0.47	2.88E-02	Not-filtered
241	ENSCAFT00000021287	MAP3K11	0.72	2.91E-02	Not-filtered
242	ENSCAFT00000048304		1.04	2.91E-02	Not-filtered
243	**ENSCAFT00000030085**	**PLIN3**	**0.99**	**2.91E-02**	**92nd**
244	ENSCAFT00000023505	TSSK1B	0.82	2.91E-02	Not-filtered
245	ENSCAFT00000022217	ATP2A3	0.78	2.91E-02	Not-filtered
246	ENSCAFT00000010557	ARID5A	0.75	2.91E-02	Not-filtered
247	ENSCAFT00000030728	PKMYT1	0.71	2.91E-02	Not-filtered
248	ENSCAFT00000027675	NSMCE1	0.39	2.91E-02	Not-filtered
249	**ENSCAFT00000006939**		**0.60**	**2.93E-02**	**91st**
250	ENSCAFT00000046137	SPIDR	0.30	2.98E-02	Not-filtered
251	ENSCAFT00000011244	DEPP1	0.91	3.02E-02	Not-filtered
252	ENSCAFT00000000389	MAP7	-0.74	3.02E-02	Not-filtered
253	**ENSCAFT00000032429**	**ZDHHC1**	**0.67**	**3.03E-02**	**93rd**
254	ENSCAFT00000049248	OTOS	1.67	3.05E-02	Not-filtered
255	ENSCAFT00000002252	FGD2	0.52	3.11E-02	Not-filtered
256	ENSCAFT00000039695	SCLY	0.72	3.11E-02	Not-filtered
257	ENSCAFT00000031098	TMEM141	0.60	3.18E-02	Not-filtered
258	**ENSCAFT00000027334**		**0.70**	**3.26E-02**	**94th**
259	ENSCAFT00000017511	HEPACAM	0.57	3.26E-02	Not-filtered
260	ENSCAFT00000046690	SRA1	0.31	3.27E-02	Not-filtered
261	ENSCAFT00000043352	PROKR1	-1.08	3.27E-02	Not-filtered
262	ENSCAFT00000038598	DXO	0.53	3.28E-02	Not-filtered
263	ENSCAFT00000031678	PLPP7	0.69	3.28E-02	Not-filtered
264	ENSCAFT00000026637	SH3PXD2B	0.54	3.35E-02	Not-filtered
265	ENSCAFT00000004116	TNNI3	0.72	3.36E-02	Not-filtered
266	ENSCAFT00000009197	TMEM173	0.54	3.37E-02	Not-filtered
267	**ENSCAFT00000003950**	**SLC2A1**	**0.52**	**3.37E-02**	**92nd**
268	ENSCAFT00000023813	LGALS3	0.43	3.37E-02	Not-filtered
269	ENSCAFT00000011570		0.39	3.37E-02	Not-filtered
270	ENSCAFT00000043190	FBLN7	0.38	3.37E-02	Not-filtered
271	ENSCAFT00000024118		0.36	3.37E-02	Not-filtered
272	ENSCAFT00000017160	TNNI1	1.10	3.39E-02	Not-filtered
273	**ENSCAFT00000048841**	**CD99L2**	**0.70**	**3.39E-02**	**97th**
274	ENSCAFT00000046384	ENG	0.66	3.39E-02	Not-filtered
275	ENSCAFT00000014976	CHAC1	0.63	3.39E-02	Not-filtered
276	ENSCAFT00000026288	SMOC1	0.57	3.39E-02	Not-filtered
277	**ENSCAFT00000047941**		**0.42**	**3.39E-02**	**93rd**
278	ENSCAFT00000007430	AGBL5	0.37	3.39E-02	Not-filtered
279	ENSCAFT00000015117	ABI3BP	-1.40	3.39E-02	Not-filtered
280	ENSCAFT00000014836	CCDC148	-0.81	3.39E-02	Not-filtered
281	ENSCAFT00000009407	ZIC5	-0.57	3.39E-02	Not-filtered
282	ENSCAFT00000015504	PDZD7	-0.47	3.39E-02	Not-filtered
283	ENSCAFT00000028352	ISLR	0.81	3.47E-02	Not-filtered
284	ENSCAFT00000024881	ACTA2	0.71	3.49E-02	Not-filtered
285	ENSCAFT00000002392	TSPAN33	0.43	3.51E-02	Not-filtered
286	ENSCAFT00000043744	MESP2	0.66	3.53E-02	Not-filtered
287	**ENSCAFT00000046278**	**CYB5R3**	**0.73**	**3.54E-02**	**94th**
288	ENSCAFT00000007733	TSEN54	0.57	3.54E-02	Not-filtered

289	ENSCAFT00000015914	CBLN1	0.52	3.54E-02	Not-filtered
290	**ENSCAFT00000020662**	**CYB561**	**0.60**	**3.55E-02**	**93rd**
291	ENSCAFT00000022385	BCOR	0.38	3.57E-02	Not-filtered
292	**ENSCAFT00000048276**		**0.71**	**3.63E-02**	**90th**
293	ENSCAFT00000015425	TSHB	-2.83	3.63E-02	Not-filtered
294	ENSCAFT00000015161	KCNE2	-0.67	3.63E-02	Not-filtered
295	**ENSCAFT00000031049**	**MIDN**	**0.76**	**3.64E-02**	**95th**
296	ENSCAFT00000042966	CCDC116	0.73	3.64E-02	Not-filtered
297	ENSCAFT00000048895	EPS15L1	0.27	3.64E-02	Not-filtered
298	ENSCAFT00000039155	INSC	1.04	3.65E-02	Not-filtered
299	ENSCAFT00000025901	CNIH3	0.81	3.65E-02	Not-filtered
300	ENSCAFT00000036972	BCAS4	0.80	3.65E-02	Not-filtered
301	**ENSCAFT00000047634**	**RAB11FIP5**	**0.69**	**3.65E-02**	**92nd**
302	ENSCAFT00000006437	TBXAS1	0.54	3.65E-02	Not-filtered
303	ENSCAFT00000003334	FHL2	0.40	3.65E-02	Not-filtered
304	ENSCAFT00000000329	CDO1	0.31	3.65E-02	Not-filtered
305	ENSCAFT00000045689	TBC1D7	0.25	3.65E-02	Not-filtered
306	ENSCAFT00000028047	SERPINA3	-1.03	3.65E-02	Not-filtered
307	ENSCAFT00000039388		-0.44	3.65E-02	Not-filtered
308	ENSCAFT00000016952	HYAL2	0.68	3.66E-02	Not-filtered
309	**ENSCAFT00000009467**	**RFTN1**	**0.61**	**3.67E-02**	**90th**
310	**ENSCAFT00000009662**		**0.47**	**3.68E-02**	**90th**
311	**ENSCAFT00000030879**	**PRXL2B**	**0.80**	**3.73E-02**	**91st**
312	ENSCAFT00000036196	CTSH	0.53	3.77E-02	Not-filtered
313	ENSCAFT00000016466	CYB5R1	0.47	3.77E-02	Not-filtered
314	ENSCAFT00000008331	MYO7A	0.39	3.77E-02	Not-filtered
315	ENSCAFT00000043486	RBM38	0.66	3.79E-02	Not-filtered
316	ENSCAFT00000005213	C11H9orf43	-1.43	3.80E-02	Not-filtered
317	ENSCAFT00000027507	ADAM19	0.42	3.80E-02	Not-filtered
318	**ENSCAFT00000001996**	**LY6H**	**0.83**	**3.84E-02**	**90th**
319	**ENSCAFT00000026443**	**TBKBP1**	**0.71**	**3.84E-02**	**91st**
320	ENSCAFT00000037002	TGFB1	0.60	3.84E-02	Not-filtered
321	ENSCAFT00000046716	SELENOS	0.46	3.84E-02	Not-filtered
322	**ENSCAFT00000000224**		**0.21**	**3.84E-02**	**99th**
323	ENSCAFT00000046176		-1.03	3.84E-02	Not-filtered
324	ENSCAFT00000010739	ADGRD1	-0.31	3.84E-02	Not-filtered
325	ENSCAFT00000049636	LYSMD4	-0.22	3.84E-02	Not-filtered
326	ENSCAFT00000008617		0.83	3.85E-02	Not-filtered
327	**ENSCAFT00000046781**	**COTL1**	**0.69**	**3.86E-02**	**92nd**
328	ENSCAFT00000023821	SLC2A1	0.50	3.86E-02	Not-filtered
329	ENSCAFT00000035862	ADAMTS14	0.75	3.88E-02	Not-filtered
330	**ENSCAFT00000003651**	**IGFBP7**	**0.50**	**3.88E-02**	**95th**
331	ENSCAFT00000022824		0.46	3.88E-02	Not-filtered
332	ENSCAFT00000049259	EGR4	0.71	3.89E-02	Not-filtered
333	ENSCAFT00000029080	STXBP2	0.58	3.89E-02	Not-filtered
334	ENSCAFT00000030441	MYO1C	0.45	3.89E-02	Not-filtered
335	ENSCAFT00000001361	PARVG	0.67	3.90E-02	Not-filtered
336	ENSCAFT00000044118		0.68	3.92E-02	Not-filtered
337	ENSCAFT00000022956	ZNF804A	0.62	3.92E-02	Not-filtered
338	ENSCAFT00000019651	SLC37A4	0.54	3.92E-02	Not-filtered
339	ENSCAFT00000017668	ACACB	0.39	3.92E-02	Not-filtered
340	**ENSCAFT00000023142**	**ARL4D**	**0.55**	**3.93E-02**	**94th**
341	**ENSCAFT00000017502**	**SDHA**	**0.15**	**3.94E-02**	**96th**
342	ENSCAFT00000004509	SPACA6	-0.54	3.94E-02	Not-filtered
343	ENSCAFT00000017565	ITGB2	0.71	3.99E-02	Not-filtered
344	ENSCAFT00000043567	SOHLH1	1.26	4.04E-02	Not-filtered
345	ENSCAFT00000029692	SASH3	0.68	4.04E-02	Not-filtered
346	ENSCAFT00000007662	SLC25A19	0.67	4.04E-02	Not-filtered
347	ENSCAFT00000048099	MYCBPAP	0.65	4.04E-02	Not-filtered
348	ENSCAFT00000030431	DAPK3	0.61	4.04E-02	Not-filtered
349	ENSCAFT00000010812	AMOTL2	0.45	4.04E-02	Not-filtered
350	ENSCAFT00000013897	BOLA3	0.38	4.04E-02	Not-filtered
351	ENSCAFT00000036547	AMOTL2	0.36	4.04E-02	Not-filtered
352	ENSCAFT00000023491	GULP1	-0.52	4.04E-02	Not-filtered
353	**ENSCAFT00000046003**	**CAPNS1**	**0.40**	**4.06E-02**	**92nd**
354	ENSCAFT00000049527		0.64	4.07E-02	Not-filtered
355	ENSCAFT00000006254	PLEKHA4	0.72	4.07E-02	Not-filtered
356	**ENSCAFT00000006525**	**PODXL2**	**0.66**	**4.07E-02**	**94th**
357	**ENSCAFT00000031428**	**REXO4**	**0.60**	**4.07E-02**	**95th**
358	**ENSCAFT00000048442**	**PODXL2**	**0.67**	**4.07E-02**	**94th**
359	ENSCAFT00000024558	TAS1R2	0.66	4.07E-02	Not-filtered
360	**ENSCAFT00000017006**	**ASL**	**0.55**	**4.07E-02**	**94th**
361	ENSCAFT00000037539	FAM161A	-0.57	4.07E-02	Not-filtered
362	ENSCAFT00000046651	ABI3	0.72	4.08E-02	Not-filtered
363	ENSCAFT00000043135	LDB3	0.85	4.13E-02	Not-filtered

364	ENSCAFT00000047389		0.84	4.13E-02	Not-filtered
365	ENSCAFT00000025530	MMRN2	0.82	4.13E-02	Not-filtered
366	ENSCAFT00000015021	CD82	0.78	4.13E-02	Not-filtered
367	ENSCAFT00000014092	TGM2	0.71	4.13E-02	Not-filtered
368	**ENSCAFT00000009494**	**PCYT2**	**0.69**	**4.13E-02**	**91st**
369	ENSCAFT00000047372	MAFF	0.58	4.13E-02	Not-filtered
370	ENSCAFT00000046369	ROM1	0.57	4.13E-02	Not-filtered
371	ENSCAFT00000035631	MCHR1	0.43	4.13E-02	Not-filtered
372	ENSCAFT00000019881	PFKFB4	0.42	4.13E-02	Not-filtered
373	**ENSCAFT00000018410**	**MSMO1**	**0.24**	**4.13E-02**	**91st**
374	ENSCAFT00000043610	SPACA6	-0.45	4.13E-02	Not-filtered
375	ENSCAFT00000023304	LAG3	0.63	4.16E-02	Not-filtered
376	ENSCAFT00000025466	ITPKB	0.69	4.17E-02	Not-filtered
377	ENSCAFT00000048272	METTL7A	0.44	4.18E-02	Not-filtered
378	**ENSCAFT00000046422**	**ERF**	**0.74**	**4.21E-02**	**90th**
379	ENSCAFT00000023093	SH3TC1	0.69	4.21E-02	Not-filtered
380	ENSCAFT00000018146	MAPKAPK2	0.58	4.21E-02	Not-filtered
381	ENSCAFT00000032430	TPPP3	0.58	4.21E-02	Not-filtered
382	ENSCAFT00000003309	RCL1	0.28	4.21E-02	Not-filtered
383	ENSCAFT00000038030		-1.30	4.21E-02	Not-filtered
384	ENSCAFT00000011043	FAHD2A	0.52	4.22E-02	Not-filtered
385	ENSCAFT00000000248		0.82	4.22E-02	Not-filtered
386	ENSCAFT00000029863		0.48	4.22E-02	Not-filtered
387	ENSCAFT00000013320	AGA	-0.22	4.22E-02	Not-filtered
388	**ENSCAFT00000029100**	**EVI5L**	**0.70**	**4.23E-02**	**94th**
389	ENSCAFT00000044344	GSTZ1	0.45	4.23E-02	Not-filtered
390	ENSCAFT00000047011	LMAN2	0.32	4.23E-02	Not-filtered
391	ENSCAFT00000011246	GAPDHS	0.86	4.24E-02	Not-filtered
392	ENSCAFT00000015739	POLL	0.45	4.24E-02	Not-filtered
393	ENSCAFT00000046504		1.10	4.25E-02	Not-filtered
394	**ENSCAFT00000017753**	**NRGN**	**0.62**	**4.25E-02**	**99th**
395	**ENSCAFT00000026846**	**FDPS**	**0.38**	**4.25E-02**	**91st**
396	ENSCAFT00000042959	ALDH8A1	-0.86	4.25E-02	Not-filtered
397	ENSCAFT00000009996	CDC25B	-0.40	4.25E-02	Not-filtered
398	**ENSCAFT00000028891**		**0.59**	**4.26E-02**	**95th**
399	**ENSCAFT00000007432**	**APOE**	**0.81**	**4.26E-02**	**99th**
400	ENSCAFT00000001503	SERHL2	0.35	4.26E-02	Not-filtered
401	ENSCAFT00000015110	HSD17B13	-0.74	4.26E-02	Not-filtered
402	ENSCAFT00000044769		-0.67	4.26E-02	Not-filtered
403	**ENSCAFT00000043093**	**C1QB**	**0.73**	**4.28E-02**	**91st**
404	**ENSCAFT00000026481**	**BCAN**	**0.61**	**4.28E-02**	**97th**
405	**ENSCAFT00000037889**		**0.50**	**4.28E-02**	**94th**
406	ENSCAFT00000023625	BCS1L	0.43	4.28E-02	Not-filtered
407	ENSCAFT00000000089	ITGA7	0.43	4.30E-02	Not-filtered
408	ENSCAFT00000019060	PRLHR	1.06	4.31E-02	Not-filtered
409	ENSCAFT00000002101	PDGFB	0.55	4.31E-02	Not-filtered
410	**ENSCAFT00000043363**	**OAZ2**	**0.53**	**4.31E-02**	**96th**
411	ENSCAFT00000016881	ZMYND10	0.46	4.31E-02	Not-filtered
412	**ENSCAFT00000026732**	**ATP5MC1**	**0.39**	**4.34E-02**	**94th**
413	ENSCAFT00000008171	INSIG1	0.24	4.34E-02	Not-filtered
414	ENSCAFT00000015805	STMND1	-0.54	4.34E-02	Not-filtered
415	**ENSCAFT00000023619**	**MXD4**	**0.71**	**4.35E-02**	**91st**
416	ENSCAFT00000012296	ABCB9	0.50	4.35E-02	Not-filtered
417	ENSCAFT00000007368	GEMIN7	0.47	4.35E-02	Not-filtered
418	ENSCAFT00000029317	RNF135	0.45	4.35E-02	Not-filtered
419	ENSCAFT00000006898	CCDC8	0.42	4.35E-02	Not-filtered
420	ENSCAFT00000025069	DPYSL3	0.40	4.35E-02	Not-filtered
421	ENSCAFT00000020133	SELENBP1	0.33	4.35E-02	Not-filtered
422	ENSCAFT00000046911	DCDC1	-0.98	4.35E-02	Not-filtered
423	**ENSCAFT00000004280**		**0.54**	**4.37E-02**	**98th**
424	**ENSCAFT00000028305**	**STOML1**	**0.68**	**4.40E-02**	**94th**
425	ENSCAFT00000003023	MTERF1	-0.78	4.40E-02	Not-filtered
426	ENSCAFT00000043816	TMC7	-0.48	4.40E-02	Not-filtered
427	ENSCAFT00000046420		1.94	4.42E-02	Not-filtered
428	ENSCAFT00000048762		1.09	4.42E-02	Not-filtered
429	ENSCAFT00000043336	LY6H	0.78	4.42E-02	Not-filtered
430	ENSCAFT00000022144	FAH	0.77	4.42E-02	Not-filtered
431	**ENSCAFT00000048751**	**C1QA**	**0.69**	**4.42E-02**	**93rd**
432	ENSCAFT00000025974	ADGRE5	0.69	4.42E-02	Not-filtered
433	**ENSCAFT00000049194**	**CST3**	**0.65**	**4.42E-02**	**98th**
434	ENSCAFT00000015524	NEURL2	0.64	4.42E-02	Not-filtered
435	ENSCAFT00000046320	RCN3	0.63	4.42E-02	Not-filtered
436	ENSCAFT00000009007	GAA	0.60	4.42E-02	Not-filtered
437	ENSCAFT00000045724	CTU2	0.58	4.42E-02	Not-filtered

438	ENSCAFT00000049447	CCDC102A	0.56	4.42E-02	Not-filtered
439	**ENSCAFT00000029401**	**COPRS**	**0.48**	**4.42E-02**	**91st**
440	ENSCAFT00000001756	MCHR1	0.47	4.42E-02	Not-filtered
441	ENSCAFT00000008744	SLCO2B1	0.39	4.42E-02	Not-filtered
442	ENSCAFT00000007647	MRPS7	0.38	4.42E-02	Not-filtered
443	ENSCAFT00000019165		0.22	4.42E-02	Not-filtered
444	ENSCAFT00000001433	TTLL1	0.16	4.42E-02	Not-filtered
445	ENSCAFT00000021053	NRP2	-0.26	4.46E-02	Not-filtered
446	**ENSCAFT00000028942**	**CSF1R**	**0.63**	**4.48E-02**	**91st**
447	**ENSCAFT00000025423**	**COQ8A**	**0.62**	**4.48E-02**	**91st**
448	ENSCAFT00000047462	MRC2	0.58	4.48E-02	Not-filtered
449	**ENSCAFT00000004761**	**EMC10**	**0.48**	**4.48E-02**	**94th**
450	**ENSCAFT00000026830**	**LTBP2**	**2.58**	**4.49E-02**	**91st**
451	ENSCAFT00000002835	SHISA3	1.62	4.49E-02	Not-filtered
452	**ENSCAFT00000021217**	**APOD**	**1.08**	**4.49E-02**	**99th**
453	ENSCAFT00000023419	CHST7	0.83	4.49E-02	Not-filtered
454	ENSCAFT00000047446	SELPLG	0.77	4.49E-02	Not-filtered
455	ENSCAFT00000039763	TRIB3	0.75	4.49E-02	Not-filtered
456	**ENSCAFT00000008918**	**LGALS3BP**	**0.70**	**4.49E-02**	**95th**
457	ENSCAFT00000022913		0.63	4.49E-02	Not-filtered
458	ENSCAFT00000023281	AOC3	0.61	4.49E-02	Not-filtered
459	ENSCAFT00000000764	TCF19	0.61	4.49E-02	Not-filtered
460	ENSCAFT00000025475	EPHA2	0.59	4.49E-02	Not-filtered
461	**ENSCAFT00000019526**		**0.57**	**4.49E-02**	**91st**
462	ENSCAFT00000044681	ARFGEF3	0.55	4.49E-02	Not-filtered
463	**ENSCAFT00000007975**	**ACAA1**	**0.51**	**4.49E-02**	**93rd**
464	ENSCAFT00000019042	DIRC2	0.24	4.49E-02	Not-filtered
465	ENSCAFT00000028044	SERPINA3	-1.12	4.49E-02	Not-filtered
466	ENSCAFT00000011466	ZC3H8	-0.40	4.49E-02	Not-filtered
467	ENSCAFT00000030284	HSPB11	-0.34	4.49E-02	Not-filtered
468	ENSCAFT00000016183	GUSB	0.50	4.49E-02	Not-filtered
469	**ENSCAFT00000014487**	**MT2A**	**0.49**	**4.49E-02**	**98th**
470	ENSCAFT00000038426		0.80	4.50E-02	Not-filtered
471	ENSCAFT00000017630	CD99	0.47	4.50E-02	Not-filtered
472	ENSCAFT00000045007		0.36	4.50E-02	Not-filtered
473	**ENSCAFT00000034820**	**MT-ND1**	**-0.42**	**4.50E-02**	**98th**
474	ENSCAFT00000049525	PRAG1	0.65	4.51E-02	Not-filtered
475	ENSCAFT00000038192	GSDMD	0.78	4.53E-02	Not-filtered
476	ENSCAFT00000045553	NPEPL1	0.67	4.53E-02	Not-filtered
477	**ENSCAFT00000014635**	**CHRM4**	**0.63**	**4.53E-02**	**91st**
478	**ENSCAFT00000030343**	**SNRPB**	**0.56**	**4.53E-02**	**93rd**
479	ENSCAFT00000046897	EFCAB2	0.43	4.53E-02	Not-filtered
480	ENSCAFT00000050202	TRIM62	0.40	4.53E-02	Not-filtered
481	ENSCAFT00000014157	TLL1	-1.10	4.53E-02	Not-filtered
482	ENSCAFT00000010537	ZNF382	-0.88	4.53E-02	Not-filtered
483	ENSCAFT00000020156	ARHGAP30	0.54	4.53E-02	Not-filtered
484	ENSCAFT00000028142	TBX2	0.67	4.54E-02	Not-filtered
485	**ENSCAFT00000009073**	**SIRT2**	**0.61**	**4.54E-02**	**98th**
486	ENSCAFT00000027087	PLEKHO2	0.61	4.54E-02	Not-filtered
487	ENSCAFT00000048821	ITIH3	1.34	4.54E-02	Not-filtered
488	**ENSCAFT00000030037**	**DEXI**	**0.73**	**4.54E-02**	**98th**
489	ENSCAFT00000013571	MMP15	0.63	4.54E-02	Not-filtered
490	**ENSCAFT00000026233**	**GALNT16**	**0.57**	**4.54E-02**	**95th**
491	**ENSCAFT00000019198**	**MCAM**	**0.53**	**4.54E-02**	**94th**
492	ENSCAFT00000001476	WDR46	0.47	4.54E-02	Not-filtered
493	**ENSCAFT00000024053**	**CAPZB**	**0.47**	**4.54E-02**	**96th**
494	**ENSCAFT00000027758**	**SUPT4H1**	**0.38**	**4.54E-02**	**93rd**
495	ENSCAFT00000008258	LCT	-0.55	4.54E-02	Not-filtered
496	ENSCAFT00000000403	IL20RA	-0.37	4.54E-02	Not-filtered
497	ENSCAFT00000046931	HDHD2	-0.35	4.54E-02	Not-filtered
498	ENSCAFT00000011649	KRT71	1.02	4.54E-02	Not-filtered
499	ENSCAFT00000013918	WNT1	0.67	4.54E-02	Not-filtered
500	**ENSCAFT00000030389**	**CD99L2**	**0.64**	**4.54E-02**	**96th**
501	ENSCAFT00000025515	GPR173	0.49	4.54E-02	Not-filtered
502	ENSCAFT00000039444		0.38	4.54E-02	Not-filtered
503	ENSCAFT00000046451	SULF1	-0.73	4.54E-02	Not-filtered
504	ENSCAFT00000016563	DHCR7	0.60	4.54E-02	Not-filtered
505	**ENSCAFT00000027989**	**ATP5F1A**	**0.26**	**4.54E-02**	**98th**
506	ENSCAFT00000050049	DNASE2	0.59	4.57E-02	Not-filtered
507	ENSCAFT00000002073	SIL1	0.46	4.61E-02	Not-filtered
508	**ENSCAFT00000026928**		**0.54**	**4.61E-02**	**98th**
509	ENSCAFT00000044103	PKDCC	0.45	4.61E-02	Not-filtered
510	ENSCAFT00000008885	EOMES	1.19	4.61E-02	Not-filtered
511	ENSCAFT00000023001	ETV4	0.69	4.61E-02	Not-filtered
512	ENSCAFT00000021197	MRC2	0.61	4.61E-02	Not-filtered

513	**ENSCAFT00000045076**	**CEND1**	**0.59**	**4.61E-02**	**99th**
514	ENSCAFT00000013448	MOGS	0.55	4.61E-02	Not-filtered
515	**ENSCAFT00000013822**	**ADGRG1**	**0.54**	**4.61E-02**	**93rd**
516	ENSCAFT00000043384		0.53	4.61E-02	Not-filtered
517	ENSCAFT00000044371	LRCH4	0.52	4.61E-02	Not-filtered
518	**ENSCAFT00000013340**	**SCARA3**	**0.51**	**4.61E-02**	**90th**
519	ENSCAFT00000004561	ETFB	0.44	4.61E-02	Not-filtered
520	ENSCAFT00000024087	CD9	0.42	4.61E-02	Not-filtered
521	ENSCAFT00000027168	CITED1	0.39	4.61E-02	Not-filtered
522	ENSCAFT00000025961	TCAP	0.36	4.61E-02	Not-filtered
523	ENSCAFT00000029024	SLFN14	-0.64	4.61E-02	Not-filtered
524	ENSCAFT00000020162	NPAS4	0.95	4.62E-02	Not-filtered
525	ENSCAFT00000018997	TMEM136	-0.48	4.63E-02	Not-filtered
526	ENSCAFT00000009489	PCYT2	0.69	4.63E-02	Not-filtered
527	ENSCAFT00000045894		1.16	4.63E-02	Not-filtered
528	ENSCAFT00000006386	SULT2B1	1.02	4.63E-02	Not-filtered
529	ENSCAFT00000037740		0.83	4.63E-02	Not-filtered
530	ENSCAFT00000009576	DCXR	0.75	4.63E-02	Not-filtered
531	**ENSCAFT00000024555**	**WSCD1**	**0.71**	**4.63E-02**	**90th**
532	**ENSCAFT00000035428**	**IER2**	**0.70**	**4.63E-02**	**93rd**
533	**ENSCAFT00000026529**	**CPLX1**	**0.67**	**4.63E-02**	**99th**
534	ENSCAFT00000049456	SLC39A3	0.67	4.63E-02	Not-filtered
535	ENSCAFT00000018316	HAPLN3	0.67	4.63E-02	Not-filtered
536	ENSCAFT00000006533		0.67	4.63E-02	Not-filtered
537	ENSCAFT00000026538	SLC49A3	0.61	4.63E-02	Not-filtered
538	**ENSCAFT00000027930**	**YIPF2**	**0.60**	**4.63E-02**	**92nd**
539	ENSCAFT00000012188	C1H19orf12	0.59	4.63E-02	Not-filtered
540	ENSCAFT00000026821	FAM117A	0.59	4.63E-02	Not-filtered
541	ENSCAFT00000007121	VASP	0.57	4.63E-02	Not-filtered
542	ENSCAFT00000011352		0.56	4.63E-02	Not-filtered
543	ENSCAFT00000015230	CHID1	0.55	4.63E-02	Not-filtered
544	ENSCAFT00000044230	RRP1	0.54	4.63E-02	Not-filtered
545	ENSCAFT00000035983	TTC32	0.45	4.63E-02	Not-filtered
546	ENSCAFT00000000217	RBMS2	0.40	4.63E-02	Not-filtered
547	ENSCAFT00000021389	UCK2	0.38	4.63E-02	Not-filtered
548	**ENSCAFT00000020625**	**SDHC**	**0.26**	**4.63E-02**	**90th**
549	**ENSCAFT00000016300**		**0.21**	**4.63E-02**	**92nd**
550	ENSCAFT00000000234	CCDC68	-1.06	4.63E-02	Not-filtered
551	ENSCAFT00000047464	CYP1B1	-0.94	4.63E-02	Not-filtered
552	ENSCAFT00000003375		-0.87	4.63E-02	Not-filtered
553	ENSCAFT00000037896		-0.85	4.63E-02	Not-filtered
554	ENSCAFT00000010394	HPX	-0.55	4.63E-02	Not-filtered
555	ENSCAFT00000002701	BICRAL	-0.44	4.63E-02	Not-filtered
556	ENSCAFT00000018299	TARBP1	-0.41	4.63E-02	Not-filtered
557	**ENSCAFT00000039816**	**OAT**	**-0.29**	**4.63E-02**	**95th**
558	ENSCAFT00000029092	RAD51D	0.54	4.63E-02	Not-filtered
559	**ENSCAFT00000007134**	**ME3**	**0.51**	**4.63E-02**	**96th**
560	ENSCAFT00000046733	SRI	-0.23	4.63E-02	Not-filtered
561	**ENSCAFT00000031352**	**BSG**	**0.65**	**4.64E-02**	**99th**
562	ENSCAFT00000023294	SLC25A1	0.60	4.64E-02	Not-filtered
563	ENSCAFT00000029653	SLC46A1	0.48	4.64E-02	Not-filtered
564	ENSCAFT00000026293	ITGAL	0.71	4.67E-02	Not-filtered
565	**ENSCAFT00000009418**	**ASPSCR1**	**0.65**	**4.67E-02**	**96th**
566	ENSCAFT00000029224		1.11	4.67E-02	Not-filtered
567	ENSCAFT00000044447		0.61	4.67E-02	Not-filtered
568	ENSCAFT00000030888	MOB3A	0.64	4.67E-02	Not-filtered
569	ENSCAFT00000030881	PNPLA7	0.56	4.67E-02	Not-filtered
570	ENSCAFT00000021177	MRC2	0.61	4.72E-02	Not-filtered
571	ENSCAFT00000048982	CD101	-0.71	4.72E-02	Not-filtered
572	ENSCAFT00000019844	CASTOR1	0.50	4.74E-02	Not-filtered
573	ENSCAFT00000019653	SCLY	0.63	4.76E-02	Not-filtered
574	ENSCAFT00000016804	SELENOS	0.45	4.76E-02	Not-filtered
575	ENSCAFT00000020341	TCN2	0.55	4.78E-02	Not-filtered
576	ENSCAFT00000005428	STK40	0.55	4.78E-02	Not-filtered
577	ENSCAFT00000024761	NWD1	0.36	4.78E-02	Not-filtered
578	**ENSCAFT00000032066**	**SLC2A8**	**0.67**	**4.78E-02**	**92nd**
579	**ENSCAFT00000008608**	**SERTAD1**	**0.63**	**4.79E-02**	**92nd**
580	**ENSCAFT00000015047**	**GOT1**	**0.35**	**4.79E-02**	**98th**
581	ENSCAFT00000049200	TRIB2	0.18	4.79E-02	Not-filtered
582	ENSCAFT00000031228		0.78	4.80E-02	Not-filtered
583	ENSCAFT00000045026	COL17A1	0.62	4.80E-02	Not-filtered
584	ENSCAFT00000046184	CRTAC1	0.48	4.80E-02	Not-filtered
585	ENSCAFT00000014206	LYZF2	1.27	4.82E-02	Not-filtered
586	ENSCAFT00000043197		0.53	4.82E-02	Not-filtered

587	ENSCAFT00000000131	DGKA	-0.19	4.83E-02	Not-filtered
588	ENSCAFT00000044444	LMO3	-0.58	4.83E-02	Not-filtered
589	**ENSCAFT00000022063**	**VGF**	**0.68**	**4.86E-02**	**98th**
590	**ENSCAFT00000046233**	**LTBP2**	**2.55**	**4.87E-02**	**91st**
591	ENSCAFT00000042954	SOX17	0.81	4.87E-02	Not-filtered
592	**ENSCAFT00000036976**	**EVI5L**	**0.67**	**4.87E-02**	**93rd**
593	**ENSCAFT00000045135**	**PPP1CA**	**0.61**	**4.87E-02**	**93rd**
594	ENSCAFT00000002463	RECQL4	0.55	4.87E-02	Not-filtered
595	**ENSCAFT00000023617**	**CHCHD1**	**0.51**	**4.87E-02**	**91st**
596	ENSCAFT00000015184	DNMBP	-0.40	4.87E-02	Not-filtered
597	**ENSCAFT00000021221**	**APOD**	**1.03**	**4.89E-02**	**99th**
598	**ENSCAFT00000017804**	**LRP10**	**0.75**	**4.89E-02**	**90th**
599	**ENSCAFT00000029276**	**SREBF1**	**0.68**	**4.89E-02**	**90th**
600	ENSCAFT00000044529	BOP1	0.68	4.89E-02	Not-filtered
601	ENSCAFT00000048512	KIF26B	0.67	4.89E-02	Not-filtered
602	ENSCAFT00000024534	TSPAN9	0.64	4.89E-02	Not-filtered
603	**ENSCAFT00000048052**	**PODXL2**	**0.64**	**4.89E-02**	**94th**
604	**ENSCAFT00000004100**	**HSPBP1**	**0.64**	**4.89E-02**	**97th**
605	**ENSCAFT00000031601**	**PPP1R26**	**0.62**	**4.89E-02**	**96th**
606	**ENSCAFT00000022529**	**ACSBG1**	**0.61**	**4.89E-02**	**92nd**
607	**ENSCAFT00000043153**	**CLEC2L**	**0.59**	**4.89E-02**	**94th**
608	**ENSCAFT00000026430**	**CLTB**	**0.59**	**4.89E-02**	**98th**
609	**ENSCAFT00000004187**	**TTYH1**	**0.58**	**4.89E-02**	**98th**
610	ENSCAFT00000023140	DNAJC4	0.56	4.89E-02	Not-filtered
611	ENSCAFT00000020594	NARF	0.54	4.89E-02	Not-filtered
612	ENSCAFT00000007824	PROSER2	0.51	4.89E-02	Not-filtered
613	**ENSCAFT00000019330**	**HYOU1**	**0.48**	**4.89E-02**	**95th**
614	ENSCAFT00000017479	SLC37A2	0.47	4.89E-02	Not-filtered
615	ENSCAFT00000008964	LMCD1	0.43	4.89E-02	Not-filtered
616	ENSCAFT00000031424	DENND2D	0.42	4.89E-02	Not-filtered
617	ENSCAFT00000031667	KLHDC4	0.37	4.89E-02	Not-filtered
618	ENSCAFT00000016873	HSF2BP	0.24	4.89E-02	Not-filtered
619	**ENSCAFT00000015156**		**0.23**	**4.89E-02**	**93rd**
620	ENSCAFT00000032485	SLC44A5	-0.69	4.89E-02	Not-filtered
621	ENSCAFT00000010536	LMNTD2	-0.68	4.89E-02	Not-filtered
622	ENSCAFT00000047719	HTR7	-0.44	4.89E-02	Not-filtered
623	ENSCAFT00000010002	CDC25B	-0.40	4.89E-02	Not-filtered
624	ENSCAFT00000026258	SUSD6	0.44	4.89E-02	Not-filtered
625	ENSCAFT00000035327	NR2E1	-0.48	4.90E-02	Not-filtered
626	ENSCAFT00000023033	GNB1L	0.59	4.90E-02	Not-filtered
627	**ENSCAFT00000014776**	**CRTAC1**	**0.56**	**4.91E-02**	**90th**
628	**ENSCAFT00000023116**	**TPI1**	**0.35**	**4.94E-02**	**99th**
629	ENSCAFT00000046346	PIGN	0.28	4.94E-02	Not-filtered
630	ENSCAFT00000048429		0.53	4.95E-02	Not-filtered
631	ENSCAFT00000046568	CEACAM23	0.83	4.95E-02	Not-filtered
632	ENSCAFT00000002737	CHRNA5	0.39	4.97E-02	Not-filtered
633	ENSCAFT00000023141	DNAJC4	0.50	4.98E-02	Not-filtered
634	ENSCAFT00000016048	HMGCS2	1.64	4.98E-02	Not-filtered
635	ENSCAFT00000031156	SSTR5	0.93	4.98E-02	Not-filtered
636	ENSCAFT00000008126		0.72	4.98E-02	Not-filtered
637	**ENSCAFT00000013911**	**DDN**	**0.67**	**4.98E-02**	**98th**
638	ENSCAFT00000011216	AKIP1	0.61	4.98E-02	Not-filtered
639	ENSCAFT00000046490	FN3K	0.48	4.98E-02	Not-filtered
640	ENSCAFT00000029547	TRIP10	0.46	4.98E-02	Not-filtered
641	ENSCAFT00000011194	ST5	0.39	4.98E-02	Not-filtered
642	**ENSCAFT00000026747**	**TRAPPC1**	**0.34**	**4.98E-02**	**93rd**
643	ENSCAFT00000017922	ZNF639	-0.27	4.98E-02	Not-filtered
644	ENSCAFT00000023373	MACROD1	0.69	4.99E-02	Not-filtered
645	ENSCAFT00000017596	ROBO4	0.64	4.99E-02	Not-filtered
646	ENSCAFT00000020445	FCER1G	0.54	4.99E-02	Not-filtered
647	ENSCAFT00000002605	IMPDH1	0.50	4.99E-02	Not-filtered
648	ENSCAFT00000001193	PPT2	0.45	4.99E-02	Not-filtered
649	ENSCAFT00000043269		0.28	4.99E-02	Not-filtered
650	**ENSCAFT00000002307**	**LGALS1**	**0.74**	**5.00E-02**	**92nd**
651	ENSCAFT00000015776	ADCY7	0.37	5.00E-02	Not-filtered
	ENSCAFT00000005173	SPRED2			Added 97th. 95th.90th
	ENSCAFT00000023941	MED28			Added 97th. 95th.90th
	ENSCAFT00000046536				Added 97th. 95th.90th
	ENSCAFT00000019347				Added 97th. 95th.90th
	ENSCAFT00000035659	TINAGL1			Added 97th. 95th.90th
	ENSCAFT00000020451	TMEM50A			Added 97th. 95th
	ENSCAFT00000026776	PHOSPHO1			Added 97th. 95th.90th
	ENSCAFT00000046387	FBXO17			Added 97th. 95th.90th
	ENSCAFT00000022306	ACHE			Added 97th. 95th
	ENSCAFT00000025894	PKN1			Added 97th. 95th

ENSCAFT00000046295	CBX2	Added 97th. 95th.90th
ENSCAFT00000028935	LHFPL1	Added 97th. 95th.90th
ENSCAFT00000005896	PREP	Added 97th. 95th.90th
ENSCAFT00000012109	ANKRD37	Added 97th. 95th.90th
ENSCAFT00000019616	HMBS	Added 97th. 95th.90th
ENSCAFT00000000818	SNCAIP	Added 97th. 95th.90th
ENSCAFT00000048703	DENND10	Added 97th. 95th.90th
ENSCAFT00000017817	UNC93B1	Added 97th. 95th.90th
ENSCAFT00000023184	TNS1	Added 97th. 95th.90th
ENSCAFT00000002969	CKAP4	Added 97th. 95th.90th
ENSCAFT00000001416	TSPO	Added 97th. 95th.90th
ENSCAFT00000001830	TRAPPC9	Added 97th. 95th
ENSCAFT00000030449	PNMA6A	Added 97th. 95th
ENSCAFT00000032043	NIBAN2	Added 97th. 95th.90th
ENSCAFT00000009379	SRA1	Added 97th. 95th.90th
ENSCAFT00000047650	ACOT8	Added 97th. 95th.90th
ENSCAFT00000027145	TOB1	Added 97th. 95th.90th
ENSCAFT00000045343	TARBP1	Added 97th. 95th.90th
ENSCAFT00000018419	PPP1CA	Added 97th. 95th
ENSCAFT00000003571	SUSD3	Added 97th. 95th.90th
ENSCAFT00000016049	TEX264	Added 97th. 95th
ENSCAFT00000038320	RNH1	Added 97th
ENSCAFT00000001412	TTLL12	Added 97th. 95th
ENSCAFT00000014664	PI4K2A	Added 97th. 95th.90th
ENSCAFT00000016906	CSTB	Added 97th. 95th
ENSCAFT00000009602	ZNF268	Added 97th. 95th.90th
ENSCAFT00000044467	GIMD1	Added 97th. 95th.90th
ENSCAFT00000023290	CTDSP1	Added 97th. 95th.90th
ENSCAFT00000021383	DGKI	Added 97th. 95th.90th
ENSCAFT00000047216		Added 97th. 95th.90th
ENSCAFT00000002360	KCTD17	Added 97th
ENSCAFT00000030093	TICAM1	Added 97th. 95th.90th
ENSCAFT00000031657	NTNG2	Added 97th. 95th.90th
ENSCAFT00000009186	CHMP6	Added 97th
ENSCAFT00000028660	SNX33	Added 97th. 95th.90th
ENSCAFT00000002363	TST	Added 97th.95th
ENSCAFT00000007160	FOSB	Added 97th. 95th.90th
ENSCAFT00000008360	CAPN5	Added 97th.95th
ENSCAFT00000014711	GOLGA7B	Added 97th
ENSCAFT00000023163	CPZ	Added 97th. 95th.90th
ENSCAFT00000023453	HSPG2	Added 97th. 95th.90th
ENSCAFT00000026936	BORCS6	Added 97th. 95th.90th
ENSCAFT00000029761	CAPS	Added 97th.95th
ENSCAFT00000029769	NRTN	Added 97th. 95th.90th
ENSCAFT00000035590	NRTN	Added 97th. 95th.90th
ENSCAFT00000044415	ADGRA2	Added 97th. 95th.90th
ENSCAFT00000008800	MTMR14	Added 97th. 95th.90th
ENSCAFT00000013256	DOK1	Added 97th. 95th.90th
ENSCAFT00000014170	PRADC1	Added 97th. 95th.90th
ENSCAFT00000014432	PLBD2	Added 97th
ENSCAFT00000024336	SLC38A5	Added 97th. 95th.90th
ENSCAFT00000028320	MRM1	Added 97th. 95th.90th
ENSCAFT00000028651	SIN3A	Added 97th. 95th.90th
ENSCAFT00000036757	TSPAN13	Added 97th.95th
ENSCAFT00000043367	TRAPPC6A	Added 97th. 95th.90th
ENSCAFT00000043331	NCKAP1L	Added 97th. 95th.90th
ENSCAFT00000031387		Added 97th. 95th.90th
ENSCAFT00000019530	CERS2	Added 97th.95th
ENSCAFT00000019209	CIB1	Added 97th.95th
ENSCAFT00000030547	BCAP31	Added 97th
ENSCAFT00000025240	RGR	Added 97th. 95th.90th
ENSCAFT00000000313	SHMT2	Added 97th. 95th.90th
ENSCAFT00000011116	ETV2	Added 97th.95th
ENSCAFT00000043906	H2BC18	Added 97th. 95th.90th
ENSCAFT00000044751	TRIOBP	Added 97th. 95th.90th
ENSCAFT00000002342	MFNG	Added 97th. 95th.90th
ENSCAFT00000005111	COL17A1	Added 97th. 95th.90th
ENSCAFT00000000694	PPP1R18	Added 97th. 95th.90th
ENSCAFT00000013323	NCKAP5L	Added 97th. 95th.90th
ENSCAFT00000011300	ID1	Added 97th. 95th.90th
ENSCAFT00000030472	SLC43A2	Added 97th. 95th.90th
ENSCAFT00000002700	TOM1	Added 97th. 95th.90th
ENSCAFT00000010865	PRAG1	Added 97th. 95th.90th
ENSCAFT00000023210	VAT1	Added 97th. 95th

ENSCAFT00000043059	SREBF1	Added 97th. 95th.90th
ENSCAFT00000005467	PDIA4	Added 97th. 95th
ENSCAFT00000014352	DDX54	Added 97th. 95th.90th
ENSCAFT00000019028	KIF7	Added 97th. 95th.90th
ENSCAFT00000027222	GSTZ1	Added 97th. 95th.90th
ENSCAFT00000031817	S1PR1	Added 97th. 95th.90th
ENSCAFT00000025531	MMRN2	Added 97th. 95th.90th
ENSCAFT00000027110	DNASE2	Added 97th. 95th.90th
ENSCAFT00000019200	DENND10	Added 97th. 95th.90th
ENSCAFT00000014281	HERPUD1	Added 97th. 95th.90th
ENSCAFT00000002231	SOX10	Added 95th
ENSCAFT00000048807	KIF19	Added 97th. 95th.90th
ENSCAFT00000005648	ASTN2	Added 97th. 95th.90th
ENSCAFT00000019797	RPS6KA1	Added 97th. 95th.90th
ENSCAFT00000026182	PLOD1	Added 97th. 95th.90th
ENSCAFT00000029231	NUDT14	Added 95th
ENSCAFT00000023969	ITIH3	Added 97th. 95th.90th
ENSCAFT00000030863	ARHGAP4	Added 97th. 95th.90th
ENSCAFT00000045788	LMNA	Added 97th. 95th.90th
ENSCAFT00000026478	EFNB3	Added 97th. 95th.90th
ENSCAFT00000030174	CARHSP1	Added 97th. 95th.90th
ENSCAFT00000046454	SCAMP4	Added 97th. 95th.90th
ENSCAFT00000006555	EHD2	Added 97th. 95th.90th
ENSCAFT00000049535	TARBP2	Added 97th. 95th.90th
ENSCAFT00000002299	TRIOBP	Added 97th. 95th.90th
ENSCAFT00000007452	NECTIN2	Added 97th. 95th.90th
ENSCAFT00000039596	TSPAN11	Added 97th. 95th.90th
ENSCAFT00000038921	JPH2	Added 97th. 95th.90th
ENSCAFT00000030963	ARHGEF16	Added 97th. 95th.90th
ENSCAFT00000023556	MYOZ1	Added 97th. 95th.90th
ENSCAFT00000050110		Added 97th. 95th.90th
ENSCAFT00000020548	OGFOD3	Added 97th. 95th.90th
ENSCAFT00000019412	TRIM26	Added 97th. 95th.90th
ENSCAFT00000025645	EPHX1	Added 97th. 95th.90th
ENSCAFT00000025232	CCDC22	Added 97th. 95th.90th
ENSCAFT00000019049	KREMEN1	Added 97th. 95th.90th
ENSCAFT00000017710	INPP5A	Added 97th. 95th.90th
ENSCAFT00000006272	BCAT2	Added 97th. 95th.90th
ENSCAFT00000047869	PER2	Added 97th. 95th.90th
ENSCAFT00000022273	LARS2	Added 97th. 95th.90th
ENSCAFT00000047037	TPH1	Added 95th
ENSCAFT00000003181	SPATA18	Added 95th
ENSCAFT00000028109	EEF2K	Added 95th
ENSCAFT00000017991	SELPLG	Added 97th.95th.90th
ENSCAFT00000011130	AACS	Added 95th
ENSCAFT00000049304		Added 95th
ENSCAFT00000014535	CHST14	Added 97th. 95th.90th
ENSCAFT00000045470	HSPG2	Added 97th. 95th.90th
ENSCAFT00000024856	STK11IP	Added 97th. 95th.90th
ENSCAFT00000024922	EVC2	Added 97th. 95th.90th
ENSCAFT00000001166	TNXB	Added 97th. 95th.90th
ENSCAFT00000046665	TNXB	Added 97th. 95th.90th
ENSCAFT00000048328	C17orf113	Added 97th. 95th.90th
ENSCAFT00000030678		Added 95th
ENSCAFT00000029702	UNC119	Added 97th. 95th.90th
ENSCAFT00000026151	IL27RA	Added 95th
ENSCAFT00000005227		Added 95th
ENSCAFT00000037329	HCK	Added 97th. 95th.90th
ENSCAFT00000022140	TM7SF2	Added 97th. 95th.90th
ENSCAFT00000002028	DEF6	Added 97th. 95th.90th
ENSCAFT00000011738	ATG101	Added 95th
ENSCAFT00000007181	NDUFB6	Added 95th
ENSCAFT00000032147	TAT	Added 95th
ENSCAFT00000008578	GALNT14	Added 90th
ENSCAFT00000022979	PTPN6	Added 90th
ENSCAFT00000019676	CDC42SE1	Added 90th
ENSCAFT00000022238	LIMD1	Added 90th

Supplementary Table 6.9. Correlation and outliers. Correlation values (r^2) and the root mean square error (RMSE) from the regression analysis between the final dispersion estimates and the mean of normalized counts for both case studies, wolves and dogs, and aggressive and tame foxes. The number of outliers identified by DESEQ2 are also presented. Values are shown for the non-filtered (NF) and all the filtered datasets used in differential expression analysis.

	Wolves and dogs			**Aggressive and tame foxes**		
Percentile	Outliers	r^2	RMSE	Outliers	r^2	RMSE
NF	281	0.74	0.81	20	0.49	1.12
99	262	0.76	0.79	20	0.5	1.11
98	249	0.77	0.77	19	0.51	1.11
97	233	0.78	0.76	18	0.52	1.1
96	219	0.78	0.76	18	0.53	1.1
95	218	0.79	0.75	17	0.54	1.09
94	205	0.8	0.74	16	0.55	1.09
93	198	0.8	0.73	15	0.55	1.09
92	186	0.81	0.73	15	0.56	1.08
91	179	0.81	0.72	14	0.57	1.08
90	172	0.82	0.72	14	0.58	1.07
85	141	0.83	0.7	13	0.61	1.05
80	123	0.85	0.68	13	0.65	1.03
75	118	0.86	0.67	14	0.68	1
70	102	0.87	0.65	15	0.72	0.97

Supplementary Table 6.10. Shared genes and gene families between non-filtered datasets. List of the gene families, and shared genes, that were commonly regulated in dogs and tame foxes, using the non-filtered datasets. The number, and name, of the genes within each gene family are provided, with the corresponding log2fold-change values in brackets for each species. Within each family, single genes were charecterized as shared between dogs and tame foxes, or as exclusive to each of the two groups. When more than one transcript for a specific gene was present, all the log2FC values are reported.

Gene Family	Group	N genes	Gene name and log2FC value
Up-regulated genes			
Cholinergic receptor nicotinic alpha	**Shared**	1	**CHRNA5** (1.11 in dogs, 0.39 in tame foxes)
Squalene epoxidase	**Shared**	1	**SQLE** (0.55 in dogs, 0.29 in tame foxes)
Rho GTPase activating protein	**Shared** Tame fox	1 1	**ARHGAP25** (0.86 in dogs, 0.70 in tame foxes) ARHGAP30 (0.54)
Integrin alpha subunits	Dog Tame fox **Shared**	3 1 1	ITGA6 (1.26. 1.25); ITGA8 (1.14. 0.91); ITGAX (0.97) ITGAL (0.71) **ITGA7** (0.76 in dogs, 0.47 and 0.43 in tame foxes)
Myosin	Dog Tame fox **Shared**	1 2 1	MYO3A (1.12) MYO1F (0.90); MYO1C (0.45) **MYO7A** (0.82 in dogs, 0.39 in tame foxes)
Tribbles pseudokinase	Tame fox **Shared**	2 1	TRIB1 (0.91); TRIB3 (0.75) **TRIB2** (0.62 in dogs, 0.18 in tame foxes)
EF hand calcium binding	Dog Tame fox	1 1	EFCAB1 (2.6) EFCAB2 (0.43)
Transcription factor	Dog Tame fox	1 1	TCF23 (2.05) TCF19 (0.61)
Adhesion G protein-coupled receptors	Dog Tame fox	1 1	ADGRG6 (1.46) ADGRG1 (0.54)
Patatin Like Phospholipase Domain	Dog Tame fox	1 1	PNPLA4 (1.42) PNPLA7 (0.56)
SRY-box	Dog Tame fox	1 1	SOX6 (1.27) SOX17(0.81)
Hyaluronan and proteoglycan link protein	Dog Tame fox	1 1	HAPLN1 (1.16) HAPLN3 (0.67)
Serine/threonine kinase	Dog Tame fox	2 1	STK17A (1.16. 1.14); STK32A (1.11) STK40 (0.55)
Potassium channels	Dog Tame fox	1 1	KCTD16 (0.98) KCTD15 (0.69)
Podocalyxin like	Dog Tame fox	1 1	PODXL (0.96. 0.85) PODXL2 (0.77. 0.66. 0.64)
ATP binding cassette subfamily B	Dog Tame fox	1 1	ABCB1 (0.93) ABCB9 (0.50)
Zinc finger DHHC-type	Dog Tame fox	1 1	ZDHHC15 (0.76) ZDHHC1 (0.67)
Sushi domain	Dog Tame fox	1 1	SUSD1 (0.69) SUSD6 (0.44)
TBC1 domain family	Dog Tame fox	1 1	TBC1D5 (0.55) TBC1D7 (0.25)
Mitogen-activated protein kinase kinase kinases	Dog Tame fox	1 1	MAP3K5 (0.51) MAP3K11 (0.72)
Spermatogenesis Associated	Dog Tame fox	1 1	SPATA5 (0.88) SPATA24 (0.51)
Down-regulated genes			
Stathmin domain	**Shared**	1	**STMND1** (-1.18 in dogs, -0.54 in tame foxes)
Oligoadenylate synthetase like	**Shared**	1	**OASL** (-0.40 in dogs, -0.54 in tame foxes)
Heat shock protein family B	Dog Tame fox	1 1	HSPB8 (-0.69) HSPB11 (-0.34)

Chapter VII.

General Discussion

This dissertation presents an integrative approach, combining genomics with gene expression data, used to uncover the evolutionary trajectory of wolves living in highly human-dominated landscapes, using the Iberian wolf as a case study. Support for distinct evolutionary mechanisms that might be operating to allow adaptation to the fast human-induced changes was uncovered. Within Chapter III, we showed that the Iberian wolf genome is enriched with genetic variants from the domestic dog, which display signatures of positive selection in Iberian wolves. Given the association of introgressed variants with a gene, MAST4, involved with several neurological processes in model vertebrate species, we believe dog introgression could be associated with behavioral responses toward humans and their activities. In parallel, in Chapter IV it was shown that several genes associated with traits linked with the domestic behavioral phenotype demonstrated clear differences in expression levels within the frontal cortex of Iberian wolves when compared with other wolf populations. Some of these genes had similar expression patterns to those found in dogs, which has been previously associated with reduced anxiety, increased cognitive abilities, and reduced exploratory behavior. This result could be the first indication of a self-domestication process in Iberian wolves, as a result of strong human pressure. In Chapter V, we provided an unprecedented view into the historical genetic composition of Iberian wolves, showing how population decline can affect hybridization dynamics and genetic diversity. Both genetic diversity and levels of hybridization with domestic dogs varied according to distinct population range dynamics and both were associated with demographic scenarios of population contraction and expansion. Ancestry analysis of the Iberian wolf population, both historical and contemporary, in Chapters III and V, revealed remarkably low levels of dog admixture (< 1%) across the wolf genome, confirming that Iberian wolves remain clearly differentiated from domestic dogs despite events of hybridization between the two. Finally, in this dissertation we also provided important genomic resources, such as the capture array we have developed and validated for historical samples in Chapter II, allowing the integration with samples containing high-quality DNA; as well as a software tool, TVscript, that can be used to quantify levels of intra-condition count variation and remove

hypervariable transcripts from a reference set before performing differential expression analysis (Chapter VI). We showed that the maintenance of hyper variable transcripts in the reference set has a negative impact on the detection of differentially expressed genes due to an inevitable level of ambiguity in determining expression levels.

7.1. Introgression as a driver of adaptation of wolves to human-dominated landscapes

Pervasive events of gene flow have been described within the genus *Canis*, at both ancient and contemporary scales, including genetic contribution from as-yet-unrecognized lineages (Gopalakrishnan et al. 2018; Wang et al. 2020). Complex and intertwined histories of interbreeding have revealed a dynamic exchange of genes between wolves and dogs following divergence (Freedman et al. 2014; Fan et al. 2016; Bergström et al. 2022). Such a rapid and wide availability of genomic content from a closely related species can lead to a wholesale change in a few generations, and promote opportunity for adaptation to new environments (Grant and Grant 2019). Adaptive introgression in canids has proven to be both responsible for the adaptation of Tibetan gray wolves and Tibetan Mastiff dogs to hypoxia in high altitudes (Wang et al. 2020), as well as for the increased immunological response of black coat color wolves in North America (Schweizer et al. 2018). This made us hypothesize it could also be a powerful mechanism driving adaptation of wolves to human-dominated landscapes.

The Iberian wolf population, as we now concluded, represents an excellent choice to study this “introgression-fueled adaptation” hypothesis. Wolves and dogs have co-occurred in the Iberian Peninsula for thousands of years, and we knew that hybridization with domestic dogs had occurred in the past (Gómez-Sánchez et al. 2018) and was still ongoing (Godinho et al. 2011). Most importantly, domestic dogs, as the first animal species to be domesticated, possess remarkable adaptations to live in close proximity to humans. Therefore, under high selection pressure to survive in densely human populated areas and landscapes profoundly altered by humans and agricultural practices, Iberian wolves could profit from having specific behavioral traits found in domestic dogs.

Uncovering the genetic composition of species over multiple timescales offers a unique opportunity to understand how distinct evolutionary mechanisms, such as hybridization and adaptive introgression, could have affected species’ history (Grabenstein and Taylor

2018). To this end, integrating ancient and/or historical samples with contemporary data is fundamental to reconstruct the evolutionary histories of species (Der Sarkissian et al. 2015; Card et al. 2021). As such, we sampled a large number of Iberian wolf specimens contained in NHCs to be able to perform a complementary approach where historical and contemporary data could be compared. However, we understood that i) clear guidelines on how to sample NHCs were missing within the scientific community, and ii) genotype historical and contemporary samples in a comparable manner still poses several methodological challenges (e.g., DNA in historical samples is often of low quality and quantity and presents shorter fragments, and high levels of contamination are also expected; Raxworthy and Smith 2021). We addressed these limitations in Chapter II, by showing that skin tissues should be prioritized when collecting samples from mammalian carnivores NHCs, given the increased amount of endogenous DNA they provided and the minimal impact on the general appearance of the specimens, in relation to bone samples; and, we have also designed and made publicly available a capture array to sequence a set of 100,000 genome-wide SNPs present in the commercial Illumina Canine HD Beadchip. Following this genotyping technique, we showed that it is possible to genotype the same set of SNPs across samples with low- and high-quality DNA. This hybrid approach has allowed us to incorporate and directly compare historical and contemporary samples, which turned out to be fundamental for analyzing introgression patterns (Chapter III), rates of wolf-dog hybridization and levels of genetic diversity following the 20th-century bottleneck (Chapter V). We hope this methodological work will help researchers design more effective sampling strategies of NHCs, which ultimately, will also positively impact the management of these precious biological resources. Additionally, we would like to point that the target capture array is not restricted to historical samples, as it can be also applied to non-invasive or environmental samples, where DNA is usually of low quality.

Profiting from this historical-contemporary integrative approach, in Chapter III we uncovered patterns of dog introgression in chromosome 2 of ~30% of the analyzed Iberian wolves. Our population genomic approach also benefited largely from the extensive collection of dog and Iberian wolf samples present at the host institute, CIBIO-InBIO/BIOPOLIS, collected over many years and that cover the entire wolf range, as well as from publicly available datasets of other European wolves (Stronen et al. 2015) and purebred dogs (Lequarré et al. 2011; Vaysse et al. 2011) genotyped for the same set of SNPs using the Canine HD BeadChip from Illumina. The idea was to exhaustively sample the genetic diversity present in dogs in order to attempt to find the donor of

introgression. The European wolves were used as a control population to validate the singularity of the genomic patterns found in Iberian wolves. Local ancestry analysis allowed us to map the ancestry of the Iberian wolf genome to either wolf or dog origin. Our results confirmed the documented genetic uniqueness of Iberian wolves in relation to other European wolf populations (Silva et al. 2020) and high levels of differentiation with domestic dogs. Despite the presence of small dog blocks stochastically distributed across the genome, we did not find any pattern like the one in chromosome 2 across the remaining chromosomes nor across the analyzed European wolves. All wolves with the introgression in chromosome 2 belong to the Northwestern population that has persisted in Iberia, even during the 1970s when the population has reached its minimum (Valverde 1971; Garzón 1979; Blanco et al. 1990; Rico and Torrente 2000; Clavero et al. 2022; Nores and López-Bao 2022). No evidence of this introgressed dog block was found in wolves from the South of Iberia, where the population went extinct. This finding might suggest that introgressed dog variants enhanced the adaptive potential of Iberian wolves in the North allowing their persistence, while in the South, wolves without such genetic contribution from dogs were unable to adapt. However, this idea needs further testing as our historical data does not allow us to measure different adaptive potentials between individuals in the North and South. Sampling a large number of individuals in the South would be required, as well as having samples from periods predating the introgression (estimated date of ca. 3.6 ka), to better understand its geographical origin and further expansion. Future genomic studies analyzing ancient Iberian wolf genomes could provide fundamental insights into this question.

The technique we used to genotype ~170,000 genome-wide SNPs across several Iberian wolf and dog samples, was revealed to be useful to provide a genome-wide perspective and detect general patterns of introgression, at a relatively low cost. However, the low SNP density across the genome (one SNP every 25 kb, on average) provided a low resolution over the introgressed region. Sequencing the entire genome of some Iberian wolves at moderate coverage (13.5x), offered a much-increased resolution and uncovered a haplotype of 90 kb inside the introgressed region shared by Iberian wolves with the introgression and village dogs from the Iberian Peninsula. The level of sequence identity observed between wolves and dogs across the 90 kb haplotype ruled out a scenario of incomplete lineage sorting and instead lead us to believe that only introgression can explain such a pattern of diversity. The fact that dogs in Iberia (village dogs but also purebreds such as the Portuguese water dog and Portuguese Podengo) are the ones sharing high sequence similarity in this haplotype with Iberian wolves indicates this event took place in Iberia Peninsula long after the split

of wolves and dogs. Still, determining the direction of admixture can be difficult, as most methods developed to detect introgression do not define the donor and recipient (Wang et al. 2020). In our work, we are confident that introgression was from dogs to wolves, as the haplotype in dogs carries a higher number of polymorphisms, whereas in wolves, it only differs by a single SNP. The introgressed 90 kb haplotype is positioned within an intron of a 5' UTR region of the MAST4 gene. Although we could not directly estimate the functional impact of this gene, there is evidence that MAST4 is associated with hippocampal volumes (Hibar et al. 2017) and juvenile myoclonic epilepsy (Steffens et al. 2012) in humans. MAST4 was also found to be differentially expressed in the prefrontal cortex of atypical cases of frontotemporal lobar degeneration (Martins-de-Souza et al. 2012), and knockout *Drosophila* showed enhanced predisposition to develop Alzheimer's disease (Shulman et al. 2014). A recent study (Satizabal et al. 2019) has shown that in human brain structures, MAST4 interacts with DLG2, which encodes a protein involved in the activation of NMDA glutamate receptors and postsynaptic events (GeneCards). Given the association with several neurocognitive processes, as well as the multiple evidence of selection found in our work (including demographic simulations of neutral vs selection scenarios), it is possible to assume that the introgressed variants in MAST4 played an important role in shaping some of the behavioral traits found in Iberian wolves, which could have allowed their adaptation to human-dominated landscapes. Furthermore, in Chapter IV we showed that MAST4 is down-regulated in the brain of Iberian wolves and dogs in relation to non-Iberian wolves, which could imply a role of the introgressed variants at the 5'UTR intron in gene regulation. In fact, 5'UTR introns have been proposed to be of higher significant functional consequence for the regulation of mammalian genes than was previously thought (Bicknell et al. 2012).

Importantly, despite the volume of circumstantial evidence pointing to the role of dog introgression in the adaptation of Iberian wolves, convincingly proving that introgression offers enhanced adaptive potential is extremely challenging. First, we do not have a clear and obvious phenotype, such as coat color (Anderson et al. 2009), or hypoxia-adapted (Miao et al. 2017; Wang et al. 2020) to associate with introgressed genetic variants. Furthermore, behavioral phenotypes are often underpinned by complex genetic interactions, as is evident from extensive research done on domestic animals (e.g., Carneiro et al. 2014; Li et al. 2014; Hou et al. 2020; Sato et al. 2020), which hampers the identification of a clear genetic basis. Additionally, despite the observed intermediate frequency of the introgression, the entire current population of Iberian wolves exhibits the same behavioral traits, which may indicate that other introgressed signatures

associated with the 90 kb haplotype, such as structural variation, might be present at a higher frequency but remained undetected due to the limited resolution of our datasets. It is also likely that other evolutionary mechanisms and ecological and social-environmental factors have contributed to the emergence of the behavioral traits in Iberian wolves. Finally, unlike model vertebrate species, wolves living in the wild cannot be easily incorporated into controlled experimental designs to estimate measures of fitness. This is mostly associated with the challenges of surveying and capturing wolves (Blanco and Cortés 2012), as well as with the high mortality rate reported for wolves in the Iberian Peninsula (Campos 2018) which makes it difficult to monitor individuals for long periods of time.

The results of this work reflect one of the few reported cases of adaptive introgression between dogs and wolves, and we believe it will contribute to advancing scientific knowledge on the role of introgression from domestic to wild species in an increasingly human-changed world.

7.2. Evidence for self-domestication in Iberian wolves

In parallel to the “introgression-fueled adaptation” hypothesis, we also proposed that adaptation to human-dominated landscapes would have triggered a process of self-domestication in Iberian wolves. Self-domestication was initially described in bonobo apes to refer to behavioral features, such as less aggression, that were naturally selected in wild animals and that resemble those found in domestic animals (Hare et al. 2012). However, this process does not refer to a process of domestication itself, which involves human intervention. In wolves, self-domestication would be applied by favoring behaviors to co-exist with humans, such as those described in Iberian wolves. Given the growing body of literature supporting that behavioral phenotypic differentiation between wild and domestic animals is mostly underlined by gene regulation rather than variation in the protein-coding sequence (Li et al. 2014; Sato et al. 2020), we thought that a reasonable way to approach this hypothesis would be to study gene expression profiles in the brain of Iberian wolves. Unlike dogs, for whom there is a quite diverse amount of RNA-seq data available for several tissues/organs, the availability of such data for wolves living in their natural habitats was to date non-existent. There is, however, transcriptomic data from the frontal cortex of six gray wolves kept in captivity (Albert et al. 2012). Such data scarcity is associated with the unstable nature of the RNA molecule which has a much faster degradation rate than DNA, requiring its collection and

preservation almost immediately after the animals' death (Wang et al. 2009). This impedes the collection of biological samples for RNA isolation from animal carcasses, which usually, are the most abundant source of biological samples for wolves. In this work, we took advantage of a population density control program occurring in Asturias (Spain), to collect samples immediately after wolves were killed (after February of 2021 wolves are fully protected in Spain). Despite our efforts to collect samples from wolves living in wilder environments with lower human density, namely from those found in Slovenia or Estonia, several bureaucratic and practical constraints undermined the possibility of applying a proper case-control comparison (wolves living in human-dominated landscapes vs. wild environments without human disturbance). As such, we compared gene expression profiles in the brain of Iberian wolves with those already publicly available for non-Iberian wolves kept in captivity (Albert et al. 2012), as well as for dogs (Broad Institute 2019) to have a background reference for the domestic behavioral phenotype.

Knowing that sources of unwanted variation can emerge in RNA experiments even under highly controlled environments and that intra-condition variation can be greatly increased within datasets that were generated in different studies (Hansen et al. 2011; Conesa et al. 2016), we developed a new software tool, TVscript (Chapter VI), to aid in the analysis of RNA-seq data from Iberian wolves, non-Iberian wolves, and dogs. In Chapter VI, we showed the effectiveness of TVscript in estimating levels of intra-condition variation, which can be present even within datasets generated in the same study. Moreover, the removal of transcripts associated with high levels of intra-condition count variation with TVscript proved to increase the sensitivity of differential expression analysis. We demonstrated the usefulness of this tool on simulated data, but also in a case study involving publicly available RNA-seq data from the brain of wolves (Albert et al. 2012), dogs (Albert et al. 2012; Roy et al. 2013; Hoeppner et al. 2014; Fushan et al. 2015), and tame and aggressive foxes from the farm-fox experiment (Wang et al. 2018). This combined analysis of expression profiles in the brain of tamed canids advanced our knowledge of candidate genes associated with the tame behavior, which was fundamental for interpreting the expression profiles in the Iberian wolf's brain (Chapter IV). We encourage the use of TVscript, or the general development of similar approaches, prior to performing differential expression analysis in order to aid in the presentation of end results, consisting of lists of differentially expressed genes, through the deciphering of ambiguity associated with intra-condition variation on the threshold p-values applied. To date, the context of such ambiguity is often poorly discussed, despite

the importance of such results. Additionally, we hope that our tool will be a small step forward in helping researchers take advantage of the growing repertoire of publicly available RNA-seq datasets.

In Chapter IV, by using publicly available data of non-Iberian wolves to perform differential expression analysis with Iberian wolves, we found expressional changes in several genes associated with the domestic behavioral phenotype (i.e., genes displaying convergent signatures of selection and/or gene expression in several domestic animals), especially those involved in the neurotransmission of serotonin and glutamate. In Chapter III we had already highlighted the role of glutamate receptors in the possible adaptation of Iberian wolves given the association of the introgressed MAST4 with the activation of NMDA glutamate receptors. Although this overlap *per se* does not prove evidence for convergent selection on glutamate genes, we believe that it is strongly suggestive that this neurotransmitter played an important role in the evolution of the behavioral traits found in Iberian wolves in response to anthropogenic pressure. For example, domestication-related behaviors that have been associated with glutamate genes, such as reduced anxiety, harm avoidance, and increased cognitive abilities (O'Rourke and Boeckx 2020), could explain the increased tolerance of Iberian wolves towards humans and their reduced levels of chronic stress. In addition, when we looked for genes displaying overlap signatures of allele frequency changes and differential expression, we also found two genes associated with neurological processes, including GPRC5C that belongs to the glutamate receptor subfamily (GeneCards). This could represent the first evidence that the behavioral traits found in Iberian wolves conform, at least partially, to a process of self-domestication mediated by glutamate genes. But again, this hypothesis will require extensive additional testing to be conclusive about the functional impact of these genes. Future studies on domestic animals could provide helpful insights into the exact role of glutamate genes in determining specific behavioral phenotypes. Novel techniques of genome editing, such as CRISPR-Cas9 (Cong et al. 2013), can also provide an unprecedented view of the functional role of some of the candidate genes we identified.

As we anticipated, only a small portion of the differential expressed genes detected between Iberian and non-Iberian wolves carried polymorphisms in the protein-coding sequence. Similar to what was previously proposed for wild and domestic animals (Li et al. 2014; Sato et al. 2020), it is likely that the behavioral differentiation between Iberian and non-Iberian wolves is largely explained by variation in gene regulation. Because our analysis was based on RNA-seq data, we were limited to variation in the exome. Future

studies should aim to uncover signatures of selection across the entire genome of Iberian wolves, to account for polymorphisms and structural variation in promoter regions that could affect gene expression. Although we have generated a sufficient sample size of Iberian wolf genomes (Chapter III) to perform genome-wide scans of selection, we lack an equally well sampled European wolf population for comparison. An ideal scenario would be to have WGS data of some of the individuals or populations we analyzed in Chapter III using the SNP chip technology. With the increasing availability of wolf genomes, we believe we will be able to fully address this question in the near future.

7.3. Genomic signatures following a population bottleneck in Iberian wolves

Iberian wolves were heavily persecuted in the last two centuries, where a severe reduction of the population size and distribution range in the 1970s left the Iberian wolf population reduced to its minimum (Clavero et al. 2022; Nores and López-Bao 2022). Wolves in the South of Iberia were almost completely extirpated, except for a small population that has remained isolated in Sierra Morena at least since the 1980s (Sastre et al. 2011; López-Bao et al. 2015). However, the situation in the North of Iberia was very different, where wolves managed to persist and due to the implementation of strong conservation measures in recent decades, the population has been expanding (Blanco and Cortés 2009; Chapron et al. 2014; López-Bao et al. 2018). However, little is known about the signatures left in the Iberian wolf genome by such a massive bottleneck, and if they differ under such different population range scenarios of contraction and expansion. Multiple and repeated evidence of genomic erosion following a population bottleneck has been described in several species, for example, the Iberian lynx (Abascal et al. 2016) and the mountain gorilla (Xue et al. 2015), including some that were not able to recover and end up extinct, like many rhinoceros species (Liu et al. 2021) and lions in North Africa (de Manuel et al. 2020). Increased rates of hybridization between different species have also been reported following habitat homogenization (Grabenstein and Taylor 2018). In wolves, there is evidence that hybridization with domestic dogs is associated with the disruption of their social structure and that is more frequent in peripheral areas of the wolf range (Godinho et al. 2011; Galaverni et al. 2017). So, how have genetic diversity and dynamics of interspecific gene flow in Iberian wolves been affected by the population decline?

In Chapter V we addressed this question by analyzing historical and contemporary samples of Iberian wolves. Once again, the integration of historical and contemporary samples, which were widely distributed across time and geographical space, was revealed to be fundamental to understanding how anthropogenic disturbance has affected the evolutionary trajectory of Iberian wolves. Our findings confirmed a rapid and sudden loss of genetic diversity following the decade of 1970, and revealed that the level of genetic variability in the historical population was similar to that found today in large and outbred Eastern European wolf populations (Pilot et al. 2014). This result is particularly relevant given that Iberian wolves are thought to have diverged from the other European populations ca. 10.3 ka, with negligible levels of gene flow ever since (Silva et al. 2020). Therefore, contrasting with the inferred absence of gene flow with other populations, the wolf was likely highly connected in Iberia over the last thousand years to ensure the maintenance of high levels of variability, at least until the mid-20th century. As expected, we showed that the loss of genetic diversity was more severe under a scenario of population contraction. However, the considerable difference in sample size between the North and South regions hampered further testing for significant diversity patterns. Similar to what was previously reported for other large carnivore populations facing extreme declines, like the Apennine bear (Benazzo et al. 2017) and the Iberian lynx (Kleinman-Ruiz et al. 2022), we detected a process of active maintenance of genetic diversity in Iberian wolves, specifically in genomic regions enriched with genes associated with immune response.

An opposite dynamic was observed for the rate of hybridization between wolves and dogs, as the highest number of hybrids was detected in the contemporary Northwestern population. In line with previous evidence, our work shows a higher prevalence of hybrids in geographical areas of current expansion. The area of expansion covers the Southeastwards region of Castilla y Leon, which is characterized by intensively modified agricultural plains, largely comprised of cereal and maize fields and highly human-populated (Blanco and Cortés 2009). This area was previously thought to be an unsuitable area for wolves, but it turned out to be one of the areas that wolves recolonized more successfully (Blanco and Cortés 2009). Given these landscape features, it is plausible to assume that the number of free-ranging dogs will be also higher, which can promote ample opportunities for hybridization. Our results also showed low levels of admixture in the genome of wolves from Sierra Morena in the past, revealing that dog introgression did not increase in this region as an effect of the southern wolf population decline. This finding is particularly relevant to uncovering the evolutionary trajectory of this isolated population in the last decades, given that the only genomic

assessment previously performed was based on a single genome and reported massive levels of dog introgression (Gómez-Sánchez et al. 2018). However, our sample size in this region is low and sparse in space and time, and it would be relevant to address this issue with a larger number of samples from all decades since 1970, mainly after 1990, which corresponds to our last sampling point and when we still did not observe signs of a population threatened by dog introgression.

To date, the past genomic composition of the Iberian wolf remained largely understudied. Our work provided the first population genomic assessment combining historical and contemporary samples to characterize the levels of genetic diversity and admixture in wolves under distinct population range dynamics, following a severe bottleneck. We believe our findings will contribute to understanding aspects of the evolutionary trajectory of wolves in the past decades in Iberia and might also provide important insights for other large carnivore species that have, or still are, facing similar demographic changes. Our findings will also contribute to advancing our knowledge of the ecological settings that can contribute to, or promote, hybridization between wolves and dogs. Characterizing the levels of admixture in the genome of wolves from the historical and contemporary populations (Chapters III and V, respectively) is also of high relevance for conservation-wise practices, as it showed that the Iberian wolf is not threatened by hybridization with domestic dogs.

7.4. Workflow to detect hybridization and introgression

Over the last few decades, major technical and theoretical advances (i.e., sequence technologies, availability of whole-genome data, and development of new analytical tools) have allowed an unprecedented understanding of the frequency and extent of events of gene flow between species (Aguillon et al. 2022). Several methods and tools have been made available to detect gene flow and characterize the ancestry within species' genome. In general, all these methods were designed with the premise that gene flow can be detected by analyzing patterns of increased shared genetic variation between donor and recipient species (Yuan et al. 2017). Detection of gene flow is usually performed at the global (whole-genome) and local (specific genomic region) level, and both approaches (usually) require the availability of genomic information from both donor and recipient species (Aguillon et al. 2022). However, the immense variety of methods available can be overwhelming for researchers that are planning to perform such

analysis. In fact, we understood the need for a clear and systematic workflow when developing the work presented in Chapters III and V, where we aimed to detect and characterize hybridization and introgression patterns in Iberian wolves. As such, we designed a workflow (**Figure 7.1**) with clear steps to identify and quantify levels of admixture between two populations, or species (hereinafter the term population will be used to refer to populations or species), based on the work we have performed. We have divided it into **population-based** or **individual-based** methods.

7.4.1. Population-based methods

In this context, population-based analyses were defined as those using genome-wide SNP data of several individuals from two or more populations and putative admixed individuals, either one way or both ways admixture. A population-based analysis may be performed using global and/or local ancestry methods.

Global ancestry methods (**Figure 7.1,** Step 1) allow to identify individuals with hybrid origin when analyzed together with individuals from parental populations, and different parental populations can be evaluated at the same time. Two different types of global ancestry methods are generally implemented, Clustering and Dimensionality-reduction methods. **Clustering methods** are those methods that estimate the genome admixture proportion of single individuals to each parental population applying maximum-likelihood or Bayesian algorithms. Individuals are classified as belonging to one of the parental populations (two parental populations in the example provided in Step 1 of **Figure 7.1**) or being admixed, with different ancestry proportions. One of the most widely used software for this analysis is ADMIXTURE (Alexander et al. 2009), but FASTSTRUCTURE (Raj et al. 2014) and INSTRUCT (Gao et al. 2007) are among other examples; **Dimensionality-reduction methods** are methods that reduce the dimensionality of large datasets by creating new uncorrelated variables that successively maximize the variance of the dataset. Among these methods, Principal Component Analysis (PCA) is the one widely implemented. Individual values for the new variables are projected onto the axes to identify patterns and differentiate groups of parental individuals from different populations, where admixed individuals generally appear at intermediary positions. However, caution is needed when inferring admixture solely based on PCA results, as this is not a formal test (Patterson et al. 2012). Global ancestry methods do not require prior information about parental populations, which makes them a useful tool for a first and blind assessment of the admixture levels. After this step, reference individuals (non-

admixed) and individuals carrying some levels of admixture are identified and can be used in local ancestry methods.

Local ancestry methods (**Figure 7.1,** Step 2) screen the genome of individuals and estimate the most likely ancestry of small genomic regions along the chromosome, considering two or more parental populations. These methods are usually categorized based on the implemented algorithm, whether it models linkage disequilibrium (LD) or not, and if it requires phased haplotype data (Geza et al. 2019; Mazandu et al. 2019). **Non-LD-based methods** usually use the allele frequencies of ancestry informative markers (AIMs) in admixed and parental populations to estimate the ancestry of each locus. Phased haplotype information is not required, and because these methods do not model LD, linked SNPs are automatically excluded. LAMP (Sankararaman et al. 2008) and WINPOP (Pasaniuc et al. 2009) are examples of software that implement such methods, among many others. The major difference between these two methods relies on the ability to self-defining the window size to conduct the analysis. LAMP implements windows of fixed size (N SNPs) along the chromosome, assuming no recombination; in contrast, WINPOP adjusts the window size accordingly to the local genetic structure of the parental populations, allowing for recombination events within windows. Both methods can be used in the absence of known parental populations (or when parental populations were not genotyped) if individuals have resulted from recent admixture events. Non-LD-based methods are less accurate than LD-based methods but can be particularly useful when the SNP density is not too high (e.g., for historical samples). **LD-based methods** model LD using Hidden Markov Models (HMMs), offering highly accurate estimates of local ancestry. ELAI (Guan 2014) and LAMP-LD (Baran et al. 2012) are examples of tools that apply such methods. LAMP-LD requires phased haplotype data for the parental populations and assumes a fixed window size along the chromosome. Given the lack of high-quality haplotype information for most non-model organisms, haplotypes must be reconstructed using phasing computational methods, such as BEAGLE (Browning and Browning 2007) or SHAPEIT (Delaneau et al. 2012). The phasing process can result in ambiguous reconstructions of the haplotypes, which may introduce bias during ancestry estimation. To account for the unavailability of phased reference panels, ELAI can alternately be used. This software does not require phased data (but can also use it) and self-defines the window size accordingly with the local recombination rate. It offers additional features, for example, models genetic structure within the admixed population, estimates local recombination rates, and infers ancestry from missing parental populations if the admixed population carries a high ancestry proportion

from them. After performing local ancestry analysis, it is possible to plot the ancestry along the genome to easily localize and characterize introgressed genomic regions. In the example provided in **Figure 7.1**, Step 2, the karyotype of a single individual (N chromosomes) reveals most of the ancestry with origin in population A and small introgressed genomic regions from population B. Usually, the size of the introgressed regions is associated with time since introgression, the smaller the block, the older the introgression event. To detect regions with a significantly higher proportion of introgression across individuals, delta (Δ) statistics (Tang et al. 2007) can be calculated. Δ statistics for parental population B at marker *m* is defined as $q_B^m - q_B$, where q_B^m is the mean ancestry at the marker *m* and q_B represents the mean genome-wide ancestry over all individuals. Positive values represent core hotspot regions for introgression from population B in the genome. After implementing local ancestry analysis, ancestry fractions at the chromosome level are estimated per individual. The size, position, and abundance of the introgressed genomic regions within the genome can be reported and visualized.

7.4.2. Individual-based methods

Individual-based analyses compare sequence alignments at introgressed genomic regions or chromosome/genome levels of individuals from different populations. This analysis can be implemented using traditional phylogenetic methods, network methods, and/or methods searching for genomic regions of high sequence similarity.

Phylogenetic methods (**Figure 7.1,** Step 3) can be used to detect inconsistencies in phylogenetic relationships of individuals at genome-wide or local genomic regions. The example in **Figure 7.1**, Step 3, shows a species tree using whole-genome data, where an admixed individual clusters together with an individual from population A, and a tree using data from an introgressed genomic region (identified by local ancestry methods) showing that the admixed individual clusters instead with an individual from population B. This type of inconsistencies may be indicative of introgression and can be searched using a wide range of software, for example, DENSITREE (Bouckaert 2010), TASSEL (Bradbury et al. 2007), RAXML (Stamatakis 2014), and TREEMIX (Pickrell and Pritchard 2012). Additionally, to visualize reticulate relationships between individuals, SPLITSTREE (Huson 1998) may be implemented. SPLITSTREE uses haplotype sequence alignments to build split networks between the main branches of a tree. The presence of loops in the network is compatible with the patterns of multiple topologies obtained using the methods mentioned above (**Figure 7.1**, Step 3). Although we included this approach in

the individual-based analyses section, such methods can also be performed at a population level (using pairwise genetic distances between individuals or population allele counts). In addition to those, there are also methods specially designed to detect archaic introgression using genome-wide data, such as *D* statistics and its derivates (Green et al. 2010; Durand et al. 2011; Martin et al. 2015). This method determines whether there is an excess of shared variation in the genome between two populations, P2-P3, that is not shared between P1-P3, considering the phylogenetic conformation (((P1, P2), P3), O), where O stands for an outgroup. Positive values are indicative of introgression between P2 and P3 post-divergence (Green et al. 2010; Durand et al. 2011; Martin et al. 2015). This method was initially designed to compare individual genomes but can also be applied to populations.

Sequence similarity methods (**Figure 7.1**, Step 4) are used to identify genomic regions of high genetic similarity between haplotypes of a group of individuals. The similarity is searched in windows along the chromosome (or entire genome) and the position, size, and percentage of sequence similarity are reported. HYBRIDCHECK (Ward and van Oosterhout 2016) is an example of such a method, which calculates pairwise sequence similarity in a set of triplets using sliding windows of N base pairs to determine the number and length of introgressed genomic regions. HYBRIDCHECK also reports the number of substitutions between haplotypes of individuals within each genomic region and attributes a statistical value to support introgression. The example in **Figure 7.1**, Step 4, shows a chromosome graph produced by HYBRIDCHECK, where it is possible to observe a clear signature of high sequence similarity between the admixed individual and an individual from population B in a region of the chromosome. This software also estimates divergence times between haplotypes of individuals within genomic regions of high sequence similarity, which can be useful to distinguish between signatures of incomplete lineage sorting and introgression.

With this workflow, we hope to help researchers delineate a clear stepwise approach to identify the number of admixed individuals in a population, estimate the ancestry proportion attributed to each parental population, and determine the position and length of introgressed genomic regions. The development of new methods in this area continues apace, where recent advances in machine learning promise to revolutionize the way we analyze genomic introgression (Aguillon et al. 2022).

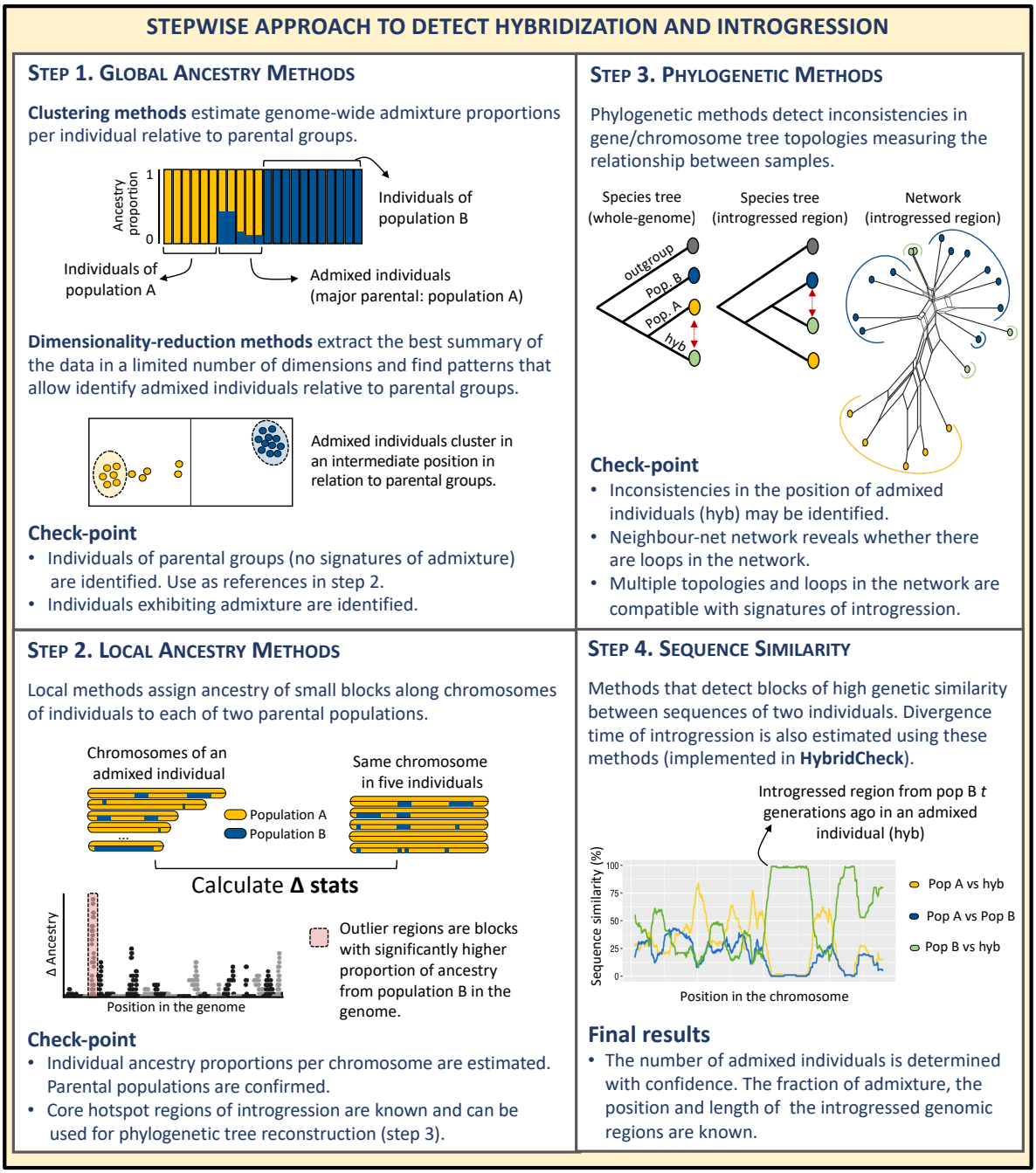


Figure 7.1. Stepwise approach with methods to detect hybridization and genetic introgression between two populations, or species. This workflow was designed for genomic datasets containing a high-density panel of genome-wide SNPs. Having a well-annotated reference genome and chromosome-level information can help in some of this analysis (e.g., Step 2, local ancestry methods).

7.5. Concluding remarks and perspectives

In the barely 70 years since the discovery of the DNA molecule, the field of evolutionary biology has been revolutionized by spectacular improvements in high-throughput genomic technologies. The current increased accessibility to sequence the entire genome of almost every species at a relatively low cost made it possible to design and develop high-resolution genomic studies with hundreds of individuals, to capture population signatures of genetic diversity and evolutionary processes across time and space. Such historical progress led us to develop the work we present in this dissertation. Although several evolutionary mechanisms underlying adaptation to extreme life settings have been uncovered in the last few years, the adaptation to extreme human-modified landscapes was much less understood.

We hope the findings of this dissertation will help to fill the knowledge gap on this issue through the integrative genomic and transcriptomic approach that we have applied to uncover the evolutionary trajectory of wolves living in highly human-dominated landscapes. Distinct evolutionary mechanisms involving adaptive introgression and self-domestication might have been key drivers of adaptation of Iberian wolves to humans and their activities in the Iberian Peninsula. These mechanisms seem to be intrinsically associated with the metabolism of glutamate, which has been proposed to be implicated in the domestication of several animal species. Our findings provide the first glimpse into the functional impact of these evolutionary mechanisms, but future studies taking advantage of emerging techniques of genome editing, such as CRISPR-Cas9, will be necessary to validate and fully characterize the genotype-phenotype associations that our results seem to suggest. Sequencing Iberian wolf genomes with long-read HiFi sequencing could additionally allow us to find and characterize structural variation, which can provide insightful information about genomic signatures of evolution in Iberian wolves.

Our unprecedented characterization of the genetic composition of the historical Iberian wolf population also validated previous demographic trends of significant declines, revealing a clear decrease in genetic diversity associated with the likely increase in human persecution. Hence, the results of this dissertation should be also framed for conservation-wise purposes, especially in the encouragement of the maintenance of conservation measures applied to the wolves in the Iberian Peninsula.

Within this dissertation, we generated and will make available the first brain transcriptomes of wild wolves, which can enable innovative research in the fields of molecular and phenotypic evolution and be of extreme relevance to solving long-standing questions about the main drivers of differentiation between wolves and their domestic counterparts. We also provide the most comprehensive genomic dataset with whole-genomes and SNP data for the Iberian wolf population, which includes historical and contemporary data. In fact, this is likely the most complete dataset for a gray wolf population to date. Given the most recent hypothesis about a dual ancestry of dogs in East and Western Eurasia (Bergström et al. 2022), this new set of 12 Iberian wolf genomes may be used to add a small piece to the puzzle of dog domestication.

All things considered, I hope that the genomic resources we have generated, the software tool we have provided, and the extensive empirical work we have developed on Iberian wolves, which together make up this dissertation, will provide an insightful contribution to researchers in the field of evolutionary biology, particular to those dedicated to understanding the wonders of wolves. These last few years were a truly fantastic journey, and as future research progresses, I can only expect outstanding developments in our understanding of the evolutionary processes that shape life on Earth.

7.6. References

Abascal F, Corvelo A, Cruz F, Villanueva-Cañas JL, Vlasova A, Marcet-Houben M, Martínez-Cruz B, Cheng JY, Prieto P, Quesada V, et al. 2016. Extreme genomic erosion after recurrent demographic bottlenecks in the highly endangered Iberian lynx. Genome Biol. 17:251.

Aguillon SM, Dodge TO, Preising GA, Schumer M. 2022. Introgression. Curr. Biol. 32:R865–R868.

Albert FW, Somel M, Carneiro M, Aximu-Petri A, Halbwax M, Thalmann O, Blanco-Aguiar JA, Plyusnina IZ, Trut L, Villafuerte R, et al. 2012. A Comparison of Brain Gene Expression Levels in Domesticated and Wild Animals. PLoS Genet. 8:e1002962.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res.19:1655–1664.

Anderson TM, VonHoldt BM, Candille SI, Musiani M, Greco C, Stahler DR, Smith DW, Padhukasahasram B, Randi E, Leonard JA, et al. 2009. Molecular and Evolutionary History of Melanism in North American Gray Wolves. Science 323:1339–1343.

Baran Y, Pasaniuc B, Sankararaman S, Torgerson DG, Gignoux C, Eng C, Rodriguez-Cintron W, Chapela R, Ford JG, Avila PC, et al. 2012. Fast and accurate inference of local ancestry in Latino populations. Bioinformatics 28.

Benazzo A, Trucchi E, Cahill JA, Maisano Delser P, Mona S, Fumagalli M, Bunnefeld L, Cornetti L, Ghirotto S, Girardi M, et al. 2017. Survival and divergence in a small group: The extraordinary genomic history of the endangered Apennine brown bear stragglers. Proc. Natl. Acad. Sci. 114:E9589–E9597.

Bergström A, Stanton DWG, Taron UH, Frantz L, Sinding M-HS, Ersmark E, Pfrengle S, Cassatt-Johnstone M, Lebrasseur O, Girdland-Flink L, et al. 2022. Grey wolf genomic history reveals a dual ancestry of dogs. Nature 607:313-320.

Bicknell AA, Cenik C, Chua HN, Roth FP, Moore MJ. 2012. Introns in UTRs: Why we should stop ignoring them. BioEssays 34:1025–1034.

Blanco JC, Cortés Y. 2009. Ecological and social constraints of wolf recovery in Spain. In: Musiani M, Boitani L, Paquet P, editors. A New Era for Wolves and People: Wolf Recovery, Human Attitudes, and Policy. Calgary: University of Calgary Press. p. 41–66.

Blanco JC, Cortés Y. 2012. Surveying wolves without snow: a critical review of the methods used in Spain. Hystrix, Ital. J. Mammal. 23:35–48.

Blanco JC, de la Cuesta L, Reig S. 1990. El lobo en España: una visión global. In: Blanco JC, de la Cuesta L, Reig S, editors. El lobo (Canis lupus) en España. Situación, problemática y apuntes sobre su ecología. Madrid: ICONA. p. 69–94.

Bouckaert RR. 2010. DensiTree: Making sense of sets of phylogenetic trees. Bioinformatics 26:1372–1373.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. 2007. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635.

Browning SR, Browning BL. 2007. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. Am. J. Hum. Genet. 81:1084–1097.

Card DC, Shapiro B, Giribet G, Moritz C, Edwards SV. 2021. Museum Genomics. Annu. Rev. Genet. 55:633–659.

Carneiro M, Rubin CJ, Palma F Di, Albert FW, Alföldi J, Barrio AM, Pielberg G, Rafati N, Sayyab S, Turner-Maier J, et al. 2014. Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. Science 345:1074–1079.

Chapron G, Kaczensky P, Linnell JDC, von Arx M, Huber D, Andrén H, López-Bao JV, Adamec M, Álvares F, Anders O, et al. 2014. Recovery of large carnivores in Europe's modern human-dominated landscapes. Science 346:1517–1519.

Clavero M, García-Reyes A, Fernández-Gil A, Revilla E, Fernández N. 2022. Where wolves were: setting historical baselines for wolf recovery in Spain. Anim. Conserv.

Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, et al. 2016. A survey of best practices for RNA-seq data analysis. Genome Biol.13

Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, et al. 2013. Multiplex Genome Engineering Using CRISPR/Cas Systems. Science 339:819–823.

Delaneau O, Marchini J, Zagury JF. 2012. A linear complexity phasing method for thousands of genomes. Nat. Methods 9:179–181.

Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for Ancient Admixture between Closely Related Populations. Mol. Biol. Evol. 28:2239–2252.

Fan Z, Silva P, Gronau I, Wang S, Armero AS, Schweizer RM, Ramirez O, Pollinger J, Galaverni M, Ortega Del-Vecchyo D, et al. 2016. Worldwide patterns of genomic variation and admixture in gray wolves. Genome Res. 26:163–173.

Freedman AH, Gronau I, Schweizer RM, Ortega-Del Vecchyo D, Han E, Silva PM, Galaverni M, Fan Z, Marx P, Lorente-Galdos B, et al. 2014. Genome Sequencing Highlights the Dynamic Early History of Dogs. PLoS Genet. 10:e1004016.

Fushan AA, Turanov AA, Lee SG, Kim EB, Lobanov A V., Yim SH, Buffenstein R, Lee SR, Chang KT, Rhee H, et al. 2015. Gene expression defines natural changes in mammalian lifespan. Aging Cell 14:352–365.

Galaverni M, Caniglia R, Pagani L, Fabbri E, Boattini A, Randi E. 2017. Disentangling timing of admixture, patterns of introgression, and phenotypic indicators in a hybridizing Wolf population. Mol. Biol. Evol. 34:2324–2339.

Gao H, Williamson S, Bustamante CD. 2007. A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. Genetics 176:1635–1651.

Garzón J. 1979. La apasionada geografia del lobo. Trofeo 104:26–28.

Geza E, Mugo J, Mulder NJ, Wonkam A, Chimusa ER, Mazandu GK. 2019. A comprehensive survey of models for dissecting local ancestry deconvolution in human genome. Brief. Bioinform. 20:1709–1724.

Godinho R, Llaneza L, Blanco JC, Lopes S, Álvares F, García EJ, Palacios V, Cortés Y, Javier

T, Ferrand N. 2011. Genetic evidence for multiple events of hybridization between wolves and domestic dogs in the Iberian Peninsula. Mol. Ecol. 20:5154–5166.

Gómez-Sánchez D, Olalde I, Sastre N, Enseñat C, Carrasco R, Marques-Bonet T, Lalueza-Fox C, Leonard JA, Vilà C, Ramírez O. 2018. On the path to extinction: Inbreeding and admixture in a declining grey wolf population. Mol. Ecol. 27:3599–3612.

Gopalakrishnan S, Sinding MHS, Ramos-Madrigal J, Niemann J, Samaniego Castruita JA, Vieira FG, Carøe C, Montero M de M, Kuderna L, Serres A, et al. 2018. Interspecific Gene Flow Shaped the Evolution of the Genus Canis. Curr. Biol. 28:3441-3449.e5

Grabenstein KC, Taylor SA. 2018. Breaking Barriers: Causes, Consequences, and Experimental Utility of Human-Mediated Hybridization. Trends Ecol. Evol. 33:198–212.

Grant PR, Grant BR. 2019. Hybridization increases population variation during adaptive radiation. Proc. Natl. Acad. Sci. 116:23216–23224.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, et al. 2010. A Draft Sequence of the Neandertal Genome. Science 328:710–722.

Guan Y. 2014. Detecting Structure of Haplotypes and Local Ancestry. Genetics 196:625–642.

Hansen KD, Wu Z, Irizarry RA, Leek JT. 2011. Sequencing technology does not eliminate biological variability. Nat. Biotechnol. 29:572–573.

Hare B, Wobber V, Wrangham R. 2012. The self-domestication hypothesis: evolution of bonobo psychology is due to selection against aggression. Anim. Behav. 83:573–585.

Hibar DP, Adams HHH, Jahanshad N, Chauhan G, Stein JL, Hofer E, Renteria ME, Bis JC, Arias-Vasquez A, Ikram MK, et al. 2017. Novel genetic loci associated with hippocampal volume. Nat. Commun. 8:13624.

Hoeppner MP, Lundquist A, Pirun M, Meadows JRS, Zamani N, Johnson J, Sundström G, Cook A, FitzGerald MG, Swofford R, et al. 2014. An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. PLoS One 9:91172.

Hou Y, Qi F, Bai X, Ren T, Shen X, Chu Q, Zhang X, Lu X. 2020. Genome-wide analysis reveals molecular convergence underlying domestication in 7 bird and mammals. BMC Genomics 21:1–20.

Huson DH. 1998. SplitsTree: Analyzing and visualizing evolutionary data. Bioinformatics 14:68–73.

Kleinman-Ruiz D, Lucena-Perez M, Villanueva B, Fernández J, Saveljev AP, Ratkiewicz M, Schmidt K, Galtier N, García-Dorado A, Godoy JA. 2022. Purging of deleterious burden in the endangered Iberian lynx. Proc. Natl. Acad. Sci. 119:e2110614119.

Lequarré A-S, Andersson L, André C, Fredholm M, Hitte C, Leeb T, Lohi H, Lindblad-Toh K, Georges M. 2011. LUPA: A European initiative taking advantage of the canine genome architecture for unravelling complex disorders in both human and dogs. Vet. J. 189:155–159.

Li Y, Wang GD, Wang MS, Irwin DM, Wu DD, Zhang YP. 2014. Domestication of the dog from the Wolf was promoted by enhanced excitatory synaptic plasticity: A hypothesis. Genome Biol. Evol. 6:3115–3121.

Liu S, Westbury MV, Dussex N, Mitchell KJ, Sinding M-HS, Heintzman PD, Duchêne DA, Kapp JD, von Seth J, Heiniger H, et al. 2021. Ancient and modern genomes unravel the evolutionary history of the rhinoceros family. Cell 184:4874-4885.e16.

López-Bao JV, Blanco JC, Rodríguez A, Godinho R, Sazatornil V, Alvares F, García EJ, Llaneza L, Rico M, Cortés Y, et al. 2015. Toothless wildlife protection laws. Biodivers. Conserv. 24:2105–2108.

López-Bao JV, Fleurke F, Chapron G, Trouwborst A. 2018. Legal obligations regarding populations on the verge of extinction in Europe: Conservation, Restoration, Recolonization, Reintroduction. Biol. Conserv. 227:319–325.

de Manuel M, Barnett R, Sandoval-Velasco M, Yamaguchi N, Garrett Vieira F, Zepeda Mendoza ML, Liu Shiping, Martin MD, Sinding M-HS, Mak SST, et al. 2020. The evolutionary history of extinct and living lions. Proc. Natl. Acad. Sci. 117:10927–10934.

Martin SH, Davey JW, Jiggins CD. 2015. Evaluating the Use of ABBA–BABA Statistics to Locate Introgressed Loci. Mol. Biol. Evol. 32:244–257.

Martins-de-Souza D, Guest PC, Mann DM, Roeber S, Rahmoune H, Bauder C, Kretzschmar H, Volk B, Baborie A, Bahn S. 2012. Proteomic Analysis Identifies Dysfunction in Cellular Transport, Energy, and Protein Metabolism in Different Brain Regions of Atypical Frontotemporal Lobar Degeneration. J. Proteome Res. 11:2533–2543.

Mazandu G, Geza E, Seuneu M, Chimusa E. 2019. Orienting Future Trends in Local Ancestry Deconvolution Models to Optimally Decipher Admixed Individual Genome Variations. In: Bioinformatics Tools for Detection and Clinical Interpretation of Genomic Variations.

Miao B, Wang Z, Li Y. 2017. Genomic analysis reveals hypoxia adaptation in the tibetan mastiff by introgression of the gray Wolf from the tibetan plateau. Mol. Biol. Evol. 34:734–743.

Nores C, López-Bao JV. 2022. Historical data to inform the legal status of species in Europe: An example with wolves. Biol. Conserv. 272:109639.

O'Rourke T, Boeckx C. 2020. Glutamate receptors in domestication and modern human evolution. Neurosci. Biobehav. Rev. 108:341–357.

Pasaniuc B, Sankararaman S, Kimmel G, Halperin E. 2009. Inference of locus-specific ancestry in closely related populations. Bioinformatics 25:i213–i221.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient admixture in human history. Genetics 192:1065–1093.

Pickrell JK, Pritchard JK. 2012. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. PLoS Genet. 8:e1002967.

Pilot M, Greco C, VonHoldt BM, Jędrzejewska B, Randi E, Jędrzejewski W, Sidorovich VE, Ostrander EA, Wayne RK. 2014. Genome-wide signatures of population bottlenecks and diversifying selection in European wolves. Heredity 112:428–442.

Raj A, Stephens M, Pritchard JK. 2014. FastSTRUCTURE: Variational inference of population structure in large SNP data sets. Genetics 197.

Raxworthy CJ, Smith BT. 2021. Mining museums for historical DNA: advances and challenges in museomics. Trends Ecol. Evol. 36:1049–1060.

Rico M, Torrente J. 2000. Caza y rarificación del lobo en España: investigación histórica y conclusiones biológicas. Galemys 12:163–179.

Roy M, Kim N, Kim K, Chung WH, Achawanantakun R, Sun Y, Wayne R. 2013. Analysis of the canine brain transcriptome with an emphasis on the hypothalamus and cerebral cortex. Mamm. Genome 24:484–499.

Sankararaman S, Sridhar S, Kimmel G, Halperin E. 2008. Estimating Local Ancestry in Admixed Populations. Am. J. Hum. Genet. 82:290–303.

Der Sarkissian C, Allentoft ME, Ávila-Arcos MC, Barnett R, Campos PF, Cappellini E, Ermini L, Fernández R, da Fonseca R, Ginolhac A, et al. 2015. Ancient genomics. Philos. Trans. R. Soc. B Biol. Sci. 370:20130387.

Sastre N, Vilà C, Salinas M, Bologov V V., Urios V, Sánchez A, Francino O, Ramírez O. 2011. Signatures of demographic bottlenecks in European wolf populations. Conserv. Genet. 12:701–712.

Satizabal CL, Adams HHH, Hibar DP, White CC, Knol MJ, Stein JL, Scholz M, Sargurupremraj M, Jahanshad N, Roshchupkin G V., et al. 2019. Genetic architecture of subcortical brain structures in 38,851 individuals. Nat. Genet. 51:1624–1636.

Sato DX, Rafati N, Ring H, Younis S, Feng C, Blanco-Aguiar JA, Rubin C-J, Villafuerte R, Hallböök F, Carneiro M, et al. 2020. Brain Transcriptomics of Wild and Domestic Rabbits Suggests That Changes in Dopamine Signaling and Ciliary Function Contributed to Evolution of Tameness. Genome Biol. Evol.12:1918–1928.

Schweizer RM, Durvasula A, Smith J, Vohr SH, Stahler DR, Galaverni M, Thalmann O, Smith DW, Randi E, Ostrander EA, et al. 2018. Natural Selection and Origin of a Melanistic Allele in North American Gray Wolves. Mol. Biol. Evol. 35:1190–1209.

Shulman JM, Imboywa S, Giagtzoglou N, Powers MP, Hu Y, Devenport D, Chipendo P, Chibnik LB, Diamond A, Perrimon N, et al. 2014. Functional screening in Drosophila identifies Alzheimer's disease susceptibility genes and implicates Tau-mediated mechanisms. Hum. Mol. Genet. 23:870–877.

Silva P, Galaverni M, Ortega-Del Vecchyo D, Fan Z, Caniglia R, Fabbri E, Randi E, Wayne R, Godinho R. 2020. Genomic evidence for the Old divergence of Southern European wolf populations. Proc. R. Soc. B Biol. Sci. 287:20201206.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Steffens M, Leu C, Ruppert A-K, Zara F, Striano P, Robbiano A, Capovilla G, Tinuper P, Gambardella A, Bianchi A, et al. 2012. Genome-wide association analysis of genetic generalized epilepsies implicates susceptibility loci at 1q43, 2p16.1, 2q22.3 and 17q21.32. Hum. Mol. Genet. 21:5359–5372.

Stronen AV, Jędrzejewska B, Pertoldi C, Demontis D, Randi E, Niedziałkowska M, Borowik T, Sidorovich VE, Kusak J, Kojola I, et al. 2015. Genome-wide analyses suggest parallel selection for universal traits may eclipse local environmental selection in a highly mobile carnivore. Ecol. Evol. 5:4410–4425.

Tang H, Choudhry S, Mei R, Morgan M, Rodriguez-Cintron W, Burchard EG, Risch NJ. 2007. Recent Genetic Selection in the Ancestral Admixture of Puerto Ricans. Am. J. Hum. Genet. 81:626–633.

Valverde J. 1971. El lobo español. Montes 159:229–241.

Vaysse A, Ratnakumar A, Derrien T, Axelsson E, Rosengren Pielberg G, Sigurdsson S, Fall T, Seppälä EH, Hansen MST, Lawley CT, et al. 2011. Identification of Genomic Regions Associated with Phenotypic Variation between Dog Breeds using Selection Mapping.

PLoS Genet. 7:e1002316.

Wang MS, Wang S, Li Y, Jhala Y, Thakur M, Otecko NO, Si JF, Chen HM, Shapiro B, Nielsen R, et al. 2020. Ancient Hybridization with an Unknown Population Facilitated High-Altitude Adaptation of Canids. Mol. Biol. Evol. 37:2616–2629.

Wang X, Pipes L, Trut L, Herbeck Y, Vladimirova A, Gulevich R, Kharlamova A, Johnson J, Acland G, Kukekova A, et al. 2018. Genomic responses to selection for tame/aggressive behaviors in the silver fox (Vulpes vulpes). Proc. Natl. Acad. Sci. 115:10398–10403.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: A revolutionary tool for transcriptomics. Nat. Rev. Genet. 10:57–63.

Ward BJ, van Oosterhout C. 2016. hybridcheck: software for the rapid detection, visualization and dating of recombinant regions in genome sequence data. Mol. Ecol. Resour. 16:534–539.

Xue Y, Prado-Martinez J, Sudmant PH, Narasimhan V, Ayub Q, Szpak M, Frandsen P, Chen Y, Yngvadottir B, Cooper DN, et al. 2015. Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. Science 348:242–245.

Yuan K, Zhou Y, Ni X, Wang Y, Liu C, Xu S. 2017. Models, methods and tools for ancestry inference and admixture analysis. Quant. Biol. 5:236–250.

Appendix

Appendix I.

frontiers | Frontiers in Ecology and Evolution

TYPE Original Research
PUBLISHED 08 September 2022
DOI 10.3389/fevo.2022.970249

Check for updates

OPEN ACCESS

EDITED BY
Jimmy McGuire,
University of California, Berkeley,
United States

REVIEWED BY
Shaoqing Wen,
Fudan University, China
Nicolas Dussex,
Stockholm University, Sweden

*CORRESPONDENCE
Raquel Godinho
rgodinho@cibio.up.pt

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Ecology and Evolution

RECEIVED 15 June 2022
ACCEPTED 11 August 2022
PUBLISHED 08 September 2022

CITATION
Pacheco C, Lobo D, Silva P, Álvares F, García EJ, Castro D, Layna JF, López-Bao JV and Godinho R (2022) Assessing the performance of historical skins and bones for museomics using wolf specimens as a case study. *Front. Ecol. Evol.* 10:970249. doi: 10.3389/fevo.2022.970249

COPYRIGHT
© 2022 Pacheco, Lobo, Silva, Álvares, García, Castro, Layna, López-Bao and Godinho. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Assessing the performance of historical skins and bones for museomics using wolf specimens as a case study

Carolina Pacheco[1,2,3†], Diana Lobo[1,2,3†], Pedro Silva[1,3], Francisco Álvares[1,3], Emilio J. García[4], Diana Castro[1,3], Jorge F. Layna[5], José Vicente López-Bao[4] and Raquel Godinho[1,2,3,6*]

[1]CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, *InBIO* Laboratório Associado, Universidade do Porto, Vairão, Portugal, [2]Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Porto, Portugal, [3]BIOPOLIS, Program in Genomics, Biodiversity and Land Planning, CIBIO, Vairão, Portugal, [4]Biodiversity Research Institute (CSIC), Oviedo University, Mieres, Spain, [5]Consultores en Iniciativas Ambientales, S.L., Madrid, Spain, [6]Department of Zoology, Centre for Ecological Genomics and Wildlife Conservation, University of Johannesburg, Johannesburg, South Africa

Advances in the field of museomics have promoted a high sampling demand for natural history collections (NHCs), eventually resulting in damage to invaluable resources to understand historical biodiversity. It is thus essential to achieve a consensus about which historical tissues present the best sources of DNA. In this study, we evaluated the performance of different historical tissues from Iberian wolf NHCs in genome-wide assessments. We targeted three tissues—bone (jaw and femur), maxilloturbinal bone, and skin—that have been favored by traditional taxidermy practices for mammalian carnivores. Specifically, we performed shotgun sequencing and target capture enrichment for 100,000 single nucleotide polymorphisms (SNPs) selected from the commercial Canine HD BeadChip across 103 specimens from 1912 to 2005. The performance of the different tissues was assessed using metrics based on endogenous DNA content, uniquely high-quality mapped reads after capture, and enrichment proportions. All samples succeeded as DNA sources, regardless of their collection year or sample type. Skin samples yielded significantly higher amounts of endogenous DNA compared to both bone types, which yielded equivalent amounts. There was no evidence for a direct effect of tissue type on capture efficiency; however, the number of genotyped SNPs was strictly associated with the starting amount of endogenous DNA. Evaluation of genotyping accuracy for distinct minimum read depths across tissue types showed a consistent overall low genotyping error rate (<7%), even at low (3x) coverage. We recommend the use of skins as reliable and minimally destructive sources of endogenous DNA for whole-genome and target enrichment approaches in mammalian carnivores. In addition, we

provide a new 100,000 SNP capture array validated for historical DNA (hDNA) compatible to the Canine HD BeadChip for high-quality DNA. The increasing demand for NHCs as DNA sources should encourage the generation of genomic datasets comparable among studies.

KEYWORDS

endogenous DNA, historical DNA, museomics, natural history collections, SNP genotyping errors, target enrichment

Introduction

Natural history collections (NHCs) have been gathered since the seventeenth century, motivated by human curiosity about our planet's biodiversity and the breakthrough in preserving perishable material (Farrington, 1915). Currently, national, regional, and private collections worldwide own irreplaceable natural resources, offering wide perspectives across distinct temporal and spatial scales that inspire research in many scientific areas (Casas-Marce et al., 2012; Tsangaras and Greenwood, 2012; Lopez et al., 2020; Pearson et al., 2020). Such collections provide unique overviews of historical biodiversity, from both extinct and extant species, and are also essential resources for addressing questions about species that have sampling limitations due to financial, bureaucratic, or conservation constraints (Burrell et al., 2015). Advances in molecular genetics and sequencing technology have promoted the use of the naturally fragmented DNA of historical specimens, transforming NHCs into invaluable sources of material for investigating genetics-related questions among different fields, including phylogenetics, biogeography, and conservation (Suarez and Tsutsui, 2004; Holmes et al., 2016; Bi et al., 2019).

Whereas a growing body of studies using historical DNA (hDNA) illustrate the potential of NHCs in genetic research, their regular use still poses different challenges. Sampling of genetic material is often destructive, eventually compromising the integrity of specimens and jeopardizing their future use. Therefore, because sampling techniques that minimize damages are prioritized (Pálsdóttir et al., 2019), the amount of genetic material collected is often limited (Horváth et al., 2005). Additionally, traditional taxidermy practices, which commonly use hazardous chemicals, and general carelessness in protecting specimens from environmental damage, do not favor DNA preservation. Thus, historical samples often yield limited and highly degraded DNA (Raxworthy and Smith, 2021) that presents major challenges during laboratory and analytical procedures (Allentoft et al., 2012; Dabney et al., 2013). Moreover, hDNA extracts can contain a non-negligible proportion of exogenous DNA from pre- or post-mortem sources, frequently in overwhelming ratios (Weiß et al., 2016; McDonough et al., 2018; Eisenhofer et al., 2019). These factors may explain why genetic studies using hDNA have often relied on the amplification of short nuclear or mitochondrial fragments (e.g., Schwartz et al., 2007; Maebe et al., 2016; Lonsinger et al., 2019). However, the use of hDNA is nowadays facilitated by high-throughput sequencing and by recent developments in molecular methods (e.g., Rowe et al., 2011; Staats et al., 2013; Hung et al., 2014). Methods like sequence capture of target loci, which limits the representation of the genome to specific loci, are among the most used in genome-wide studies of low-quality DNA to achieve large and cost-effective datasets (Jones and Good, 2016; McCormack et al., 2016; Derkarabetian et al., 2019). Furthermore, despite requiring *a priori* availability of the target genome to design specific baits, target enrichment has been shown to be successful using bait designs based on closely related species (Vallender, 2011).

Mammals have been traditionally preserved in NHCs by archiving skins, bones, teeth, or mounted specimens (Rowe et al., 2011). It is thus not surprising that most of the available genetic studies using hDNA rely on these tissues (Raxworthy and Smith, 2021). However, DNA yields may vary greatly among different tissues and be dependent on curation history (Burrell et al., 2015). Hard tissues, such as teeth and bones, were at first thought to provide higher-quality DNA (Wandeler et al., 2007; Casas-Marce et al., 2010), encouraging proposals to use hard tissues assumed to minimize sampling damage, such as maxilloturbinal bone (Wisely et al., 2004). Yet, recent studies have shown conflicting results (Rowe et al., 2011; Lonsinger et al., 2019; Tsai et al., 2020), revealing soft tissues to be good sources of hDNA when preserved appropriately (Burrell et al., 2015). To date, few studies have implemented genomic resources to assess differences in the performance of distinct tissues, and those tackling this question are often based on very low sample sizes that hamper reliable statistical comparisons (e.g., Rowe et al., 2011; McDonough et al., 2018). Thus, the tissue of choice for increasing the quality of genomic data, while sampling mammal NHCs with minimal damaging, remains an open question (Raxworthy and Smith, 2021).

In this work, we sought to evaluate the performance of different mammalian carnivore tissues generally available at NHCs in genome-wide assessments. Using the Iberian wolf (*Canis lupus signatus*) as a case study, we first collected bones (jaw and femur), maxilloturbinal bones, and skins from

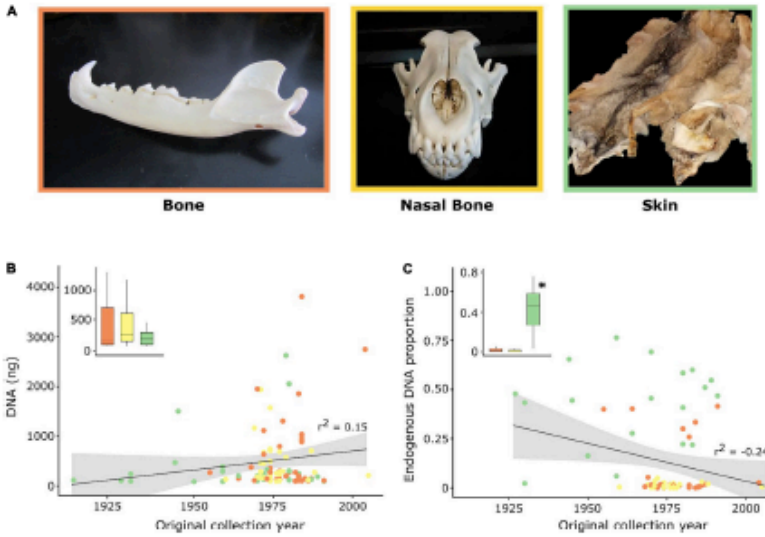


FIGURE 1
Characterization of samples and hDNA used in this work. (A) Photographs illustrating the specimens used for the collection of each tissue type (Image credits Raquel Godinho). (B,C) Correlation (r^2) between DNA yield (ng) following extraction and the proportion of endogenous DNA content with the original collection date of the specimens. Each dot represents a historical sample and inset boxplots display the median (central line) and distribution per tissue type. Significant differences ($p < 0.05$) between tissues are identified with an asterisk. Colors depict the three different tissue types used in this study: orange—bones (jaw and femur); yellow—maxilloturbinal (nasal) bones; and, green—skins.

103 historical specimens (Figure 1A). Then, we performed a shotgun sequencing of these samples to characterize endogenous DNA content. Third, we developed and tested a capture array of 100,000 regions overlapping single nucleotide polymorphisms (SNPs) contained in the commercial Illumina Canine HD BeadChip, ensuring compatibility between datasets generated with both low- and high-quality DNA. We intended to answer three main questions: Does endogenous DNA content differ across historical tissues? How is capture efficiency affected by historical tissue type? What is the effect of read depth on SNP genotyping error rates from hDNA?

Materials and methods

Sampling, DNA extraction, and shotgun sequencing

We collected 103 samples from Iberian wolf specimens housed at the three largest museum collections in the Iberian Peninsula and at 15 private NHCs, consisting of 43 bones (jaw and femur), 31 maxilloturbinal bones (hereafter nasal bones), and 29 skins. The original collection years for these specimens ranged from 1912 to 1990, except for two samples from 2004 and 2005 (Figure 1, Supplementary Table 1). All our samples conform to the definition of historical samples by Raxworthy and Smith (2021), which restricts hDNA to that fortuitously obtained from traditional museum specimens not intended to serve as sources of DNA. Jaw and femur bones were sampled by drilling ca. 1 g of bone powder with a Dremel tool (Dremel, WI, United States; Supplementary Figure 1). Femurs were sampled in the patellar surface region, whereas jaws were sampled in the posterior lower region of the mandibula (see Supplementary Figure 1 for examples of the drilling location in each bone type). Nasal bones were collected following Wisely et al. (2004), by inserting sterilized forceps into the nasal cavity of the skull to extract the bones. Nasal bone material was posteriorly crushed into small fragments. Bones were not bleached prior to DNA extraction. Skin samples were collected from pelts or mounted specimens by extracting a patch of approximately 2 cm^2. Collecting tools, including drill bits, were cleaned with bleach, and flamed with 96% ethanol between samples to minimize cross contamination.

We prepared DNA extracts using 50 mg of sample following Dabney and Meyer (2019). For skins, we favored the inner

layer for DNA extraction and discarded hairs to minimize contamination with external DNA sources. DNA concentration was measured using the Qubit fluorometer dsDNA HS Assay Kit (Thermo Fisher Scientific, MA, United States). We used 100–300 ng of DNA to prepare blunt-end dual-indexed DNA sequencing libraries using a full-uracil-DNA-glycosylase treatment following Meyer and Kircher's (2010) protocol with the modifications described in Kircher et al. (2012). To limit DNA contamination, DNA extractions and library preparations were conducted in dedicated rooms under sterile conditions and positive air pressure, and negative controls were used alongside the procedures. DNA libraries were diluted based on concentration measurements obtained with the Qubit fluorometer dsDNA HS Assay Kit, and library size ranges were characterized using a Bioanalyzer 2100 with High Sensitivity DNA kits (Agilent Technologies, CA, United States). To characterize the endogenous DNA content of each sample, we performed a shotgun sequencing run using one lane of an Illumina HiSeq X instrument (Illumina, CA, United States) in PE 150 bp mode. For this, we selected libraries with concentrations > 15 ng/μl ($N = 79$; $N_{bone} = 29$; $N_{nasal\ bone} = 30$; $N_{skin} = 20$) to ensure that enough library material remained for the following steps (based on the required starting amount for target capture enrichment of 14–72 ng/μl).

We also generated genomic information from two contemporary wolf muscle samples to be used as positive controls to validate the implemented approach. No animals were killed or injured for this study. We isolated DNA from these samples using the DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany). DNA quantification and library preparation were performed following the same procedures as described above for historical samples but with two modifications during library preparation: (i) a shearing step using a Bioruptor Pico sonication device (Diagenode, NJ, United States) was performed to obtain fragments of ~250 bp, and (ii) the USER enzyme treatment was not performed. These two samples were genotyped using two different approaches: the Illumina CanineHD BeadChip (Illumina, CA, United States; ~170,000 SNPs) and the capture array of 100,000 regions developed in this work (see next sections).

Bait design

We targeted a set of autosomal genome-wide SNPs whose positions were defined based on the coordinates available on the CanineHD BeadChip. With this experimental design, we ensured the compatibility of historical datasets with other datasets generated using the same SNP chip. From the ~170,000 SNPs in the CanineHD BeadChip, only uniquely mapped autosomal SNPs were considered for probe design. Based on this list of putative SNPs, we custom designed a sequence capture panel containing a final set of 100,000 SNPs distributed across the wolf genome using MYbaits Target Capture technology (Arbor Biosciences, Ann Arbor, MI, United States). The implemented methods for probe design followed the strategy described by Haak et al. (2015) and Cruz-Dávalos et al. (2017), in which four probes of 60 bp each were designed to target fragments of 120 bp per SNP: one upstream of the SNP, one downstream, and one for each possible allele, with the allele positioned at the center of the probe. RNA probes were designed and synthesized by Arbor Biosciences (product code #302016). The selection of the final set of MYbaits RNA probes was based on their capture efficiency (stringent criteria $\leq$ 25% repeat masked). A FASTA file providing the sequences of all 400,000 synthesized probes is available online.

Capture enrichment

We performed target enrichment using the MYbaits Custom Target Capture Kit following the manufacturer protocol (v.3.02). DNA libraries with similar endogenous DNA content were pooled in equimolar sets of 8–10 samples per capture reaction (600 ng in the final pool). Hybridization between RNA probes and the DNA library occurred at 65°C for 40 h. Real-time PCRs were performed to determine the number of amplification cycles required to obtain sufficient molarity for sequencing. Post-enrichment libraries were amplified using KAPA HiFi Hotstart Ready Mix (KAPA Biosystems, MA, United States), following the manufacturer recommendations, with an annealing temperature of 60°C. The amplified enriched DNA libraries were purified in 20 μl of EB buffer using the MinElute PCR Purification kit (Qiagen, Hilden, Germany) and quantified by the Qubit dsDNA HS Assay Kit. Library size ranges were characterized using a Bioanalyzer 2100 with High Sensitivity DNA kits and pooled in equimolar ratios. Pooled libraries were then sequenced using 2 lanes of an Illumina HiSeq X instrument in PE 150 bp mode.

Processing of sequencing reads and single nucleotide polymorphism calling

Raw sequences from shotgun sequencing and target enrichment were processed similarly in different time periods. First, sequence reads were demultiplexed and quality assessments were done using FastQC (Andrews, 2018) and MultiQC (Ewels et al., 2016). Sequence reads were then processed and aligned using the PALEOMIX v.1.2.13.2 BAM pipeline (Schubert et al., 2014). Briefly, the pipeline consisted of the following steps: (i) adapter sequences were removed, low-quality and N bases were trimmed, and overlapping read pairs were collapsed using AdapterRemoval v.2.2.2 (Schubert et al., 2016) with the default parameters (on average, 96 and 95%

of reads were collapsed for shotgun and capture sequencing, respectively); (ii) all sequence reads were then aligned against the CanFam3.1 dog reference genome (Lindblad-Toh et al., 2005) using BWA-MEM v.0.7.17 (Heng, 2013); and (iii) PCR duplicates for each library were marked by the "paleomix rmdup_collapsed" tool but were not used for the following steps. The final alignment file was subjected to local realignment around indels using the GATK v.3.8 IndelRealigner tool (DePristo et al., 2011). The endogenous content was determined by the ratio of unique reads (no duplicates) that mapped to the dog reference genome with mapping quality above 20 (MQ > 20) to the total number of available reads (also referred as library complexity by some authors; Dehasque et al., 2022). For the target enrichment experiment, the number of reads mapping on target was defined as the total number of unique high-quality (MQ > 20) reads overlapping at least 1 base of the 120 bp target region. Enrichment success was determined by the ratio of reads on target in relation to (i) the total number of available reads and (ii) the total number of uniquely and high-quality mapped reads. These metrics were retrieved from the PALEOMIX summary report and, additionally, with the help of SAMtools v.1.9 (Li et al., 2009). Fold enrichment was determined by the ratio of the number of on-target reads to the total number of uniquely and high-quality mapped reads, divided by the expected representation of the target regions without enrichment (i.e., the ratio of genome length, 2.4 Gb, to target length, 100,000 × 120 bp, corresponding to 0.5%).

Following target enrichment, genotypes were called using BCFtools v.1.10.2 (Li, 2011) mpileup/call -m tools, with minimum Phred-scaled thresholds of 20 for base quality and read mapping quality. At the end, to evaluate the effect of sequencing depth on genotyping quality rates, we considered genotypes supported by at least three (DP ≥ 3), four (DP ≥ 4), or more reads (DP ≥ 5) using the custom python script gtvalues2plink.py, which was also used to convert the final VCF file to plink format. The distribution and density of SNPs in our dataset (using DP ≥ 4) was visualized using the R/Bioconductor package karyoploteR (Gel and Serra, 2017) in R (R Development Core Team, 2017).

Genotypes were ultimately validated by estimating their concordance rates with the genotypes obtained from the control samples using the Canine HD BeadChip. In this last approach, genotype calling was performed using GenomeStudio software (Illumina), following Illumina's recommendations. Sex-chromosome-related SNPs and non-uniquely mapped SNPs (SNPs with multiple positions attributed) were removed from the dataset using PLINK v.1.9 (Purcell et al., 2007), resulting in a final dataset of ~ 121,000 SNPs. Concordance rates were calculated for the entire set of genome-wide SNPs obtained across the three depth thresholds for each control sample, using a second custom script, SNP_concordance.py. We also calculated concordance rates for SNPs called exclusively with 3x and 4x coverage in each control sample and estimated the associated error rates. Additionally, to assess the genotype quality across all historical samples, we estimated the potential genotyping error rate associated with low read depth using the ErrorCount.sh script from the dDocent pipeline (Puritz et al., 2014). Briefly, this script reports a low range based on a 50% binomial probability of observing the second allele in a heterozygote and a high range based on a 25% probability. All genotyped SNPs were considered for this analysis without any filter for missing data.

Statistical analysis

To determine whether the amount of endogenous DNA and capture efficiency were influenced by sample type (bone, nasal bone, and skin), we implemented a set of generalized linear models (GLMs). We ran four different GLMs with the following dependent variables: (i) proportion of endogenous DNA, i.e., number of reads that mapped uniquely with MQ > 20 in relation to the total number of available reads, following shotgun sequencing; (ii) mapped reads after capture, i.e., proportion of the number of reads that mapped uniquely and with MQ > 20 to the total number of reads available after target enrichment; (iii) reads on target (all), i.e., number of unique and high-quality (MQ > 20) reads that mapped on target regions in relation to the total number of reads available; and (iv) reads on target (mapped), i.e., the same as above, but in relation to the total number of mapped reads. All GLMs were fitted with a binomial error distribution and a logit link. The fit of each model was further assessed using the Pearson's χ^2 residuals, which test whether any significant patterns remain in the residuals. Given the unavailability of shotgun sequencing data for all the samples, we tested levels of correlation between the proportion of endogenous DNA and the proportion of reads mapping after capture to understand if the latter could be interpreted as a proxy for endogenous content. We also tested correlation coefficients using the following variables: original collection year, DNA yield (ng) following extraction, endogenous DNA proportion, fragment length (average length of filtered reads from shotgun sequencing), and mapping length (average length of mapped and unique reads, with MQ > 20). The effect of sample type on DNA concentration was also evaluated using a Kruskal-Wallis test. All the previously mentioned tests were performed in R, and all the plots were constructed using the R package ggpubr (Kassambara, 2020).

Results

Endogenous DNA content

We successfully obtained DNA extracts for all 103 historical samples, with an average DNA yield of 469.91 ± 63.18 (s.e.)

ng; (range 84–3,812 ng). DNA yield was not correlated with the original specimen collection year (r^2 = 0.15) nor with the sample type (Kruskal-Wallis test, p = 0.256; Figure 1B). Nevertheless, even using the same quantity of starting material for DNA extraction, we cannot rule out greater effects due to histological differences between hard and soft tissues. The initial sample characterization by shotgun sequencing resulted in an average fragment length of 97 bp (range 42–145 bp). Endogenous DNA content across all samples varied from 0.05 to 76.35% (mean: 14.48 ± 2.47%; Figure 1C) and was not correlated with the fragment length (r^2 = −0.42; Supplementary Figure 2A) nor with the original collection year (r^2 = −0.24; Figure 1C). The average mapping length across all samples was 75 bp (range 40–116 bp) and was not correlated with specimen original collection year (r^2 = 0.15; Supplementary Figure 2B). Among sample types, skin samples retrieved the highest proportion of endogenous DNA (43.49 ± 4.76%, p = 0.008; Figure 1C and Supplementary Tables 2, 3) in relation to bones (8.14 ± 2.66%) and nasal bones (1.28 ± 0.22%).

After capture enrichment, the proportion of reads mapping to the reference genome ranged from 0.13 to 78.26% (mean: 19.89 ± 2.59%; Supplementary Table 2). Consistently, skin samples presented a significantly higher proportion of mapped reads (49.01 ± 4.69%, p = 0.001; Figure 2 and Supplementary Tables 2, 4) relative to bones (12.57 ± 2.99%) and nasal bones (2.79 ± 0.64%). Across contemporary control samples, the average proportion of reads mapping after capture was 32.85 ± 0.56% (Supplementary Table 5), with 65.4% of filtered reads being duplicates. Endogenous DNA content and the proportion of reads mapping to the reference genome following capture were highly correlated (r^2 = 0.93; Supplementary Figure 3A).

Capture efficiency across sample types

Following capture enrichment, the proportion of reads mapping on target regions ranged from 0.01 to 27.79% (mean: 4.75 ± 0.73%; Supplementary Table 2) in relation to all available reads and from 1 to 37% (mean: 19.82 ± 0.84%; Supplementary Table 2) in relation to all mapped reads. The average fold enrichment was 39.64x, with 91% of all samples presenting an enrichment > 10x. Among sample types, skin samples showed the highest proportion of reads mapping on target regardless of the metric used (12.35 ± 1.58% of all reads; 22.49 ± 1.65% of mapped reads; Figure 2). Bones (2.68 ± 0.78% of all reads; 18.13 ± 1.41% of mapped reads) and nasal bones (0.50 ± 0.11% of all reads; 19.68 ± 1.18% of mapped reads) worked less successfully; however, differences were not significant among the three sample types (Figure 2 and Supplementary Tables 6, 7). For the two contemporary control samples, the proportion of reads mapping on target regions was 17.54 ± 0.01% of all available reads and 53.42 ± 0.69% of mapped reads (Supplementary Table 5).

Genotyping errors and coverage effect

We were able to generate genotypes for a panel of 99,982 genome-wide SNPs (Supplementary Figure 4) using the target enrichment approach, with distinct levels of missing data across samples. For contemporary control samples, concordance rates between genotypes obtained from the capture array developed in this study and those from the Canine HD BeadChip were above 99%. Error rates for the two control samples (i.e., rates of genotype discordance) were almost negligible but increased with decreasing coverage at the target sites (0.64 and 0.91%

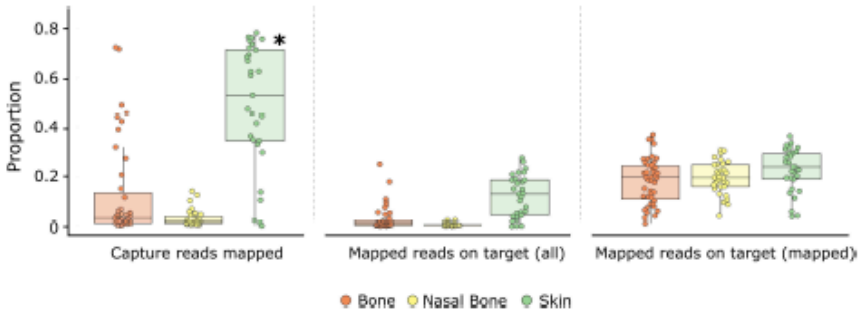


FIGURE 2
Genomic metrics assessed from capture enrichment for each tissue type: bones (orange), maxilloturbinal (nasal) bones (yellow), and skins (green). Boxplots display the proportion of reads mapping after capture enrichment and reads mapping on target region in relation to all or mapped reads for each tissue. Within boxplots, dots represent historical samples, and the central line indicates the median value. Significant differences (p < 0.05) between tissues for each metric are identified with an asterisk.

error rate, at coverage ≥ 5x and ≥ 3x, respectively; Figure 3A and Supplementary Table 8). When calculating the genotype concordance rates using SNPs called exclusively with 3x and 4x, we found a decrease to 94.1 and 96.4%, respectively (Supplementary Table 8). The most common error found in genotypes called with the lowest coverage (3x) was the dropout of a second allele (miscalled homozygous in relation to the SNP chip), corresponding to 79.5% of the observed discordances. The number of loci obtained in contemporary control samples increased as the coverage decreased; for example, the average number of SNPs obtained across both samples declined from 93,948 to 88,173 at coverage ≥ 3x and ≥ 5x, respectively (Figure 3B and Supplementary Table 9).

We found the same pattern across historical samples, with increased estimates of potential genotyping error rates associated with SNPs called with decreasing coverage (Figure 3A). Nevertheless, the highest error rate estimate for coverage ≥ 3x did not reach 7%. Concordantly, we also found an increase in the average number of SNPs across all samples for lower read depths (24,663 and 18,523 SNPs at coverage ≥ 3x and ≥ 5x; Figure 3B and Supplementary Table 9). Overall, the average number of SNPs obtained across historical samples was substantially lower than in contemporary control samples. Still, the highest genotyping success rates were found among three historical samples, which presented very similar rates (>95%) to those found in the contemporary samples (Figure 3B and Supplementary Table 9). Skin samples presented the highest average numbers of SNPs genotyped (Figure 3B). Genotyping success rate (number of genotyped SNPs) was positively correlated with the proportion of reads mapping to the genome after capture ($r^2 = 0.85$; Supplementary Figure 3B).

Discussion

In this study, we evaluated the performance of different historical tissues—bones, nasal bones, and skins—across

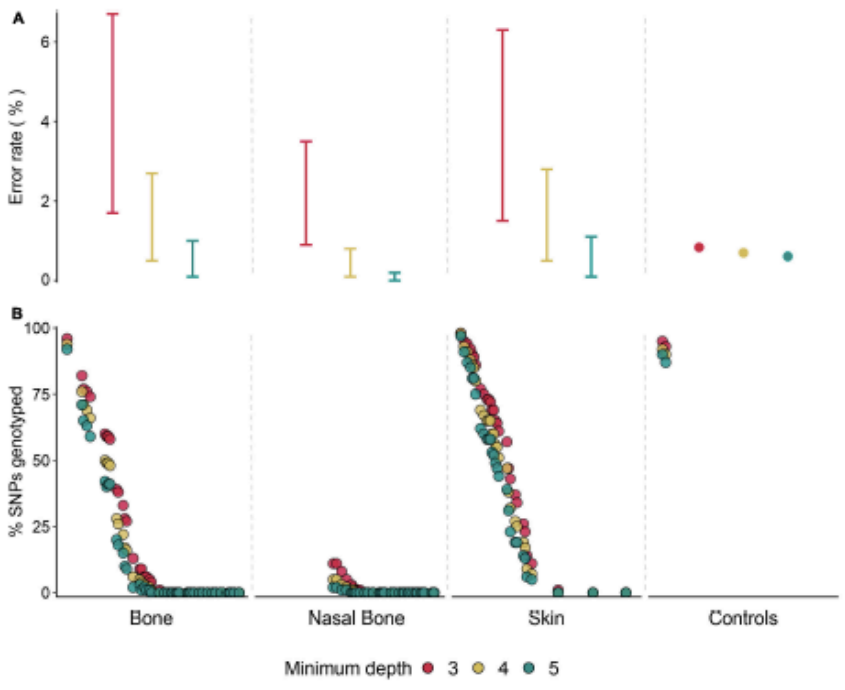


FIGURE 3
Genotyping performance of hDNA from three sample types. (**A**) Interval of potential genotyping error rate estimates (%) for the three hDNA sample types, and average empirical genotyping error rate for control samples, considering the three minimum depth levels. (**B**) Percentage of genotyped SNPs (N = 99,982) considering a minimum depth of 3x, 4x, or 5x per SNP across 103 samples. Each dot represents a sample.

a population sampling of Iberian wolf NHCs, in whole-genome sequencing and target enrichment. Specifically, we analyzed metrics based on endogenous DNA content, uniquely high-quality mapped reads after capture, and enrichment proportions.

Bones, nasal bones, and skins all succeeded as DNA sources but differed in endogenous DNA content. Based on equal starting amounts of DNA for shotgun sequencing, skins yielded significantly higher amounts of endogenous DNA than did the other sample types. This concords with other studies that have previously shown that soft tissues of mammals (van der Valk et al., 2017) and birds (Tsai et al., 2020) provide higher endogenous DNA content than hard tissues, although these are based on less comprehensive sample sizes than the one used here. Although nasal bones were initially recommended for presenting higher genotyping success rates than other bones (Wisely et al., 2004), a recent study showed conflicting results (Lonsinger et al., 2019). Here, we assessed for the first time the proportion of endogenous DNA retrieved from nasal bones, showing that this bone type provides lower content than other commonly used bones. Despite the observed significant effect of tissue type on endogenous DNA, we cannot disregard the impact of distinct collection histories in this result (Raxworthy and Smith, 2021). We did not observe significant associations between endogenous DNA content, or its mapping length, and the original collection year, suggesting no substantial DNA degradation across the period under evaluation.

There was no direct effect of tissue type on capture efficiency, with most samples presenting an enrichment > 10x. This result emphasizes the efficiency of target capture as a powerful method for building genomic datasets from hDNA and is in line with other studies using capture approaches (Carpenter et al., 2013; van der Valk et al., 2017). An *a priori* understanding of the importance of endogenous DNA content in the success of a capture experiment (Hernandez-Rodriguez et al., 2018) has driven our decision to perform shotgun sequencing prior to capture enrichment. Using this approach, we were able to demonstrate a perfect association between the number of resulting SNPs and the initial amount of endogenous DNA, emphasizing that this is a practical and cost-effective way to select samples prior to capture experiments. Furthermore, knowing the endogenous DNA content across samples also allows to minimize variation within each sequencing experiment, ensuring low rates of index hopping and reducing possible bias in downstream analysis (van der Valk et al., 2020). Our SNP genotyping results demonstrate that historical samples with high amounts of endogenous DNA can behave similarly to, or even better than, contemporary samples in capture procedures, further supporting the use of capture enrichment to genotype thousands of genome-wide SNPs in hDNA samples (Bi et al., 2013; Smith et al., 2014; Harvey et al., 2016; Lim and Braun, 2016).

The final number of SNPs in a dataset can be a trade-off between coverage and genotyping error rates: relaxing the minimum read depth to increase the number of SNPs is accompanied by higher uncertainty in genotype calling. Still, we were able to generate an average of ca. 25,000 genome-wide SNPs per sample, called with $\geq$ 3x coverage and with low potential error rates (<7%). Reducing coverage led to an increase in the number of SNPs, mostly across samples with intermediate genotyping success, i.e., those where SNPs were captured but read depth was generally low. In such cases, decreasing the coverage from $\geq$ 4x to $\geq$ 3x represented an increase of ~11,000 SNPs. Regardless of the threshold used for minimum coverage, genotypes obtained for the contemporary control samples were confirmed and validated against the SNP chip genotypes. The average error rates found in the controls were within, or very close to, the potential error rate intervals estimated for the historical dataset. This overlap suggests that the use of such an analytical approach (Puritz et al., 2014) can be a reliable alternative to estimate potential error rates when no control samples are available. Observed error rates in our work for $\geq$ 5x coverage, which is a widely accepted threshold for SNP calling (Yi et al., 2010), were much lower than those reported by Fountain et al. (2016) at the same coverage for fresh tissue samples.

The most common error found among control samples was the dropout of a second allele after target enrichment. Given the overall low error rates, we predict minimal impact associated with allelic dropouts in downstream analysis. However, even low rates of genotyping error tend to overestimate genetic variation and can affect population genetic studies in different ways, or, to a greater extent linkage and association studies (Pompanon et al., 2005; Gautier et al., 2013). Since most studies do not have control samples available to calculate empirical concordance rates and validate genotypes (Fountain et al., 2016), predicting the effects of errors might be difficult, and in this case higher minimal depth thresholds (>5x) may be considered. Choosing between increasing the number of loci or having high reliability in the genotypes should be considered on a case-by-case basis to ensure compatibility with downstream applications. Alternatively, historical datasets can be analyzed based on genotype likelihoods instead of genotypes, in order to take the inherent uncertainty of the genotypes into account (Nielsen et al., 2011).

Reducing sample damage without compromising data recovery

Skin samples were revealed to be a good source of endogenous DNA while still being minimally destructive to specimens, as small patches can be easily sampled from non-unique morphological features of hides or skin mounts. Bone sampling is generally more destructive and does not

necessarily translate into higher endogenous DNA content. This applies in particular to nasal bones, which yield the lowest endogenous DNA content while relying on a quite destructive sampling process in which a large part of the structure is removed. Although we acknowledge that the consumptive sampling of nasal bones does not compromise the utility of a specimen for morphometric or character studies (Wisely et al., 2004), our results discourage their sampling for genomic studies of mammalian carnivores.

As the demand for NHCs in molecular studies increases, conscious sampling of specimens should now become routine to not compromise their future use. The drilling procedure we used for bone sampling left a small hole in the bone tissue (Figure 1A and Supplementary Figure 1) but did not hamper future morphometric studies. This sampling approach is less damaging than cutting bone fragments (Supplementary Figure 1) and still ensures that enough material is collected for genomic analysis without further manipulation. Sampling the petrous bone, a recognized excellent source of endogenous DNA (Pinhasi et al., 2015), would only be recommendable for NHC specimens if entailing the use of damaged skulls, as its sampling is highly destructive to the skull (Charlton et al., 2019). The starting amount of endogenous DNA can also be maximized through wet lab procedures by performing several rounds of DNA extraction, creating multiple and differentially indexed DNA libraries to increase complexity levels, and/or captures per sample (Hernandez-Rodriguez et al., 2018; White et al., 2019; Fontsere et al., 2021; von Seth et al., 2021). Understanding that a reduced fragment of the appropriate tissue is enough to successfully recover molecular data without compromising the reusability of the specimen is of remarkable importance for managing NHCs.

Limitations and opportunities for further development

In contrast to current practices for preserving biological material in controlled environments and using sophisticated resources that prevent DNA damage, older traditional methods did not prioritize DNA integrity (Hall et al., 1997; Burrell et al., 2015; Card et al., 2021). Thus, these preservation techniques and storage conditions can greatly impact the quality and quantity of DNA. In this work, we used wolf specimens from three museums and 15 private collections distributed across a 90-year period, where most samples (~60%) were collected between 1960 and 1980. These samples likely have different collection histories, particularly those in private collections, where less-standardized preservation methods can be expected. Such heterogeneity may explain the observed levels of variability in endogenous DNA content, although our sample size hampers a statistical evaluation of these metrics across sampling origins. We acknowledge that it would be relevant to assess the effects of different preservation techniques on DNA degradation; unfortunately, information about the method applied to each specimen used for this study was not available, as is often the case in NHCs.

One of the major challenges the museum community currently faces is improving the documentation of collection histories for individual specimens and to make it accessible to the scientific community. Global initiatives have been promoting the digitization of specimens' metadata and digital images to make them available in electronic databases that can be easily accessed (e.g., The Global Information Facility; Robertson et al., 2014). As this process evolves, it is essential that researchers working in the field of museomics synergize genomic data with NHCs metadata to enhance the scientific impact and traceability of their studies by, for example, always providing the catalog numbers of specimens used (Card et al., 2021).

Concluding remarks

To our knowledge, this study uses the most comprehensive dataset to date—in terms of sample size and genome representation—to test the performance of three wolf tissues as sources of hDNA. Based on our findings, we recommend the use of skins for sampling mammalian carnivore specimens, as these are reliable and minimally destructive sources of endogenous DNA suitable for whole-genome and target enrichment approaches. This study should also encourage future research with the same aims but targeting different vertebrate and invertebrate groups. In addition, we provide a validated genome-wide SNP tool (i.e., probe design) that allows for direct comparison between historical and contemporary data. Although the enrichment approach presented here was based on canid genomes, its conceptual design can be implemented in any species for which SNP chips are available. We believe that the increasing demand for NHCs as DNA sources, and the requirements for minimal damage to the specimens, should encourage the generation of genomic datasets comparable among studies.

Data availability statement

Individual SNP genotypes and the sequence of the 400,000 RNA probes are available from the OSF repository: https://osf.io/j3r7x/?view_only=a747acf5297c49309bbb89f2e7414104. Raw sequence reads from whole-genome resequencing are available on NCBI (accession SRA PRJNA860381): https://www.ncbi.nlm.nih.gov/sra/PRJNA860381. Custom python scripts (gtvalues2plink.py and SNP_concordance.py) used in this study are available at https://github.com/pdroslva84/SNPcap.

Author contributions

RG coordinated the project. RG, CP, DL, and PS designed the study. RG, JL-B, and FÁ coordinated historical sample collection efforts. RG, JL-B, EG, FÁ, JL, and DL collected samples from historical specimens. CP, DL, and DC conducted laboratory work. CP and DL performed data analysis under the guidance of RG and PS. PS did the bioinformatic scripting. CP, DL, and RG wrote the manuscript with input from all the other authors. All authors contributed to the article and approved the submitted version.

Funding

This work was co-funded by the Portuguese Foundation for Science and Technology (FCT) under the project PTDC/BIA-EVF/2460/2014 and the project NORTE-01-0246-FEDER-000063, the Norte Portugal Regional Operational Programme (NORTE2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF). DL, CP, RG, and PS were supported by the FCT (PD/BD/132403/2017, PD/BD/135026/2017, 2021/00647/CEECIND, and PTDC/BIA-EVL/31902/2017, respectively). JL-B was supported by the Spanish Ministry of Economy, Industry and Competitiveness (RYC-2015-18932), and the GRUPIN research grant from the Regional Government of Asturias (IDI/2021/000075).

Acknowledgments

We acknowledge the Museu Nacional de História Natural e da Ciência (Lisbon, Portugal), Museo Nacional de Ciencias Naturales (Madrid, Spain), Estación Biológica de Doñana (Seville, Spain) and all owners of private collections who shared historical wolf samples. The two contemporary tissue samples were provided by the Portuguese Institute for Nature Conservation and Forestry (ICNF) and the Consejería de Medio Ambiente de la Junta de Castilla y León. We are grateful to R. Pedra and S. Roque for helping collecting samples, and to S. Mourão and M. Magalhães for lab assistance.

Conflict of interest

JL was employed by Consultores en Iniciativas Ambientales, S.L.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2022.970249/full#supplementary-material

References

Allentoft, M. E., Collins, M., Harker, D., Haile, J., Oskam, C. L., Hale, M. L., et al. (2012). The half-life of DNA in bone: Measuring decay kinetics in 158 dated fossils. *Proc. R. Soc. B Biol. Sci.* 279, 4724–4733. doi: 10.1098/rspb.2012.1745

Andrews, S. (2018). *FastQC: A quality control tool for high throughput sequence data (v. 0.11.7)*. Berlin: ScienceOpen, Inc.

Bi, K., Linderoth, T., Singhal, S., Vanderpool, D., Patton, J. L., Nielsen, R., et al. (2019). Temporal genomic contrasts reveal rapid evolutionary responses in an alpine mammal during recent climate change. *PLoS Genet.* 15:e1008119. doi: 10.1371/journal.pgen.1008119

Bi, K., Linderoth, T., Vanderpool, D., Good, J. M., Nielsen, R., and Moritz, C. (2013). Unlocking the vault: Next-generation museum population genomics. *Mol. Ecol.* 22, 6018–6032. doi: 10.1111/mec.12516

Burrell, A. S., Disotell, T. R., and Bergey, C. M. (2015). The use of museum specimens with high-throughput DNA sequencers. *J. Hum. Evol.* 79, 35–44. doi: 10.1016/j.jhevol.2014.10.015

Card, D. C., Shapiro, B., Giribet, G., Moritz, C., and Edwards, S. V. (2021). Museum Genomics. *Annu. Rev. Genet.* 55, 633–659. doi: 10.1146/annurev-genet-071719-020506

Carpenter, M. L., Buenrostro, J. D., Valdiosera, C., Schroeder, H., Allentoft, M. E., Sikora, M., et al. (2013). Pulling out the 1%: Whole-Genome capture for the targeted enrichment of ancient dna sequencing libraries. *Am. J. Hum. Genet.* 93, 852–864. doi: 10.1016/j.ajhg.2013.10.002

Casas-Marce, M., Revilla, E., and Godoy, J. A. (2010). Searching for DNA in museum specimens: A comparison of sources in a mammal species. *Mol. Ecol. Resour.* 10, 502–507. doi: 10.1111/j.1755-0998.2009.02784.x

Casas-Marce, M., Revilla, E., Fernandes, M., Rodriguez, A., Delibes, M., and Godoy, J. A. (2012). The value of hidden scientific resources: Preserved animal specimens from private collections and small museums. *Bioscience* 62, 1077–1082. doi: 10.1525/bio.2012.62.12.9

Charlton, S., Booth, T., and Barnes, I. (2019). The problem with petrous? A consideration of the potential biases in the utilization of pars petrosa for ancient DNA analysis. *World Archaeol.* 51, 574–585. doi: 10.1080/00438243.2019.1694062

Cruz-Dávalos, D. I., Llamas, B., Gaunitz, C., Fages, A., Gamba, C., Soubrier, J., et al. (2017). Experimental conditions improving in-solution target enrichment for ancient DNA. *Mol. Ecol. Resour.* 17, 508–522. doi: 10.1111/1755-0998.12595

Dabney, J., and Meyer, M. (2019). Extraction of highly degraded DNA from ancient bones and teeth. *Methods Mol. Biol.* 1963, 25–29. doi: 10.1007/978-1-4939-9176-1_4

Dabney, J., Knapp, M., Glocke, I., Gansauge, M. T., Weihmann, A., Nickel, B., et al. (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. U. S. A.* 110, 15758–15763. doi: 10.1073/pnas.1314445110

Dehasque, M., Pečnerová, P., Kempe Lagerholm, V., Ersmark, E., Danilov, G. K., Mortensen, P., et al. (2022). Development and Optimization of a Silica Column-Based Extraction Protocol for Ancient DNA. *Genes* 13:687. doi: 10.3390/genes13040687

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498. doi: 10.1038/ng.806

Derkarabetian, S., Benavides, L. R., and Giribet, G. (2019). Sequence capture phylogenomics of historical ethanol-preserved museum specimens: Unlocking the rest of the vault. *Mol. Ecol. Resour.* 19, 1531–1544. doi: 10.1111/1755-0998.13072

Eisenhofer, R., Minich, J. J., Marotz, C., Cooper, A., Knight, R., and Weyrich, L. S. (2019). Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends Microbiol.* 27, 105–117. doi: 10.1016/j.tim.2018.11.003

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32, 3047–3048. doi: 10.1093/bioinformatics/btw354

Farrington, O. C. (1915). The rise of Natural History Museums. *Science* 42, 197–208. doi: 10.1126/science.42.1076.197

Fontsere, C., Alvarez-Estape, M., Lester, J., Arandjelovic, M., Kuhlwilm, M., Dieguez, P., et al. (2021). Maximizing the acquisition of unique reads in noninvasive capture sequencing experiments. *Mol. Ecol. Resour.* 21, 745–761. doi: 10.1111/1755-0998.13300

Fountain, E. D., Pauli, J. N., Reid, B. N., Palsbøll, P. J., and Peery, M. Z. (2016). Finding the right coverage: The impact of coverage and sequence quality on single nucleotide polymorphism genotyping error rates. *Mol. Ecol. Resour.* 16, 966–978. doi: 10.1111/1755-0998.12519

Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhué, C., Pudlo, P., et al. (2013). The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Mol. Ecol.* 22, 3165–3178. doi: 10.1111/mec.12089

Gel, B., and Serra, E. (2017). KaryoploteR: An R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* 33, 3088–3090. doi: 10.1093/bioinformatics/btx346

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., et al. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522, 207–211. doi: 10.1038/nature14317

Hall, L. M., Willcox, M. S., and Jones, D. S. (1997). Association of enzyme inhibition with methods of museum skin preparation. *Biotechniques* 22, 928–934. doi: 10.2144/97225st07

Harvey, M. G., Smith, B. T., Glenn, T. C., Faircloth, B. C., and Brumfield, R. T. (2016). Sequence Capture versus Restriction Site Associated DNA Sequencing for Shallow Systematics. *Syst. Biol.* 65, 910–924. doi: 10.1093/sysbio/syw036

Heng, L. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. *arXiv* [Preprint]. doi: 10.48550/arXiv.1303.3997

Hernandez-Rodriguez, J., Arandjelovic, M., Lester, J., de Filippo, C., Weihmann, A., Meyer, M., et al. (2018). The impact of endogenous content, replicates and pooling on genome capture from faecal samples. *Mol. Ecol. Resour.* 18, 319–333. doi: 10.1111/1755-0998.12728

Holmes, M. W., Hammond, T. T., Wogan, G. O. U., Walsh, R. E., Labarbera, K., Wommack, E. A., et al. (2016). Natural history collections as windows on evolutionary processes. *Mol. Ecol.* 25, 864–881. doi: 10.1111/mec.13529

Horváth, M. B., Martínez-Cruz, B., Negro, J. J., Kalmár, L., and Godoy, J. A. (2005). An overlooked DNA source for non-invasive genetic analysis in birds. *J. Avian Biol.* 36, 84–88. doi: 10.1111/j.0908-8857.2005.03370.x

Hung, C. M., Shaner, P. J. L., Zink, R. M., Liu, W. C., Chu, T. C., Huang, W. S., et al. (2014). Drastic population fluctuations explain the rapid extinction of the passenger pigeon. *Proc. Natl. Acad. Sci. U. S. A.* 111, 10636–10641. doi: 10.1073/pnas.1401526111

Jones, M. R., and Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. *Mol. Ecol.* 25, 185–202. doi: 10.1111/mec.13304

Kassambara, A. (2020). *ggpubr R Package: Ggplot2-Based Publication Ready Plots*. Available online at: https://rpkgs.datanovia.com/ggpubr/ (accessed July, 2021).

Kircher, M., Sawyer, S., and Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* 40:e3. doi: 10.1093/nar/gkr771

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993. doi: 10.1093/bioinformatics/btr509

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Lim, H. C., and Braun, M. J. (2016). High-throughput SNP genotyping of historical and modern samples of five bird species *via* sequence capture of ultraconserved elements. *Mol. Ecol. Resour.* 16, 1204–1223. doi: 10.1111/1755-0998.12568

Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., Kamal, M., et al. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438, 803–819. doi: 10.1038/nature04338

Lonsinger, R. C., Daniel, D., Adams, J. R., and Waits, L. P. (2019). Consideration of sample source for establishing reliable genetic microsatellite data from mammalian carnivore specimens held in natural history collections. *J. Mammal.* 100, 1678–1689. doi: 10.1093/jmammal/gyz112

Lopez, L., Turner, K. G., Bellis, E. S., and Lasky, J. R. (2020). Genomics of natural history collections for understanding evolution in the wild. *Mol. Ecol. Resour.* 20, 1153–1160. doi: 10.1111/1755-0998.13245

Maebe, K., Meeus, I., Vray, S., Claeys, T., Dekoninck, W., Boevé, J. L., et al. (2016). A century of temporal stability of genetic diversity in wild bumblebees. *Sci. Rep.* 6:38289. doi: 10.1038/srep38289

McCormack, J. E., Tsai, W. L. E., and Faircloth, B. C. (2016). Sequence capture of ultraconserved elements from bird museum specimens. *Mol. Ecol. Resour.* 16, 1189–1203. doi: 10.1111/1755-0998.12466

McDonough, M. M., Parker, L. D., McInerney, N. R., Campana, M. G., and Maldonado, J. E. (2018). Performance of commonly requested destructive museum samples for mammalian genomic studies. *J. Mammal.* 99, 789–802. doi: 10.1093/jmammal/gyy080

Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* 2010:db.prot5448. doi: 10.1101/pdb.prot5448

Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* 12, 443–451. doi: 10.1038/nrg2986

Pálsdóttir, A. H., Bläuer, A., Rannamäe, E., Boessenkool, S., and Hallsson, J. H. (2019). Not a limitless resource: Ethics and guidelines for destructive sampling of archaeofaunal remains. *R. Soc. Open Sci.* 6:191059. doi: 10.1098/rsos.191059

Pearson, K. D., Nelson, G., Aronson, M. F. J., Bonnet, P., Brenskelle, L., Davis, C. C., et al. (2020). Machine learning using digitized herbarium specimens to advance phenological research. *Bioscience* 70, 610–620. doi: 10.1093/biosci/biaa044

Pinhasi, R., Fernandes, D., Sirak, K., Novak, M., Connell, S., Alpaslan-Roodenberg, S., et al. (2015). Optimal ancient DNA yields from the inner ear part of the human petrous bone. *PLoS One* 10:e0129102. doi: 10.1371/journal.pone.0129102

Pompanon, F., Bonin, A., Bellemain, E., and Taberlet, P. (2005). Genotyping errors: Causes, consequences and solutions. *Nat. Rev. Genet.* 6, 847–859. doi: 10.1038/nrg1707

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Puritz, J. B., Hollenbeck, C. M., and Gold, J. R. (2014). dDocent: A RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ* 2014:e431. doi: 10.7717/peerj.431

R Development Core Team (2017). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Raxworthy, C. J., and Smith, B. T. (2021). Mining museums for historical DNA: Advances and challenges in museomics. *Trends Ecol. Evol.* 36, 1049–1060. doi: 10.1016/j.tree.2021.07.009

Robertson, T., Döring, M., Guralnick, R., Bloom, D., Wieczorek, J., Braak, K., et al. (2014). The GBIF Integrated Publishing Toolkit: Facilitating the Efficient

Publishing of Biodiversity Data on the Internet. *PLoS One* 9:e102623. doi: 10.1371/journal.pone.0102623

Rowe, K. C., Singhal, S., Macmanes, M. D., Ayroles, J. F., Morelli, T. L., Rubidge, E. M., et al. (2011). Museum genomics: Low-cost and high-accuracy genetic data from historical specimens. *Mol. Ecol. Resour.* 11, 1082–1092. doi: 10.1111/j.1755-0998.2011.03052.x

Schubert, M., Ermini, L., Der Sarkissian, C., Jónsson, H., Ginolhac, A., Schaefer, R., et al. (2014). Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat. Protoc.* 9, 1056–1082. doi: 10.1038/nprot.2014.063

Schubert, M., Lindgreen, S., and Orlando, L. (2016). AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. *BMC Res. Notes* 9:88. doi: 10.1186/s13104-016-1900-2

Schwartz, M. K., Luikart, G., and Waples, R. S. (2007). Genetic monitoring as a promising tool for conservation and management. *Trends Ecol. Evol.* 22, 25–33. doi: 10.1016/j.tree.2006.08.009

Smith, B. T., Harvey, M. G., Faircloth, B. C., Glenn, T. C., and Brumfield, R. T. (2014). Target Capture and Massively Parallel Sequencing of Ultraconserved Elements for Comparative Studies at Shallow Evolutionary Time Scales. *Syst. Biol.* 63, 83–95. doi: 10.1093/sysbio/syt061

Staats, M., Erkens, R. H. J., Vossenberg, B., van de Wiering, J. J., Kraaijeveld, K., Stielow, B., et al. (2013). Genomic Treasure Troves: Complete Genome Sequencing of Herbarium and Insect Museum Specimens. *PLoS One* 8:e69189. doi: 10.1371/JOURNAL.PONE.0069189

Suarez, A., and Tsutsui, N. (2004). The Value of Museum Collections for Research and Society. *Bioscience* 54, 66–74. doi: 10.1641/0006-35682004054

Tsai, W. L. E., Schedl, M. E., Maley, J. M., and McCormack, J. E. (2020). More than skin and bones: Comparing extraction methods and alternative sources of DNA from avian museum specimens. *Mol. Ecol. Resour.* 20, 1220–1227. doi: 10.1111/1755-0998.13077

Tsangaras, K., and Greenwood, A. D. (2012). Museums and disease: Using tissue archive and museum samples to study pathogens. *Ann. Anat.* 194, 58–73. doi: 10.1016/j.aanat.2011.04.003

Vallender, E. J. (2011). Expanding whole exome resequencing into non-human primates. *Genome Biol.* 12:R87. doi: 10.1186/GB-2011-12-9-R87

van der Valk, T., Lona Durazo, F., Dalén, L., and Guschanski, K. (2017). Whole mitochondrial genome capture from faecal samples and museum-preserved specimens. *Mol. Ecol. Resour.* 17, e111–e121. doi: 10.1111/1755-0998.12699

van der Valk, T., Vezzi, F., Ormestad, M., Dalén, L., and Guschanski, K. (2020). Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. *Mol. Ecol. Resour.* 20, 1171–1181. doi: 10.1111/1755-0998.13009

von Seth, J., Dussex, N., Diez-del-Molino, D., van der Valk, T., Kutschera, V. E., Kierczak, M., et al. (2021). Genomic insights into the conservation status of the world's last remaining Sumatran rhinoceros populations. *Nat. Commun.* 12:2393. doi: 10.1038/s41467-021-22386-8

Wandeler, P., Hoeck, P. E. A., and Keller, L. F. (2007). Back to the future: Museum specimens in population genetics. *Trends Ecol. Evol.* 22, 634–642. doi: 10.1016/j.tree.2007.08.017

Weiß, C. L., Schuenemann, V. J., Devos, J., Shirsekar, G., Reiter, E., Gould, B. A., et al. (2016). Temporal patterns of damage and decay kinetics of dna retrieved from plant herbarium specimens. *R. Soc. Open Sci.* 3:160239. doi: 10.1098/rsos.160239

White, L. C., Fontsere, C., Lizano, E., Hughes, D. A., Angedakin, S., Arandjelovic, M., et al. (2019). A roadmap for high-throughput sequencing studies of wild animal populations using noninvasive samples and hybridization capture. *Mol. Ecol. Resour.* 19, 609–622. doi: 10.1111/1755-0998.12993

Wisely, S. M., Maldonado, J. E., and Fleischer, R. C. (2004). A technique for sampling ancient DNA that minimizes damage to museum specimens. *Conserv. Genet.* 5, 105–107. doi: 10.1023/B:COGE.0000014061.04963.da

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., et al. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 329, 75–78. doi: 10.1126/science.1190371

Appendix II.

PLOS ONE

RESEARCH ARTICLE

On taming the effect of transcript level intra-condition count variation during differential expression analysis: A story of dogs, foxes and wolves

Diana Lobo[1,2,3]*, **Raquel Linheiro**[1], **Raquel Godinho**[1,2,3], **John Patrick Archer**[1,2]*

1 CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, *InBIO* Laboratório Associado, Universidade do Porto, Vairão, Portugal, 2 BIOPOLIS, Programin Genomics, Biodiversity and Land Planning, CIBIO, Vairão, Portugal, 3 Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Porto, Portugal

* diana.lobo@cibio.up.pt (DL); john.archer@cibio.up.pt (JPA)

Check for updates

OPEN ACCESS

Citation: Lobo D, Linheiro R, Godinho R, Archer JP (2022) On taming the effect of transcript level intra-condition count variation during differential expression analysis: A story of dogs, foxes and wolves. PLoS ONE 17(9): e0274591. https://doi.org/10.1371/journal.pone.0274591

Editor: Katherine James, Newcastle University, UNITED KINGDOM

Received: October 14, 2021

Accepted: August 31, 2022

Published: September 22, 2022

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0274591

Copyright: ©2022 Lobo et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All data is publically available on NCBI (https://www.ncbi.nlm.nih.gov) under the project accession numbers: PRJEB3197 (runs: ERR266355, ERR266386, ERR266395,

Abstract

The evolution of RNA-seq technologies has yielded datasets of scientific value that are often generated as condition associated biological replicates within expression studies. With expanding data archives opportunity arises to augment replicate numbers when conditions of interest overlap. Despite correction procedures for estimating transcript abundance, a source of ambiguity is transcript level intra-condition count variation; as indicated by disjointed results between analysis tools. We present TVscript, a tool that removes reference-based transcripts associated with intra-condition count variation above specified thresholds and we explore the effects of such variation on differential expression analysis. Initially iterative differential expression analysis involving simulated counts, where levels of intra-condition variation and sets of over represented transcripts are explicitly specified, was performed. Then counts derived from inter- and intra-study data representing brain samples of dogs, wolves and foxes (wolves *vs.* dogs and aggressive *vs.* tame foxes) were used. For simulations, the sensitivity in detecting differentially expressed transcripts increased after removing hyper-variable transcripts, although at levels of intra-condition variation above 5% detection became unreliable. For real data, prior to applying TVscript, ≈20% of the transcripts identified as being differentially expressed were associated with high levels of intra-condition variation, an over representation relative to the reference set. As transcripts harbouring such variation were removed pre-analysis, a discordance from 26 to 40% in the lists of differentially expressed transcripts is observed when compared to those obtained using the non-filtered reference. The removal of transcripts possessing intra-condition variation values within (and above) the 97th and 95th percentiles, for wolves *vs.* dogs and aggressive *vs.* tame foxes, maximized the sensitivity in detecting differentially expressed transcripts as a result of alterations within gene-wise dispersion estimates. Through analysis of our real data the support for seven genes with potential for being involved with selection for tameness is provided. TVscript is available at: https://sourceforge.net/projects/tvscript/.

ERR266403, ERR266382, ERR266407, ERR266371, ERR266359, ERR266374, ERR266366 and ERR266400), PRJEB4668 (run: ERR351173), PRJNA185055 (runs: SRR636937 and SRR636938), PRJNA78827 (runs: SRR388737, SRR388740, SRR388766, SRR543733, SRR536881, SRR536883) and PRJNA307604 (runs: SRR3084300, SRR3084299, SRR3084298, SRR3084297, SRR3084296, SRR3084295, SRR3084294, SRR3084293, SRR3084292, SRR3084291, SRR3084290, SRR3084289, SRR3084312, SRR3084311, SRR3084310, SRR3084309, SRR3084308, SRR3084307, SRR3084306, SRR3084305, SRR3084304, SRR3084303, SRR3084302 and SRR3084301). Further details are available in S1 Table including: Species, Publication detail (Study), Sample IDs, Project accession and Run accession.

Funding: This work was funded by the project NORTE-01-0246-FEDER-000063, supported by Norte Portugal Regional Operational Programme (NORTE2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF), and by research funding from the projects under the references PTDC/BIA-EVF/29115/2017, PTDC/BIA-EVF/2460/2014 and POCI-01-0145-FEDER-029115 co-funded by Operational Competitiveness and Internationalization Program, Portugal 2020 and the European Union via the European Regional Development Fund (ERDF) and by National Funds through FCT. DL, RG were supported by FCT (PD/BD/132403/2017 to DL, contract under DL57/2016 to RG) and JA was supported by Funds through FCT under the references POCI-01-0145-FEDER-029115 and PTDC/BIA-EVL/29115/2017. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. FCT and NORTH2020 url's: https://www.fct.pt/ and https://www.norte2020.pt.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Developments in RNA-seq technology have revolutionized transcriptomic studies by allowing for a rapid hi-resolution view of transcript expression [1]. In a typical RNA-seq experiment, transcript expression profiles are estimated for each sample using a metric based upon the number of sequenced reads associated with each transcript within a reference set [2–8]. Condition dependent expression profiles can then be used in order to identify which transcripts are differentially expressed [9, 10]. A challenge arises due to sources of variation within expression profiles that are independent of, or partially overlapping with, the condition of interest [2, 11–13]. The inclusion of biological replicates reduces the effect of such noise [14, 15], and it has been demonstrated that sufficient replicate numbers outweigh sequencing depth in terms of increasing the accuracy within differential expression experiments [14, 16]. In studies not involving highly controlled isolated environments, RNA-seq data from the rapidly growing repertoire of published works can be incorporated [17–19], if data from a matching condition to that being studied is available. This effectively increases the number replicates although variability can be amplified [20, 21].

Differential expression tools compute a statistical significance for each transcript, based upon the abundance estimates within a condition, that reflect the possibility of that transcript being differentially expressed [9, 10, 22]. To reduce the effect of intra-condition variation across biological replicates on the estimation of abundance several methods have been proposed including ALDEx2 [23], EDASeq [24] and PEER [25]. In addition to these, and more generally applied, are the abundance estimation techniques implemented within established differential expression tools such as DESeq2 [9] and EdgeR [10]. However, when methods are compared, relative to the final sets of transcripts identified as being differentially expressed, variable results are observed [15, 23, 26–29]. This is an indication that the problem of intra-condition variation relative to the detection of differentially expressed transcripts using RNA-seq data has not been completely resolved. Furthermore, there is no consensus on the best approach to use [30].

Here we explore the effects that individual transcripts associated with high levels of intra-condition count variation have on the end results of differential expression analysis using the tool DESeq2 [9]; a tool that is well established and that has consistently demonstrated reliability in identifying differentially expressed transcripts [27, 30, 31]. Our aim is to investigate the possibility of whether or not the removal of transcripts, harbouring the highest levels of intra-condition variation, from the reference set used during differential expression analysis can produce sets of differentially expressed transcripts that display an increased level of confidence. The latter being achieved through either: (a) the direct removal of transcripts previously identified as being differentially expressed, but whose expression patterns are ambiguous, or (b) the indirect addition, or removal, of transcripts to, or from, those previously identified as being differentially expressed as a consequence of alterations in p-adj values. The latter being associated with shifts in the distribution of intra-condition variation, following the removal of transcripts harbouring the highest levels of such variation. A by-product of this is the explicit quantification of the level of intra-condition abundance variation present within the final lists of differentially expressed transcripts.

To aid this exploration we present TVscript, a tool for the identification of transcripts above user-specified levels of intra-condition normalized count variation, the latter being strongly associated with transcript abundance estimation [4–8]. As input TVscript requires one file per condition-associated replicate that contains the per transcript read counts obtained following the mapping of reads from the replicate to a common reference set. As output TVscript produces a set of corresponding count files that are absent of transcripts harbouring

normalized intra-condition count variation higher than that associated with a user specified percentile. These updated count files can be subsequently used within the differential expression tool of choice, in our case DESeq2 [9]. Through multiple iterations of differential expression analysis following filtering at varying thresholds and comparisons back to differential expression analysis performed on non-filtered inputs, the effects of transcripts associated with high intra-condition variation, in relation to quantity and consistency of differentially expressed transcripts identified, can be explored.

Using TVscript we first explore the effects of intra-condition per-transcript read count variation through iterative differential expression analysis experiments involving highly controlled simulated count datasets derived from the available dog reference transcriptome [32], and where the exact level of background intra-condition count variation could be specified as well as a subset-set of transcripts to be over represented across replicates (of second conditions used within each iteration). Next, we explored the effects of intra-condition per-transcript read count variation within two distinct case-studies, involving count data obtained following the mapping of intra and inter-study RNA-seq datasets. Within these case-studies differential expression patterns arising from data derived from brain samples of dogs and wolves (inter-study scenario involving frontal cortex, cerebral cortex, prefrontal cortex and frontal lobe) [32–35], as well as tame and aggressive foxes (intra-study scenario involving prefrontal cortex) [36], generated in the scope of domestication experiments are compared at varying thresholds of intra-condition normalized read count variation exclusion.

In addition to exploring the general effects of transcripts harbouring high levels of intra-condition count variation on the outcome of differential expression analysis, we also had an interest in understanding whether or not there were genes commonly up or down regulated within the brain of both forms of domestic canids (dogs and tame foxes), but simultaneously not so within their "wild/aggressive" counter parts (wolves and wild foxes). Such genes are candidates for being associated with tameness. Domestic dogs present marked behaviour differences from wolves, their wild ancestors, due to the evolution of unique social cognitive capabilities [35, 37, 38]. Tame red foxes resulted from deliberated selection against fear and aggression over several generations of cross-breeding [39] and they present several behavioural and pheno-typical traits that resemble those found in dogs [36, 40, 41].

TVscript is open source and code, a quick start guide and test data, are available (under the GNU General Public License) through the SourceForge project page https://sourceforge.net/projects/tvscript/.

Materials & methods

RNA-seq datasets

To explore the effects of intra-condition count variation on the detection of differentially expressed transcripts using real data we used both intra and inter-study datasets. At an intra-study level, we combined publicly available RNA-seq data from brain tissue (prefrontal cortex) of 12 tame and 12 aggressive red foxes (S1 Table) generated within the same study [36]. For the inter-study case, we combined multiple publicly available RNA-seq datasets from several dogs and six wolves [32–35], also derived from brain tissue (frontal cortex, cerebral cortex, prefrontal cortex and frontal lobe) (S1 Table). In relation to the latter, dogs 1 to 6 and wolves 1 to 6 were derived from Albert *et al.*, (2012), dog 6 from Roy *et al.*, 2013, dog 7 (two replicates) from Fushan *et al.*, 2015 and dogs 8 and 9 (three replicates each) from Hoeppner *et al.*, (2014) as described within the table. All the samples were downloaded from the National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EMBL-EBI), covering a wide range of ages, both sexes as well as multiple replicates and sequencing

strategies (S1 Table). We selected these specific case studies because, firstly we were interested in evaluating the effects of intra-condition count variation both at intra and inter-study levels and the domestic dog, being a model organism, has an available high-quality reference transcriptome as well as several high-quality RNA-seq datasets generated across different studies; while secondly, we sought to perform a brief exploratory inter-study scan to investigate if the domestication of both dogs and foxes has resulted in the co-expression of a set of common brain genes, relative to their "wild/aggressive" type, since behavioural modifications are considered to have been the first target in domestication [42].

Reads from all samples were mapped to the dog reference transcriptome [32], which contained 26,107 annotated transcripts (Ensembl CanFam3.1, release 92) [43], using Bowtie v.2.3.4.1 [44]. We did not use a splice-site aware mapper, such as Tophat2 [45] or HISAT2 [46], since introns were not expected to be present within the reference transcriptome. Reads were not mapped to the dog genome as our aim was to explore the effects of intra-condition variation on a predefined set of reference transcripts, and not infer novel transcripts from these previously published datasets. We did not use the fox reference transcriptome (available on Ensembl) for the mapping the fox datasets within our overall analysis as we were interested in directly comparing differentially expressed transcripts from the same underlying reference set identified using the fox, dog and wolf datasets. We did however map all of the fox datasets to the fox reference transcriptome in order to confirm that the proportion of reads mapped was similar to that when using the dog reference transcriptome. The *pileup.sh* script from the BBmap package [47] was used to obtain per transcript abundance estimates (measured by the number of mapped reads to the corresponding transcript) in each sample. Read counts from technical replicates of "Dog_8" and "Dog_9" were averaged and merged into one file (S1 Table), while read counts from the two biological replicates of "Dog_7" were treated separately. Finally, to confirm that the final read count numbers were reliable relative to the dog reference transcriptome, all read datasets from foxes, dogs and wolves were re-mapped using the pseudo-mapper kallisto v0.46.1 that implements a rapid and accurate kmer search based strategy for estimating transcript abundance counts [48]. For each dataset an r^2 correlation value was calculated describing the linear correlation between the per-transcripts counts obtained following Bowtie2 mapping and the corresponding abundance counts obtained using kallisto.

Software

TVscript requires as input: (1) multiple files containing per transcript read counts (one per sample), (2) a file containing the lengths of the transcripts that the reads were mapped to, (3) a percentile threshold value for intra-condition variation and (4) a configuration file that indicates the locations all files as well as the condition allocations of the count files in (1). An example configuration file along with further details is available from the SourceForge project page. The steps that TVscript implements to identify transcripts associated with high levels of disjointed read counts are (see Fig 1 for a workflow): i) each input dataset, containing count values from a particular sample, is allocated to either condition A or B, as indicated within the configuration file; ii) counts are normalized by dividing them by the length of the corresponding reference transcript and by the sum of all counts for that sample; (iii) for each reference transcript (t), the absolute pairwise differences between normalized read counts across all samples within condition A are calculated; (iv) the corresponding variances are calculated; (v) steps (iii) and (iv) are repeated for condition B; (vi) variance scores from each condition are placed in ascending order and associated with corresponding percentiles; (vii) reference transcripts are removed (or filtered) if their variance score is above that associated with the user specified percentile threshold; (viii) raw read counts associated with the remaining transcripts

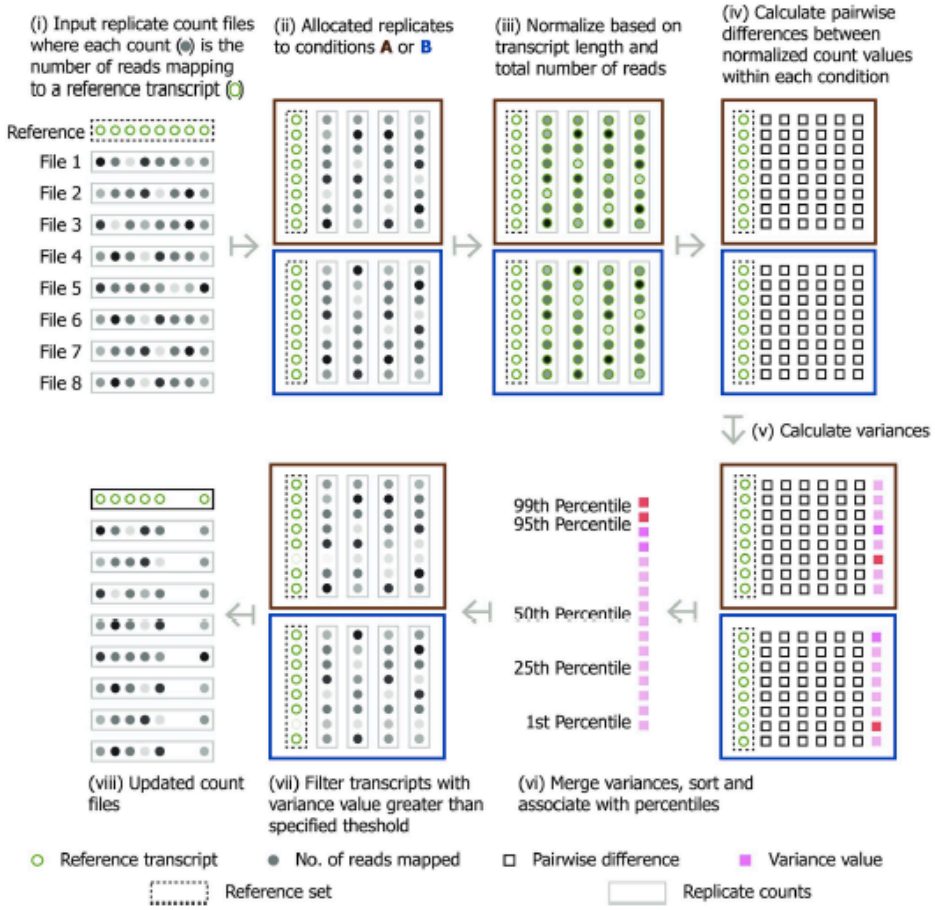


Fig 1. Diagramatic overview of how reference based transcripts are removed by TVscript. Steps (i) to (viii) indicate actions taken. Reference transcripts (green circles) are shown for diagramtic purposes in order to highlight how read counts (grey circles) across replicates are treated for each transcript independently. Read counts are grouped into individual files (gray rectangles) in accordance to replicate. These files are grouped into one of two conditions (blue and brown boxes). Remaining keys are indicated at the bottom.

https://doi.org/10.1371/journal.pone.0274591.g001

are outputted into separate files that correspond to each input dataset. As to avoid overwriting the original count files, the names of updated count files are specified within the configuration file. These updated count files can be used as input for differential expression analysis software such as DESeq2. Note: in relation to step (vi) the final list of variance scores obtained are a representation of those derived from both conditions, and during step (vii) it is the variance score that is associated with the user-defined percentile that is used. The latter means that for

separate comparisons (e.g. wolves *vs*. dogs and aggressive *vs*. tame foxes) at a given filtering threshold the number of filtered transcripts may not be identical (despite the same reference set being used); as the underlying variance score distribution calculated for each can be different.

TVscript is implemented in Java programming language and runs on all operating systems with installed Java Runtime Environment v.8.0 or higher. It is open source and available under the GNU General Public License v3.0. Source code, usage instructions and sample data can be found on the SourceForge project page: https://sourceforge.net/projects/tvscript/. Although TVscript is implemented in Java the steps involved can be readily implemented within any language (e.g. R or python), using the detailed description provided above as well as the Java source code that is fully available. There are no dependent packages where code is unavailable. At the time of development we choose Java mainly due to its platform independence, which can be an advantage within setting up analysis pipelines involving many different tools. That said we are aware that many differential expression analysis tools are R based and future demand may warrant a supported R version.

Controlled intra-condition variation within simulated data

We tested TVscript through a series of iterative differential expression analysis experiments involving highly controlled simulated count data. During each iteration the counts associated with each transcript, within each replicate dataset created, represented simulated transcript expression from the dog reference transcriptome and were obtained using CSReadGen [49]. Using the latter, the level of random background count variation away from that required for a normalized even coverage across all transcripts could be specified, as well as a subset of transcripts to be over represented within replicates of a condition. Background count variation refers to varying count levels associated with individual transcripts that are not maintained across replicates of a condition, thus effectively reflecting intra-condition noise, whilst specifying a subset of transcripts to be over represented across replicates of a condition reflects identifiable over expressed transcripts. A similar experiment to those described here, but in relation to the effects of chimerism on the results of differential expression, has been described in Linheiro and Archer (2021) [50].

As a preliminary, and to demonstrate the general reliability of DESeq2 in the absence of random intra-condition count variation, count data was simulated for 22,580 transcripts ranging in length from 300 to 5000 bp from within the dog reference transcriptome. For a single iteration ten replicates of count dataset, each representing three million read pairs (≈20X coverage), were generated in accordance with two conditions, A and B (five in each), where within condition B one hundred transcripts were selected for count over representation by a factor of two across replicates. For all other transcripts, counts were generated to represent an even distribution of reads (length 150 bp, insert size 300 bp). These count files were used as input to perform differential expression analysis between condition A and B in DESeq2 v.1.32.0, considering transcripts with p-adj < 0.05 (corrected by the Benjamini and Hochberg method) to be differentially expressed. The number of transcripts that were detected as being over-expressed was recorded. This was repeated one hundred times in two different ways: (i) the transcripts initially flagged for count over representation were kept constant throughout and (ii) during each iteration a new set of random transcripts for over representation was selected. A brief overview of the R-script we used for differential expression analysis within individual iterations is available on the Zenodo repository [51].

Next, a similar experiment was performed but where the level of random count variation introduced into the count datasets generated ranged from 1% to 10% in steps of one.

Introduced variation was not coupled between replicates, thus reflected intra-condition variation. At each level of variation one hundred iterations of the following steps were performed. (i) Ten replicates were generated and allocated into two conditions A and B (five in each), where within B one hundred selected transcripts had counts over represented. (ii) At the percent level of variation associated with the iteration, that percent of transcripts from each of the ten replicates were randomly selected for count over representation. (iii) DESeq2 was used in a similar manner to before on the count files within conditions A and B to obtain a list of over expressed transcripts. (iv) TVscript was run using a 95th percentile variance threshold to generate ten corresponding modified count files also separated into two conditions (A' and B'). (v) DESeq2 was again used on these to obtain a list of over-expressed transcripts. (vi) The lists of over-expressed transcripts obtained in (iv) and (v) were cross compared. Once again this was repeated in two different ways: (i) the one hundred transcripts initially flagged for over representation were kept constant throughout all levels of variation and for each of the associated iterations and (ii) during each level of variation and for each iteration a new set of one hundred transcripts were randomly selected.

Exploring the removal of transcripts associated with high intra-condition variation within real data

For each case study (wolves *vs.* dogs and aggressive *vs.* tame fox) we ran TVscript using the count datasets described under the section "RNA-seq datasets" and by applying variance filtering thresholds corresponding to variance values associated with the 70th up to the 90th percentiles (in steps of five), and to the 91st up to the 99th (in steps of one). Steps of one were used in the latter as to allow for transcripts associated with the highest levels of intra-condition variation to be explored in more detail. For each threshold value, only transcripts with variance below that value were maintained. During each run, we recorded the number and IDs of all transcripts that were removed so that they could be cross-compared. Following each run, we used the updated count files produced to perform differential expression analysis using DESeq2 v.1.22.2. Transcripts with p-adj < 0.05 (corrected by the Benjamini and Hochberg method) were considered to be differentially expressed. DESeq2 identifies differentially expressed transcripts by estimating gene-wise dispersions and applying shrinking methods to model counts and thus effectively normalize for individual outliers [31]. Distributions of gene-wise dispersions following normalization are conveniently accessible and provide a good metric to visualize the effects of removing transcripts associated differing filter levels. Differential expression analysis using the original non-filtered datasets was also performed. For the aggressive *vs.* tame fox case study batch effects were not considered as all data came from the same study, tissue and sequencing run, additionally no further information about sample preparation was available. For the wolves *vs.* dogs case study we tested for effects based on tissue, primarily for quality control of the final transcripts we drew biological-related conclusions about, and compared results obtained to those in the absence of batch information. In our analysis we used differential expression results based solely on the latter, as firstly, effects associated with tissue at an inter-study level are unpredictable as there are many factors involved, such as precision of dissection, time of dissection, time to dissect, state of individual tissue samples as well as individual who prepared sample, and other than publication or information mentioned for the fox case study, no further information on batches was obtainable. Secondly, although DESeq2 provides an internalized method for accommodating batch effects that we applied (~batch + condition), the results obtained at an intra-study level, with well defined batches, between alternative methods of testing are variable [52]. Lastly, we were primarily exploring the effects of removing hyper variable transcripts on the mechanics of detecting differentially

expressed transcripts and our simulations and case studies were a means to an end in achieving this. As long as input counts for a given filtering threshold within a given case study or a iteration were consistent with those of the initial input data, the effects of removing hyper variable transcripts could be observed, independent of other factors affecting the data prior to analysis. To visualize the overall effects of covariates broadly affecting the relationships between datasets within each case study, we performed a principal component analysis (PCA) using the *plotPCA* function from DESeq2 with non-filtered normalized count data.

To evaluate TVscript we used three metrics that when combined quantify the overall impact of intra-condition variation on downstream differential expression analysis. The metrics were: i) number of ambiguous positives within transcripts identified as being differentially expressed in the non-filtered datasets; ii) distributions of dispersion estimates and outliers in differential expression analysis for non-filtered and all filtered datasets; and iii) discordance in the list of differentially expressed transcripts between non-filtered and filtered datasets (selected percentiles: 97^{th}, 95^{th}, 90^{th}).

(i) Ambiguous positives. We identified transcripts appearing as being differentially expressed when using the non-filtered datasets as input to DESeq2 that were associated with the top 10 levels of intra-condition variation (above the 90^{th} percentile threshold value). These we designated as ambiguous positives. Small numbers of these, relative to the overall number of identified differentially expressed transcripts would indicate that TVScript is having little direct effect on lists of identified differentially expressed genes.

(ii) Distributions of dispersion estimates. For the non-filtered and all filtered input datasets (70^{th} up to the 90^{th} percentiles in steps of five and to the 91^{st} up to the 99^{th} in steps of one) we calculated correlation coefficients (r^2) using a linear regression analysis in R [53], between dispersion estimates and the mean of normalized counts, both the latter calculated by DESeq2 during differential expression analysis. Dispersion is inversely related to the mean, as lower mean counts are affected by variation to a higher degree. If a stronger correlation is seen for the filtered input datasets, then this would suggest that the distribution used to model differential expression could be more reliable in relation to identifying differentially expressed transcripts. In addition to this, we retrieved the number of outliers detected by DESeq2, expecting a decrease after each filtering step. Outliers are recognized by the DESeq2 as the points with extremely high dispersion values that cannot by shrunk towards the fit curve. This was performed independently for both case studies.

(iii) Discordance lists of differentially expressed transcripts between applied filter levels. We calculated the proportion of discordance between lists of differentially expressed transcripts produced when using non-filtered and filtered datasets at the 97^{th}, 95^{th} and 90^{th} percentile threshold values. Two types of observed discordances relative to the non-filtered list were considered: (a) transcripts that were lost directly due to filtering or indirectly due to p-adj values no longer being significant, and (b) transcripts that were added due to alterations in p-adj values. Quantifying the nature of these discordances provides insight into the general consistency of genes identified as being differentially expressed across varying filter thresholds. To visualize the overlap between the non-filtered and filtered lists we used the *VennDiagram* v.1.6.20 package in R.

Gene annotation and gene family analysis

For each case study, differentially expressed transcripts obtained using the non-filtered and filtered datasets were matched to the correspondent gene ID. This was done with the R package BioMart [54] using the Ensembl Gene database (version 94). To begin to identify gene families that displayed similar regulation in both dogs and tame foxes, i.e. relative to the "wild/

aggressive" type, we grouped up and down regulated genes into gene families. Genes within these families were then classified according to whether they were unique to dogs or tame foxes or shared between the two. Within each case study, a gene family was only considered if all the associated genes agreed in relation to their direction of expression (up or down regulation).

Results

Mapping success of RNA-seq data

Mapping of the 44 datasets corresponding to dogs and wolves against the dog reference transcriptome revealed an average success of 60% and 58% respectively, in terms of the number of mapped reads (S1 Fig). Similar values between wolves and dogs were expected, given their recent divergence of ~23,000 years ago [55]. Comparable proportions of reads failing to map (~40%) have been previously reported for dog brain samples [33] and are most likely associated with i) novel genes; ii) regions that are not translated despite being transcribed; iii) contamination with genomic DNA; and iv) uncharacterized chimeras and other artefacts within reference sets resulting from library preparation during sequencing [56] and various assembly errors [57]. When a different mapping approach was used for each RNA-seq dataset (i.e. kallisto) transcript abundance counts remained consistent with those obtained following Bowtie2 mapping, as indicated by high r^2 correlation values (S2 Fig). R^2 values ranged between 0.8546 and 0.9944. All per-transcript mapped read counts, obtained following each mapping approach, have been made available on the Zenodo repository [58]. For the fox datasets, an average of 50% of reads mapped to the dog reference transcriptome using Bowtie2 (S1 Fig); confirmed by the kallisto estimated abundance counts (S2 Fig). This lower percentage of mapped reads, relative to the dog and wolf datasets, could be expected due to an increased genetic divergence from dogs (~10 mya) [59] together with the other aforementioned factors. However, when these fox datasets were mapped to the Ensembl available fox reference transcriptome using both Bowtie2 an improvement in the overall mapped read counts was not observed (S3 Fig).

Controlled intra-condition variation within simulated data

When not faced with increasing levels of random intra-condition count variation DESeq2 performed exceptionally well. For 86 of the one hundred iterations performed DESeq2 recovered all transcripts that were selected for count over representation (S4a Fig). Of the other 14 iterations the lowest number recovered was 72. Similarly, for the one hundred iterations where the random transcripts selected for count over representation were re-selected during each, in 87 cases all over represented transcripts were identified as being over expressed, whilst in the remaining 13 the minimum number identified was 72 (S4b Fig). However, as levels of introduced intra-condition variation increased, the number of transcripts identified by DESeq2 fell, both before and after filtering the input counts with TVscript (Fig 2). At all levels of intra-condition variation, the post-filtered data had an increase in the number of differentially expressed transcripts identified. It should be emphasized that the reduction in the number of transcripts identified as being over expressed is not a negative reflection on the performance of DESeq2, but instead it is a consequence of purposefully increasing the level of randomness within the count data. The same pattern is true when the one hundred selected transcripts for count over representation are re-selected within each iteration (S5 Fig).

For iterations associated with each increment in random intra-condition variation, the number of transcripts commonly identified as being over expressed both prior-to and post filtering are presented in S2 Table (over represented transcripts kept constant across iterations)

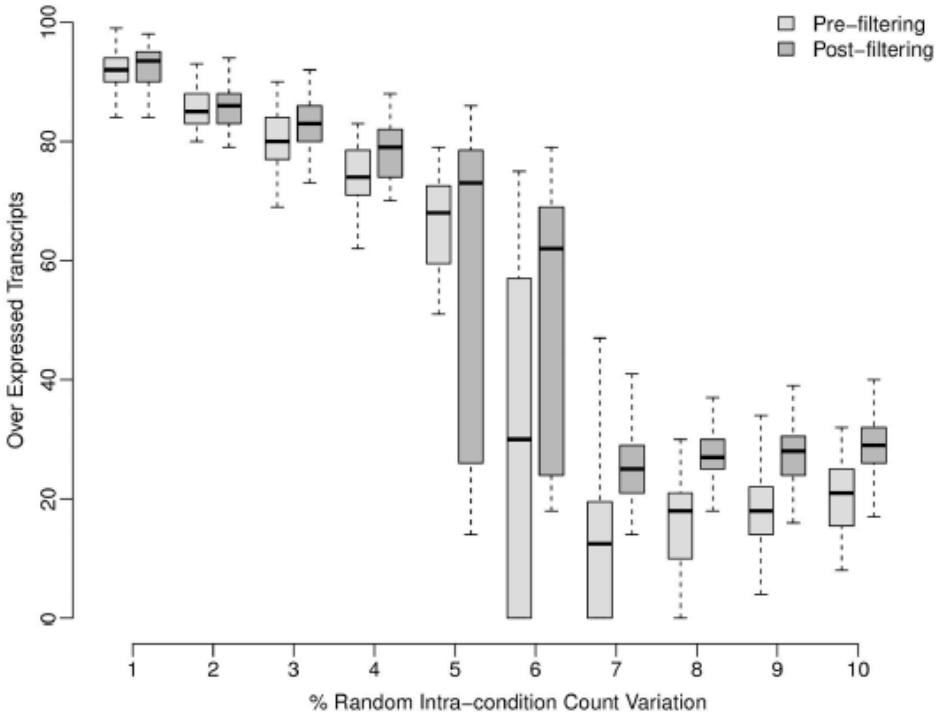


Fig 2. Over expressed transcripts pre- and post-filtering using simulated data. The number of transcripts identified by DESeq2 as being over expressed both prior to (light gray) and post (dark gray) filtering of count datasets within each of the one hundred iterations performed at each level of introduced random intra-condition count variation. Each iteration involved initially simulating ten count datasets divided into conditions A and B following which DESeq2 was run to attempt to identify the one hundred transcripts selected for over representation as described in the methods. Following this the ten simulated datasets were filtered using TVscript with a 95th percentile threshold to generate corresponding filtered datasets (divided into corresponding conditions A' and B') on which DESeq2 was re-run.

https://doi.org/10.1371/journal.pone.0274591.g002

and S3 Table (over represented transcripts re-selected for each iteration). The proportion that these numbers make up relative to the maximum number of transcripts identified as being differentially expressed, pre- and post filtering, are presented in S4 and S5 Tables (constant) (re-selected). In all cases, below a 5% level of random variation these numbers are high (constant—1 to 4% averages: 0.96, 0.98, 0.94 and 0.81; re-selected—1 to 4% averages: 0.96, 0.97, 0.93 and 0.82), indicating that on top of additional transcripts identified post-filtering with TVscript, transcripts identified pre-filtering are still found. Consequently, this suggests that additional transcripts identified as being over expressed post-filtering are not at the expense of previously detected transcripts pre-filtering. Above the 5% level of intra-condition variation the ability to successfully identify the one hundred transcripts selected for over representation within condition B diminishes within iterations (Fig 2; S5 Fig). This could be indicative of a tentative estimate on the limit of at what level of random intra-condition count variation becomes inhibitory within differential expression analysis studies.

Exploring the removal of transcripts associated with high intra-condition variation within real data

No significant difference existed between the overall distributions of the per-transcript intra-condition variation values for wolf and dog samples (Wilcoxon-test, p-value < 0.198, Fig 3a). The PCA based on the entire set of normalized non-filtered counts, revealed that the wolf samples were more aggregated than dog samples (Fig 3c). For aggressive and tame fox samples, we observed a significant difference (Wilcoxon-test, p-value < $2.2e^{-16}$, Fig 3b) between the distributions of the per-transcript intra-condition variation values, most likely resulting from an increased intra-condition variability within tame fox samples. In particular, we found five samples that were differentiated from the remaining seven in the PCA (Fig 3d), with 80% variance being explained by this clustering in PC1.

Prior to differential expression analysis, for each case study (wolves *vs.* dogs and aggressive *vs.* tame foxes), TVscript was used to remove transcripts in accordance with a series of intra-condition variance thresholds (Fig 4a and 4b; S6 Table). Initially, for wolves and dogs 184 transcripts (out of the 26,107) associated with high intra-condition variation (99th percentile and

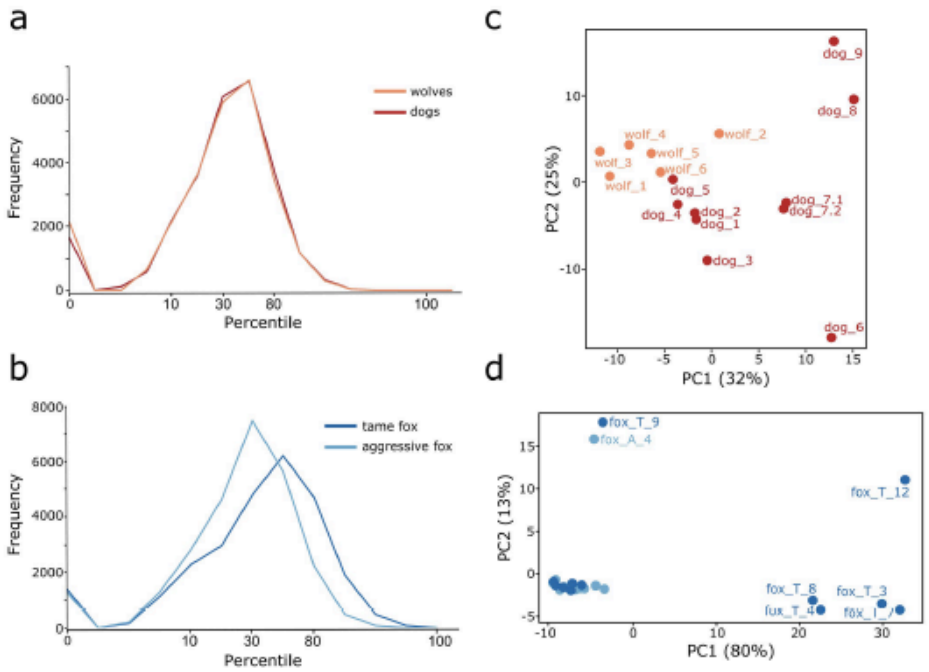


Fig 3. Characterization of intra-condition variation. Percentile range of intra-condition variation scores (x-axis) observed prior to filtering, across both case studies, a) wolves (orange) and dogs (red); b) tame (dark blue) and aggressive (light blue) foxes. PCA plots based on normalized non-filtered count data of the individual datasets comparing c) wolf and dog, and d) tame and aggressive fox. In the latter only individual samples that were positioned within a distant cluster are labelled with the sample ID.

https://doi.org/10.1371/journal.pone.0274591.g003

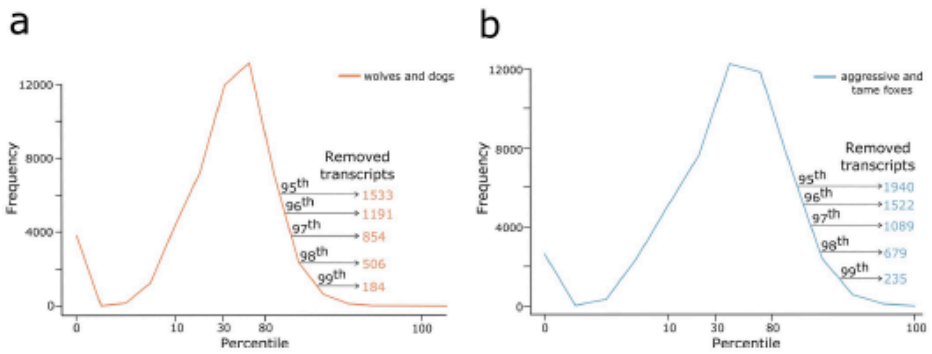


Fig 4. Removal of transcripts above specified levels of intra-condition variation. Percentile range of combined intra-condition variation scores (x-axis) present in each case study, a) wolves and dogs; b) tame and aggressive foxes. The number of transcripts removed in the top five percentiles (from the 95th to the 99th) are presented in each panel.

https://doi.org/10.1371/journal.pone.0274591.g004

above) were removed, while for the aggressive and tame fox samples, 235 transcripts were removed. The number of transcripts removed was higher for the fox samples than for those of wolf and dog, reflecting the higher intra-condition variability present. Combined across the top ten levels of intra-condition variation, 12% (n = 3134) and 14.89% (n = 3888) of the reference transcripts were removed in wolf/dog datasets and aggressive/tame fox datasets respectively (S6 Table).

Differential expression analysis

Using non-filtered datasets as input to DESeq2, 430 differentially expressed transcripts were identified between wolves and dogs (Fig 5a; S7 Table). Of those, 259 were up regulated, while 171 were down regulated in dogs. Between aggressive and tame foxes, 651 differentially expressed transcripts were identified (Fig 5a; S8 Table), of which, 532 and 119 were up and down regulated, respectively, in tame foxes. Post filtering, within the first ten steps of size one from the 99th down to the 90th percentiles, the number of differentially expressed transcripts identified peaks at the 97th (n = 430; up = 255, down = 175) and the 95th percentiles (n = 730; up = 607, down = 123) in dogs and tame foxes (Fig 5a), respectively. This indicates that for these data the removal of the 3% (n = 854) and 5% (n = 1940) of transcripts associated with the highest levels of intra-condition variation maximized the detection of differentially expressed transcripts.

Evaluation metrics

(i) Ambiguous positives. Of the transcripts that appeared as being differentially expressed, when using non-filtered datasets as input to DESeq2, 17.44 (n = 75) and 21.51% (n = 140) were associated with high intra-condition variation (above the 90th percentile threshold) within the wolves *vs*. dogs and aggressive *vs*. tame foxes respectively (Fig 5b). This was higher than the relative proportion of such transcripts with the reference set in general, where 12.08% and 14.89% of transcripts possessed intra-condition variation above the 90th percentile for the wolf/dog and the aggressive/tame fox respectively (Fig 4a and 4b; S6 Table). These were transcripts that we considered as ambiguous positives and the average across both case studies

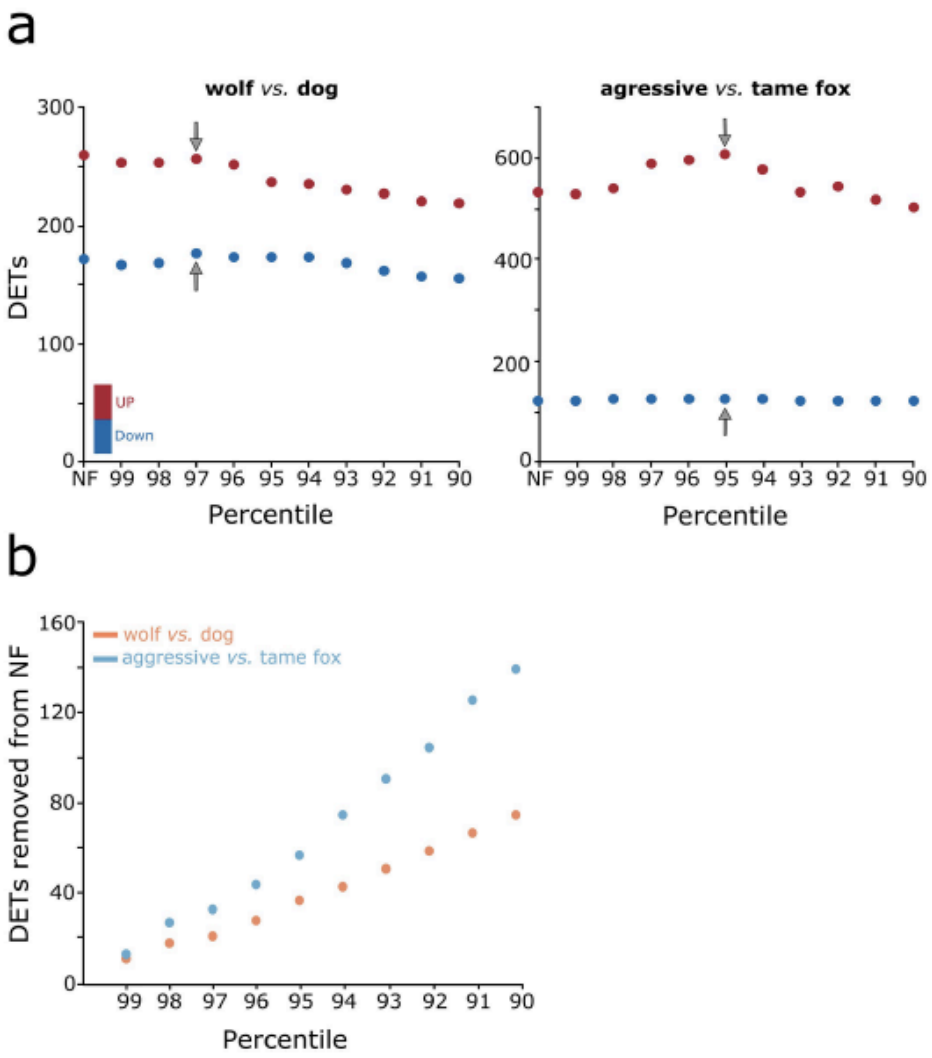


Fig 5. Effects of removing transcripts above specified levels of intra-condition variation on differential expression analysis. a) Number of differentially expressed transcripts (DETs) identified using non-filtered (NF) and filtered datasets, based on the top 10 percentiles (99^{th} to the 90^{th}), for both case studies. Up and down regulated transcripts are represented by red and blue dots respectively. Gray arrows identify the selected thresholds for which the results of subsequent corresponding differential expression analysis were used for the identification of candidate transcript associated with tameness within each case study. b) Number of transcripts identified as differentially expressed within the non-filtered datasets that were associated the highest levels of intra-condition variation (99^{th} to 90^{th}) within both case studies, wolves and dogs (orange dots), and tame and aggressive foxes (blue dots).

https://doi.org/10.1371/journal.pone.0274591.g005

was 19.45%. The number was higher within the fox datasets where elevated variability among samples was observed, suggesting that differences within intra-condition read counts could have influenced the final outcome of identified differentially expressed transcripts.

(ii) Distributions of dispersion estimates. Within both case studies the regression analysis indicates that removing transcripts associated with high levels of intra-condition variation improved correlation coefficients in relation to those from the non-filtered datasets (Fig 6a and 6b; S9 Table). Associated with the elevated levels of variation observed within the fox datasets, there was a better fit within the wolves *vs.* dogs comparison ($r^2 > 0.7$) than that of the aggressive *vs.* tame fox one ($r^2 > 0.5$). For the latter, there was visible elevation in the number of dispersed points around the line of best fit. With the removal of transcripts associated with

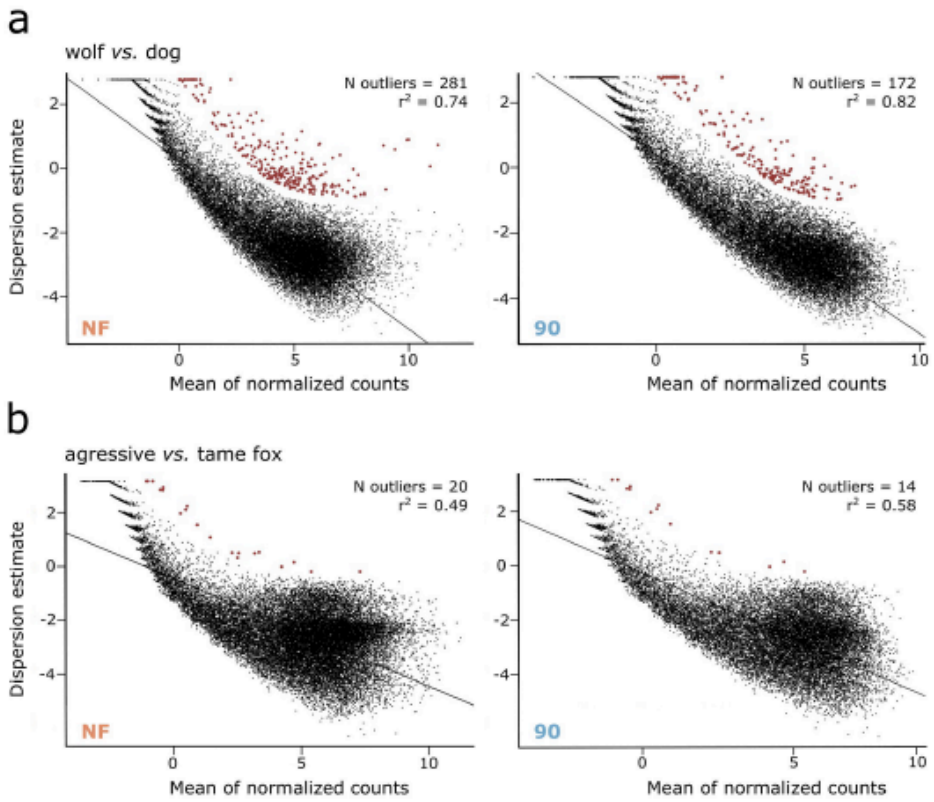
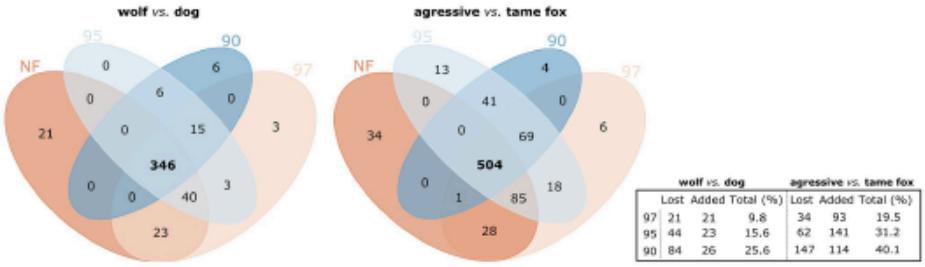


Fig 6. Distribution of dispersion estimates. Plots of final dispersion estimates for both case studies, a) wolves and dogs; b) tame and aggressive foxes, calculated using DESeq2 for the non-filtered (NF; orange) and 10% filtered datasets (90th; blue). Each black dot represents a single transcript, and red dots represent outliers. The number of outliers and correlation index (r^2) are displayed in the top right corner of each panel. Both x and y-axis are transformed into a logarithm scale. The line in each graph corresponds to the regression analysis between the mean of normalized counts and dispersion estimates.

https://doi.org/10.1371/journal.pone.0274591.g006



	wolf vs. dog			agressive vs. tame fox		
	Lost	Added	Total (%)	Lost	Added	Total (%)
97	21	21	9.8	34	93	19.5
95	44	23	15.6	62	141	31.2
90	84	26	25.6	147	114	40.1

Fig 7. Differentially expressed transcripts overlapping between non-filtered and filtered datasets. Venn diagrams representing the number of overlapping differential expressed transcripts found following differential expression analysis using non-filtered datasets and filtered datasets (97th, 95th and 90th percentiles), within each case studies. The inset table provides information about the number of differential expressed transcripts lost/added following each filter step in relation to the non-filtered dataset as well as the the percentage of total discordance.

https://doi.org/10.1371/journal.pone.0274591.g007

the highest levels of intra-condition variation a reduction in the number of outliers within both case studies was also observed (Fig 6a and 6b; S9 Table).

(iii) Discordance lists of differentially expressed transcripts between applied filter levels. Within the wolves *vs*. dogs case study, from the 430 differentially expressed transcripts identified when using non-filtered data as input to DESeq2, 346 were maintained when using input data filtered at the 97th, 95th, and 90th percentile threshold values (Fig 7 and S7 Table). 26 transcripts were added as differentially expressed following filtering. For this case study, the overall discordance between the differentially expressed transcripts identified using filtered and non-filter input data was 25.58% (Fig 7—inset table). For the second case study, aggressive *vs*. tame foxes, 504 out of the 651 differentially expressed transcripts identified using the non-filtered inputs were maintained when using filtered input data at the 97th, 95th, and 90th percentile threshold values, with up to 114 being added following filtering (Fig 7 and S8 Table). This time the overall level of discordance was 40.09% (Fig 7—inset table). Importantly, in both case studies the added transcripts were consistently maintained across the three filter levels. This reflects the general tendency observed within iterative testing using simulated data where the differentially expressed transcripts identified using lower filter levels are maintained at higher levels of filtering in addition to any newly identified transcripts (S4 and S5 Tables).

Candidate genes and gene families

By performing annotation using the filtered datasets where the number of differentially expressed transcripts was maximized (3% and 5%, in dogs and tame foxes, respectively), we found 21 gene families in common among the up-regulated genes in dogs and tame foxes. These 21 gene families contained 50 genes (Table 1), of which 19 were exclusive to dogs while 24 were exclusive to tame foxes. The remaining seven genes (RGR, CHRNA5, SQLE, ARHGAP25, ITGA7, MYO7A and TRIB2), were found to be commonly up regulated in both dogs and tame foxes. When batch effects based on tissue were considered RGR, CHRNA5, MYO7A and TRIB2 were maintained as being commonly up regulated (S6 Fig). Note: although in relation to the latter SQLE, ARHGAP25 and ITGA7 were lost, batch effects based on tissue across multiple studies where little other batch information is obtainable could be considered as a very conservative exclusion.

In addition, we also found three gene families, containing four genes, simultaneously down regulated in both groups (Table 2). Two of these genes (STMND1 and OASL) were shared

Table 1. Shared genes and gene families (Up regulation). List of the gene families, and shared genes, that were commonly up regulated in dogs and tame foxes. The number, and name, of the genes within each gene family are provided, with the corresponding log2fold-change values in brackets for each species. Within each family, single genes were charecterized as shared between dogs and tame foxes (bold), or as exclusively to each of the two groups. When more than one transcript for a specific gene was present, all the log2FC values are reported.

Gene Family	Group	N of genes	Gene name and log2FC value
Retinal G protein-coupled receptor	**Shared**	1	**RGR** (2.10 in dogs, 0.78 in tame foxes)
Cholinergic receptor nicotinic alpha	**Shared**	1	**CHRNA5** (1.1 in dogs, 0.4 in tame foxes)
Squalene epoxidase	**Shared**	1	**SQLE** (0.54 in dogs, 0.31 in tame foxes)
Rho GTPase activating protein	**Shared**	1	**ARHGAP25** (0.86 in dogs, 0.72 in tame foxes)
	Tame fox	2	ARHGAP4 (0.64); ARHGAP30 (0.57)
Integrin alpha subunits	Dog	3	ITGA6 (1.25, 1.24); ITGA8 (1.14, 0.90); ITGAX (0.97)
	Tame fox	1	ITGAL (0.73)
	Shared	1	**ITGA7** (0.76 in dogs, 0.46 and 0.49 in tame foxes)
Myosin	Dog	1	MYO3A (1.12)
	Tame fox	3	MYOZ1 (1.53); MYO1F (0.93); MYO1C (0.47)
	Shared	1	**MYO7A** (0.82 in dogs; 0.41 in tame foxes)
Tribbles pseudokinase	Tame fox	2	TRIB1 (0.94); TRIB3 (0.78)
	Shared	1	**TRIB2** (0.61 in dogs; 0.2 in tame foxes)
EF hand calcium binding	Dog	1	EFCAB1 (2.59)
	Tame fox	1	EFCAB2 (0.46)
Transcription factor	Dog	1	TCF23 (2.04)
	Tame fox	1	TCF19 (0.63)
Adhesion G protein-coupled receptors	Dog	1	ADGRG6 (1.45)
	Tame fox	1	ADGRG1 (0.57)
Patatin Like Phospholipase Domain	Dog	1	PNPLA4 (1.41)
	Tame fox	1	PNPLA7 (0.59)
SRY-box	Dog	1	SOX6 (1.26)
	Tame fox	2	SOX17 (0.84); SOX10 (0.66)
Hyaluronan and proteoglycan link protein	Dog	1	HAPLN1 (1.15)
	Tame fox	1	HAPLN3 (0.70)
Serine/threonine kinase	Dog	2	STK17A (1.15, 1.14); STK32A (1.10)
	Tame fox	1	STK40 (0.57)
Potassium channels	Dog	1	KCTD16 (0.98)
	Tame fox	1	KCTD15 (0.72)
Podocalyxin like	Dog	1	PODXL (0.95, 0.84)
	Tame fox	1	PODXL2 (0.70, 0.69, 0.67)
ATP binding cassette subfamily B	Dog	1	ABCB1 (0.93)
	Tame fox	1	ABCB9 (0.52)
Zinc finger DHHC-type	Dog	1	ZDHHC15 (0.75)
	Tame fox	1	ZDHHC1 (0.70)
Sushi domain	Dog	1	SUSD1 (0.68)
	Tame fox	2	SUSD3 (0.79); SUSD6 (0.47)
TBC1 domain family	Dog	1	TBC1D5 (0.54)
	Tame fox	1	TBC1D7 (0.27)
Mitogen-activated protein kinase kinase kinases	Dog	1	MAP3K5 (0.51)
	Tame fox	1	MAP3K11 (0.76)

https://doi.org/10.1371/journal.pone.0274591.t001

Table 2. Shared genes and gene families (Down regulation). List of the gene families, and shared genes, that were commonly down regulated in dogs and tame foxes. The number, and name, of the genes within each gene family are provided, with the corresponding log2fold-change values in brackets for each species. Within each family, single genes were characterized as shared between dogs and tame foxes, or as exclusively to each of the two groups. When more than one transcript for a specific gene was present, all the log2FC values are reported.

Gene Family	Group	Number of UE	Gene name and log2FC value
Stathmin domain	**Shared**	1	**STMND1** (-1.18 in dogs, -0.53 in tame foxes)
Oligoadenylate synthetase like	**Shared**	1	**OASL** (-0.41 in dogs, -0.52 in tame foxes)
Heat shock protein family B	Dog	1	HSPB8 (-0.70)
	Tame fox	1	HSPB11 (-0.32)

https://doi.org/10.1371/journal.pone.0274591.t002

between dogs and tame foxes, while the other two were unique to each group. When batch effects based on tissue were taken into account, STMND1 and OASL were maintained as being commonly down regulated (S6 Fig). The same analysis performed using the non-filtered datasets revealed similar results (S10 Table), although the RGR gene family which included a shared gene between dogs and tame foxes, was lost. This gene was not differentially expressed in the non-filtered fox dataset, representing an example of genes added as differentially expressed after filtering.

Discussion

Studies involving RNA-seq data often rely on the identification of one, few or many differentially expressed transcripts in order to draw conclusions about biological pathways or about general transcriptome function and evolution. The explicit quantification of intra-condition count variation associated with such transcripts is important for maintaining the context of ambiguity that may exist following differential expression analysis. This is especially true given the growing ability to base highly informative studies around archived transcriptomics datasets at an inter-study level. Here, we developed a method that quantifies intra-condition variation for each individual transcript within the reference set and that can be used to explore the effects of identifying and removing reference-based transcripts harbouring such variation above specified thresholds. By initially applying the method to extensive highly controlled simulated datasets harbouring pre-defined levels of intra-condition count variation we demonstrate the high effectiveness of DESeq2 in identifying differentially expressed transcripts, but also that it can be advantageous to reduce intra-condition variation within the count datasets in relation to identifying additional differentially expressed transcripts that could have been overlooked without such filtering (Fig 2 and S5 Fig, S4 and S5 Tables). By using highly controlled simulated datasets for initial testing, we also provide a tentative estimate on the limit of random intra-condition count variation above which the ability to reliably detect differentially expressed transcripts is diminished (Fig 2 and S5 Fig).

Our real data case study showed that, on average, nearly 20% of the transcripts identified as being differentially expressed prior to filtering contained levels of intra-condition variation equal to or above the 90th percentile value of the total distribution. This was higher than the relative proportion of such transcripts within the reference set and indicates that transcripts associated with higher intra-condition variation have a tendency to being identified as differentially expressed. When transcripts possess large amounts of such variation, some ambiguity in their identification as being differentially expressed is inevitable, since reliable expression patterns for at least one of the two conditions being compared have not been fully established; even if statistical correction is applied. This likely partially explains the level of discordance between various differential expression tools available [15, 23, 26–29], for which no consensus

on the best approach to apply exists [30]. However, more importantly, when such transcripts are used for drawing biological conclusion, the context of this uncertainty must be maintained.

We then explored the effects of removing transcripts associated with intra-condition variation, at varying threshold levels, on the gene-wise dispersion estimates, used by DESeq2. Within both case studies, as such transcripts were increasingly removed from input datasets prior to differential expression analysis, the correlation between the mean of normalized counts and dispersion estimates increased, and the number of outliers identified decreased (Fig 6a and 6b; S9 Table; S7 Fig). This, along with discordances between the lists of differentially expressed transcripts identified prior to and post filtering, suggests that transcripts were not simply removed because of physical exclusion from the input data, but that they were also removed, and added, as a result of the effects of removing intra-condition variation from the general gene-wise dispersions applied. The high rates of discordance we found, reaching 40% within the aggressive *vs*. tame fox case study (Fig 7 and S8 Table), reveal how dependent the identification of differentially expressed transcripts is on the accuracy of gene-wise dispersion estimates used; these in turn being affected by transcripts associated with high intra-condition count variation.

High intra-condition count variation at an inter, and to a lesser extent intra, study level can arise from a range of sources including i) biological differences between samples such as age, sex, diet, and health; ii) *in silica* error involving assembly tools producing poorly understood chimeras within the reference transcriptome [50, 60, 61]; iii) ambiguities in read mapping to such references [62]; iv) normalization of count data derived from such mapped reads [63]; and v) including *in vitro* error during library preparation protocols [64, 65]. Although we used DESeq2 within our study, the results of our exploration on the effects of intra-condition variation in the detection of differentially expressed transcripts likely applies to other software used for differential expression analysis that rely on per transcript count information across replicates for the estimation of transcript abundance and dispersion, for example, edgeR [10], BBSeq [66], DSS [67], baySeq [68] and ShrinkBayes [69].

Following the removal of the 3% and 5% of transcripts associated with the highest levels of variation between wolves and dogs, and aggressive and tame foxes, respectively, we observed an increase in the number of differentially expressed transcripts. This pattern is similar to what we observed within our extensive iterative differential expression analysis experiments on simulated data where the levels of intra-condition variation, as well as sets of count over represented transcripts, were explicitly controlled. Thus, this result suggests that for our case studies the removal of variation at these levels optimized the detection of differentially expressed transcripts whilst maintaining consistency. Using these 3% and 5% cut-offs, amongst the 50 over expressed genes identified, across the 21 shared gene families, seven genes were shared between dogs and tame foxes (Table 1). Of these seven genes, three main functions related to brain development, neurotransmission, and immune response were identified. These functions have been repeatedly associated with behavior selection during domestication by different approaches, such as QTL analysis [40, 70, 71], whole-genome sequencing [72–74], and RNA data both using microarrays and RNA-seq [36, 37, 75–77].

Up until recently, almost no gene overlap had been observed between gene expression profiles involving pairs of domesticated and wild animals [35]. However, a recently published paper performing population genomic and brain transcriptional comparisons in seven bird and mammal domesticated species has revealed a strong convergent pattern in genes implicated in neurotransmission and neuroplasticity [42]. These functions are compatible with those found in our analysis. The shared gene ITGA7 belongs to a gene family that is known to

play an essential role in the control of neuronal connectivity [78] and the inflammatory response [79]. Other genes from this family, for example, ITGA8, have been previously observed to be over expressed in tame foxes [76], and here we also observed its over expression in dogs providing further evidence of the family's role in tameness. Similar functions are associated with the shared genes CHRNA5 [80, 81] and TRIB2 [82] from the cholinergic and tribbles family, respectively. Additionally, we found a shared gene involved in sensing local environmental stimuli, the MYO7A, whose mutation results in loss of hearing and vision [83]. Amongst the three gene families identified as under expressed (Table 2), we found the shared gene STMND1, which deficiency in the amygdala of mice was connected to a deficiency in innate and learned fear [84], a behavior that speculatively could also have an important role in domestication. Although we are aware that this overlap analysis between genes that show the same direction of expression in both dogs and tame foxes is not a formal test for gene convergence, we identified genes involved in several functions previously validated in the scope of domestication.

In this work, we have presented TVscript, a new tool that identifies and removes transcripts associated with high levels of intra-condition variation from RNA-seq count data prior to differential expression analysis. By applying TVscript to simulated data, as well as to real data derived from brain samples of dogs, wolves, tame and aggressive foxes, we demonstrate that as hyper variable transcripts are removed the ability to detect differentially expressed transcripts increases in a robust and repeatable manner. Furthermore, we show that above a certain level of random intra-condition count variation, the identification of differentially expressed transcripts is no longer viable. We propose that studies using RNA-seq data at an inter, or intra, study level should determine whether or not transcripts identified as being differentially expressed, using pre-filtered reference sets, are still identified once filtering based on intra-condition count variation as been performed; regardless of the differential expression software used (or the method of obtaining initial counts). Discussion of such transcripts can then be presented relative to the context of such filtering, thus taking a step forward in reducing the ambiguity surrounding intra-condition count variation. Such context is likely going to be data-set specific, as indicated between differences between our case studies, as the extent of intra-condition count variation will differ between datasets and will rarely be known as a prior to analysis. The latter is further highlighted by the consistent patterns observed during the iterative simulations that we performed where levels of intra-condition variation were pre-specified. Finally, by comparing the genes that were differentially expressed in the brain of dogs and tame foxes, we provided further tentative support for candidate genes involved with several functions long known for being involved with domestication. These genes, and functions, have potential for being involved with selection for tameness, which appears to have played a crucial role in canine domestication. We use the word tentative to describe our support, as the primary aim of this study was to investigate the effects of intra-condition count variation on the detection of differentially expressed transcripts, and the identification of genes involved within an evolutionary process, such as domestication, should be supported by datasets specifically generated for that purpose, and confirmed relative to the different reference transcriptomes involved. The quality of such transcriptomes in turn, in relation to chimeras, missing transcripts and partial redundancies, must also be carefully explored.

Supporting information

S1 Fig. Alignment rates obtained using Bowtie2. Mapping success rates (%) resulting from the alignment of the 44 samples used in this study to the complete dog transcriptome. For each sample, the percentage of aligned reads is presented by the blue bars, while the percentage of

reads failing to map is represented in red (the number of raw reads is available in S1 Table).
(TIF)

S2 Fig. Correlation between per-transcript counts obtained following Bowtie2 mapping and count estimates obtained using kallisto. R^2 values describing the linear correlation between each count dataset produced from the mapped datasets presented in S1 Fig and corresponding count extimates produced when pseudo-mapping the same RNA-Seq data to the complete dog transcriptome using kallisto.
(TIF)

S3 Fig. Re-mapping fox data to the fox reference transcriptome. Read mapping rates achieved when mapping the fox RNA-Seq datasets to the fox reference transcriptome.
(TIF)

S4 Fig. Transcripts identified by DESeq2 as being over expressed in the absence of randomly introduced intra-condition variation. Across one hundred iterations the dots represent the number of transcripts identified as being over expressed between condition A and B. Each condition contained five replicates. (A) The one hundred transcripts selected for read over representation within replicates of condition B were maintained as constant and (B) the one hundred transcripts selected for read over representation within replicates of condition B were re-selected during each iteration. During each iteration the ten count datasets that were simulated each reflected even transcript coverage of 3 million read pairs with the exception of the one hundred transcripts selected for over representation in condition B whose count values were increase by a factor of two.
(TIF)

S5 Fig. Over expressed transcripts pre- and post-filtering (transcripts selected for count over representation were re-selected during each iteration). The number of transcripts identified by DESeq2 as being over expressed both prior to (light gray) and post (dark gray) filtering within each of the one hundred iterations performed at each level of introduced random intra-condition count variation. Each iteration involved initially simulating ten count datasets divided into conditions A and B following which DESeq2 was run to attempt to identify the one hundred transcripts selected for over representation as described in the methods. Following this the ten simulated datasets were filtered using TVScript with a 95^{th} percentile threshold in order to generate corresponding filtered datasets (divided into corresponding conditions A' and B') on which DESeq2 was re-run.
(TIF)

S6 Fig. Confirmation of shared genes within differential expression analysis taking tissue effects into account. The upper dark grey circle contains the nine genes identified as being either commonly over, or under, expressed simultaniously within dogs and tame foxes using filter levels the 95^{th} and 97^{th} percentiles whilst only accounting for condition (wolves *vs.* dogs and aggressive *vs.* tame fox). Six of these genes (RGR, CHRNA5, MYO7A, TRIB2, STMND1 and OASL) are present when DESeq2 is run whilst also accounting for differences in tissue (light grey left oval). SQLE, ARHGAP25 and ITGA7 are observed only within the differentially expressed transcript list that is based solely on condition (dark grey right oval).
(TIF)

S7 Fig. Distribution of dispersion estimates. Plots of dispersion estimates in relation to the mean of normalized counts for both case studies, wolves and dogs (left panels), and tame and aggressive foxes (right panels). Estimates were calculated using DESeq2 for the non-filtered (NF) and all filtered datasets (99^{th}, 95^{th} and 90^{th} are shown as an example). Gray dots represent

the gene-wise maximum likelihood estimates (MLE), the red curve shows the fit to the MLEs, and blue dots identify the final maximum *a posteriori* (MAP) estimates of dispersion. Red dots represent the outliers detected by DESeq2. Both x and y-axis are transformed into a logarithm scale.
(TIF)

S1 Table. Dataset description. Full details of all datasets, including the location of the relative tissue, age, and sex of each individual, replicate information and sequencing details (FC–frontal cortex; CC–cerebral cortex; PFC–prefrontal cortex; FL–frontal lobe; NS–not specified; F–female; M–male; AD–adult; ya–years old; PE–paired-end; SE–single end).
(DOCX)

S2 Table. Common over expressed transcripts pre- and post-filtering (when transcripts selected for count over representation are kept constant). The number of transcripts from the dog reference set that are commonly identified by DESeq2 as being over expressed within condition B both prior to and post filtering for each of the one hundred iterations performed at each level of introduced random intra-condition count variation. Each iteration involved simulating ten count datasets divided into conditions A and B following which DESeq2 was run to attempt to identify the one hundred transcripts selected for over representation as described in the methods section. Filtering involved running TVScript with a 95th percentile threshold on the non-filtered datasets to generate corresponding filtered datasets (divided into corresponding conditions A' and B') following which DESeq2 was re-run and the results compared back to those obtained for the non filtered data.
(DOCX)

S3 Table. Common over expressed transcripts pre- and post-filtering (when transcripts selected for count over representation are re-selected during each iteration). Same as S2 Table but where the one hundred transcripts selected for over representation within condition B are re-selected during each iteration.
(DOCX)

S4 Table. Ratio between the common number of over expressed transcripts pre- and post-filtering and the maximum number detected when transcripts selected for count over representation are kept constant. Numbers in S2 Table were divided by the maximum number of over expressed transcripts detected within each correspomding iteration i.e. the maximum number detected using corresponding non-filtered and filtered datasets.
(DOCX)

S5 Table. Ratio between the common number of over expressed transcripts pre- and post-filtering and the maximum number detected when transcripts selected for count over representation are re-selected during each iteration. Numbers in S3 Table were divided by the maximum number of over expressed transcripts detected within each correspomding iteration i.e. the maximum number detected using corresponding non-filtered and filtered datasets.
(DOCX)

S6 Table. Removal of intra-condition variation. Number of transcripts kept and removed from the reference in each case study, wolves and dogs, and aggressive and tame foxes, across the filtered levels used (from the 99th to the 70th percentile). The first ten percentiles were explored in greater detail in steps of one, while the remaining were performed in steps of 5.
(DOCX)

S7 Table. Differentially expressed transcripts in wolf *vs*. dog. Complete list of differentially expressed transcripts in dogs when compared to wolves, identified using non-filtered datasets, and those that got removed (red) within the highest 10% of intra-condition variation, as well as those added (green) as differentially expressed across selected filtered datasets (97^{th}, 95^{th}, and 90^{th} percentiles). The correspondent annotated gene ID, log2FC values and p-values are provided.
(DOCX)

S8 Table. Differentially expressed transcripts in aggressive *vs*. tame fox. Complete list of differentially expressed transcripts in tame foxes when compared to aggressive foxes, identified using non-filtered datasets, and those that got removed (red) within the highest 10% of intra-condition variation, as also those added (green) as differentially expressed across selected filtered datasets (97^{th}, 95^{th}, and 90^{th} percentiles). The correspondent annotated gene ID, log2FC values and p-values are provided.
(DOCX)

S9 Table. Correlation and outliers. Correlation values (r^2) and the root mean square error (RMSE) from the regression analysis between the final dispersion estimates and the mean of normalized counts for both case studies, wolves and dogs, and aggressive and tame foxes. The number of outliers identified by DESeq2 are also presented. Values are shown for the non-filtered (NF) and all the filtered datasets used in differential expression analysis.
(DOCX)

S10 Table. Shared genes and gene families between non-filtered datasets. List of the gene families, and shared genes, that were commonly regulated in dogs and tame foxes, using the non-filtered datasets. The number, and name, of the genes within each gene family are provided, with the corresponding log2fold-change values in brackets for each species. Within each family, single genes were charecterized as shared between dogs and tame foxes, or as exclusive to each of the two groups. When more than one transcript for a specific gene was present, all the log2FC values are reported.
(DOCX)

Author Contributions

Conceptualization: Diana Lobo, Raquel Godinho, John Patrick Archer.

Data curation: Diana Lobo, Raquel Godinho.

Formal analysis: Diana Lobo, Raquel Linheiro, John Patrick Archer.

Funding acquisition: Raquel Godinho, John Patrick Archer.

Investigation: Diana Lobo, Raquel Godinho, John Patrick Archer.

Methodology: Diana Lobo, John Patrick Archer.

Project administration: John Patrick Archer.

Resources: Raquel Godinho.

Software: Diana Lobo, John Patrick Archer.

Supervision: Raquel Godinho, John Patrick Archer.

Validation: Diana Lobo, Raquel Linheiro, John Patrick Archer.

Visualization: Diana Lobo, Raquel Linheiro.

Writing – original draft: Diana Lobo, Raquel Godinho, John Patrick Archer.

Writing – review & editing: Diana Lobo, Raquel Linheiro, Raquel Godinho, John Patrick Archer.

References

1. Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. Nat Rev Genet. 2009; 10: 57–63. https://doi.org/10.1038/nrg2484 PMID: 19015660
2. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016; 17. https://doi.org/10.1186/s13059-016-0881-8 PMID: 26813401
3. Kukurba KR, Montgomery SB. RNA Sequencing and Analysis. Cold Spring Harb Protoc. 2015; 2015: 951. https://doi.org/10.1101/pdb.top084970 PMID: 25870306
4. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008 57. 2008; 5: 621–628. https://doi.org/10.1038/nmeth.1226 PMID: 18516045
5. Lee S, Li S, Seo CH, Lim B, Yang JO, Oh J, et al. Accurate quantification of transcriptome from RNA-Seq data by effective length normalization. Nucleic Acids Res. 2011; 39. https://doi.org/10.1093/nar/gkq1015 PMID: 21059678
6. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010; 28: 511–515. https://doi.org/10.1038/nbt.1621 PMID: 20436464
7. Jiang H, Wong WH. Statistical inferences for isoform expression in RNA-Seq. Bioinformatics. 2009; 25: 1026–1032. https://doi.org/10.1093/bioinformatics/btp113 PMID: 19244387
8. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science (80-). 2008; 320: 1344–1349. https://doi.org/10.1126/science.1158441 PMID: 18451266
9. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15:550. https://doi.org/10.1186/s13059-014-0550-8 PMID: 25516281
10. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26: 139–140. https://doi.org/10.1093/bioinformatics/btp616 PMID: 19910308
11. Hansen KD, Wu Z, Irizarry RA, Leek JT. Sequencing technology does not eliminate biological variability. Nat Biotechnol. 2011; 29: 572–573. https://doi.org/10.1038/nbt.1910 PMID: 21747377
12. Xu Z, Asakawa S. Physiological RNA dynamics in RNA-Seq analysis. Brief Bioinform. 2019; 20: 1725–1733. https://doi.org/10.1093/bib/bby045 PMID: 30010714
13. McIntyre LM, Lopiano KK, Morse AM, Amin V, Oberg AL, Young LJ, et al. RNA-seq: Technical variability and sampling. BMC Genomics. 2011; 12:293. https://doi.org/10.1186/1471-2164-12-293 PMID: 21645359
14. Liu Y, Zhou J, White KP. RNA-seq differential expression studies: More sequence or more replication? Bioinformatics. 2014; 30: 301–304. https://doi.org/10.1093/bioinformatics/btt688 PMID: 24319002
15. Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA. 2016; 22: 839–851. https://doi.org/10.1261/rna.053959.115 PMID: 27022035
16. Robles JA, Qureshi SE, Stephen SJ, Wilson SR, Burden CJ, Taylor JM. Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. BMC Genomics. 2012; 13: 484. https://doi.org/10.1186/1471-2164-13-484 PMID: 22985019
17. Lachmann A, Torre D, Keenan AB, Jagodnik KM, Lee HJ, Wang L, et al. Massive mining of publicly available RNA-seq data from human and mouse. Nat Commun 2018 91. 2018; 9: 1–10. https://doi.org/10.1038/s41467-018-03751-6 PMID: 29636450
18. Zoabi Y, Shomron N. Processing and Analysis of RNA-seq Data from Public Resources. Methods in Molecular Biology. Methods Mol Biol; 2021. pp. 81–94. https://doi.org/10.1007/978-1-0716-1103-6_4 PMID: 33606253
19. Sudmant PH, Alexis MS, Burge CB. Meta-analysis of RNA-seq expression data across species, tissues and studies. Genome Biol 2015 161. 2015; 16: 1–11. https://doi.org/10.1186/s13059-015-0853-4 PMID: 26694591

20. Rau A, Marot G, Jaffrézic F. Differential meta-analysis of RNA-seq data from multiple studies. BMC Bioinformatics. 2014; 15:91. https://doi.org/10.1186/1471-2105-15-91 PMID: 24678808
21. Jeng SL, Chi YC, Ma MC, Chan SH, Sun HS. Gene expression analysis of combined RNA-seq experiments using a receiver operating characteristic calibrated procedure. Comput Biol Chem. 2021; 93: 107515. https://doi.org/10.1016/j.compbiolchem.2021.107515 PMID: 34044204
22. McDermaid A, Monier B, Zhao J, Liu B, Ma Q. Interpretation of differential gene expression results of RNA-seq data: review and integration. Brief Bioinform. 2019; 20: 2044. https://doi.org/10.1093/bib/bby067 PMID: 30099484
23. Quinn TP, Crowley TM, Richardson MF. Benchmarking differential expression analysis tools for RNA-Seq: normalization-based vs. log-ratio transformation-based methods. BMC Bioinformatics. 2018; 19: 274. https://doi.org/10.1186/s12859-018-2261-8 PMID: 30021534
24. Risso D, Schwartz K, Sherlock G, Dudoit S. GC-Content Normalization for RNA-Seq Data. BMC Bioinformatics. 2011. https://doi.org/10.1186/1471-2105-12-480 PMID: 22177264
25. Stegle O, Parts L, Durbin R, Winn J. A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput Biol. 2010; 6(5): e100. https://doi.org/10.1371/journal.pcbi.1000770 PMID: 20463871
26. Li S, Labaj PP, Zumbo P, Sykacek P, Shi W, Shi L, et al. Detecting and correcting systematic variation in large-scale RNA sequencing data. Nat Biotechnol. 2014; 32: 888–895. https://doi.org/10.1038/nbt.3000 PMID: 25150837
27. Wang T, Li B, Nelson CE, Nabavi S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. BMC Bioinformatics. 2019; 20: 40. https://doi.org/10.1186/s12859-019-2599-6 PMID: 30658573
28. Mou T, Deng W, Gu F, Pawitan Y, Vu TN. Reproducibility of Methods to Detect Differentially Expressed Genes from Single-Cell RNA Sequencing. Front Genet. 2020; 10. https://doi.org/10.3389/fgene.2019.01331 PMID: 32010190
29. Arora S, Pattwell SS, Holland EC, Bolouri H. Variability in estimated gene expression among commonly used RNA-seq pipelines. Sci Rep. 2020; 10: 2734. https://doi.org/10.1038/s41598-020-59516-z PMID: 32066774
30. Stupnikov A, McInerney CE, Savage KI, McIntosh SA, Emmert-Streib F, Kennedy R, et al. Robustness of differential gene expression analysis of RNA-seq. Comput Struct Biotechnol J. 2021; 19: 3470–3481. https://doi.org/10.1016/j.csbj.2021.05.040 PMID: 34188784
31. Costa-Silva J, Domingues D, Lopes FM. RNA-Seq differential expression analysis: An extended review and a software tool. PLoS One. 2017; 12:e019015. https://doi.org/10.1371/journal.pone.0190152 PMID: 29267363
32. Hoeppner MP, Lundquist A, Pirun M, Meadows JRS, Zamani N, Johnson J, et al. An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. PLoS One. 2014; 9(3):91172. https://doi.org/10.1371/journal.pone.0091172 PMID: 24625832
33. Roy M, Kim N, Kim K, Chung WH, Achawanantakun R, Sun Y, et al. Analysis of the canine brain transcriptome with an emphasis on the hypothalamus and cerebral cortex. Mamm Genome. 2013; 24: 484–499. https://doi.org/10.1007/s00335-013-9480-0 PMID: 24202129
34. Fushan AA, Turanov AA, Lee SG, Kim EB, Lobanov AV, Yim SH, et al. Gene expression defines natural changes in mammalian lifespan. Aging Cell. 2015; 14: 352–365. https://doi.org/10.1111/acel.12283 PMID: 25677554
35. Albert FW, Somel M, Carneiro M, Aximu-Petri A, Halbwax M, Thalmann O, et al. A Comparison of Brain Gene Expression Levels in Domesticated and Wild Animals. Akey JM, editor. PLoS Genet. 2012; 8: e1002962. https://doi.org/10.1371/journal.pgen.1002962 PMID: 23028369
36. Wang X, Pipes L, Trut L, Herbeck Y, Vladimirova A, Gulevich R, et al. Genomic responses to selection for tame/aggressive behaviors in the silver fox (Vulpes vulpes). Proc Natl Acad Sci. 2018; 115: 10398–10403. https://doi.org/10.1073/pnas.1800889115 PMID: 30228118
37. Li Y, Wang GD, Wang MS, Irwin DM, Wu DD, Zhang YP. Domestication of the dog from the Wolf was promoted by enhanced excitatory synaptic plasticity: A hypothesis. Genome Biol Evol. 2014; 6: 3115–3121. https://doi.org/10.1093/gbe/evu245 PMID: 25377939
38. Li Y, VonHoldt BM, Reynolds A, Boyko AR, Wayne RK, Wu DD, et al. Artificial selection on brain-expressed genes during the domestication of dog. Mol Biol Evol. 2013; 30: 1867–1876. https://doi.org/10.1093/molbev/mst088 PMID: 23660689
39. Lord KA, Larson G, Coppinger RP, Karlsson EK. The History of Farm Foxes Undermines the Animal Domestication Syndrome. Trends Ecol Evol. 2019. https://doi.org/10.1016/j.tree.2019.10.011 PMID: 31810775

40. Kukekova A, Trut L, Chase K, Kharlamova A, Johnson J, Temnykh S, et al. Mapping loci for fox domestication: Deconstruction/Reconstruction of a behavioral phenotype. Behav Genet. 2011; 41: 593–606. https://doi.org/10.1007/s10519-010-9418-1 PMID: 21153916

41. Hekman J, Johnson J, Edwards W, Vladimirova A, Gulevich R, Ford A, et al. Anterior Pituitary Transcriptome Suggests Differences in ACTH Release in Tame and Aggressive Foxes. G3; Genes|Genomes|Genetics. 2018; 8: 859–873. https://doi.org/10.1534/g3.117.300508 PMID: 29378821

42. Hou Y, Qi F, Bai X, Ren T, Shen X, Chu Q, et al. Genome-wide analysis reveals molecular convergence underlying domestication in 7 bird and mammals. BMC Genomics. 2020; 21: 1–20. https://doi.org/10.1186/s12864-020-6613-1 PMID: 32131728

43. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al. Ensembl 2020. Nucleic Acids Res. 2020; 48: D682–D688. https://doi.org/10.1093/nar/gkz966 PMID: 31691826

44. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012. https://doi.org/10.1038/nmeth.1923 PMID: 22388286

45. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013; 14: 1–13. https://doi.org/10.1186/gb-2013-14-4-r36 PMID: 23618408

46. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 2019 378. 2019; 37: 907–915. https://doi.org/10.1038/s41587-019-0201-4 PMID: 31375807

47. Bushnell, Brian. BBMap: A Fast, Accurate, Splice-Aware Aligner. Conference: 9th Annual Genomics of Energy Environment Meeting. 2014.

48. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 2016 345. 2016; 34: 525–527. https://doi.org/10.1038/nbt.3519 PMID: 27043002

49. Linheiro R, Archer J. CSReadGen website. https://sourceforge.net/projects/csreadgen/

50. Linheiro R, Archer J. CStone: A de novo transcriptome assembler for short-read data that identifies non-chimeric contigs based on underlying graph structure. PLOS Comput Biol. 2021; 17: e1009631. https://doi.org/10.1371/journal.pcbi.1009631 PMID: 34813594

51. Lobo D, Linheiro R, Godinho R, Archer J. On taming the effect of transcript level intra-condition count variation during differential expression analysis: a story of dogs, foxes and wolves: Example R Script for using DESeq2. 2022 [cited 22 Jun 2022]. https://doi.org/10.5281/ZENODO.6676483

52. Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol. 2020; 21: 1–32. https://doi.org/10.1186/s13059-019-1850-9 PMID: 31948481

53. RDevelopment Core Team. R: A language and environment for statistical computing. Vienna, Austria. 2017. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

54. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. Bioinformatics. 2005; 21: 3439–3440. https://doi.org/10.1093/bioinformatics/bti525 PMID: 16082012

55. Perri AR, Feuerborn TR, Frantz LAF, Larson G, Malhi RS, Meltzer DJ, et al. Dog domestication and the dual dispersal of people and dogs into the Americas. Proc Natl Acad Sci U S A. 2021; 118: 1–8. https://doi.org/10.1073/pnas.2010083118 PMID: 33495362

56. Tu J, Guo J, Li J, Gao S, Yao B, Lu Z. Systematic Characteristic Exploration of the Chimeras Generated in Multiple Displacement Amplification through Next Generation Sequencing Data Reanalysis. PLoS One. 2015; 6: e0139857. https://doi.org/10.1371/journal.pone.0139857 PMID: 26440104

57. Rizzi R, Beretta S, Patterson M, Pirola Y, Previtali M, Della Vedova G, et al. Overlap graphs and de Bruijn graphs: data structures for de novo genome assembly in the big data era. Quant Biol. 2019; 7: 278–292. https://doi.org/10.1007/s40484-019-0181-x

58. Lobo D, Linheiro R, Godinho R, Archer J. On taming the effect of transcript level intra-condition count variation during differential expression analysis: a story of dogs, foxes and wolves: Bowtie2 counts and kallisto abundances. 2022 [cited 29 Jun 2022]. https://doi.org/10.5281/ZENODO.6778429

59. Wayne RK, Geffen E, Girman DJ, Koepfli K-P, Lau LM, Marshall CR. Molecular Systematics of the Canidae. Syst Biol. 1997; 46: 622–653. https://doi.org/10.1093/sysbio/46.4.622 PMID: 11975336

60. Linheiro R, Archer J. Quantification of the effects of chimerism on read mapping, differential expression and annotation following short-read *de novo* assembly. F1000Research 2022 11120. 2022; 11: 120. https://doi.org/10.12688/f1000research.108489.1

61. Hsieh PH, Oyang YJ, Chen CY. Effect of de novo transcriptome assembly on transcript quantification. Sci Rep. 2019; 9. https://doi.org/10.1038/s41598-019-44499-3 PMID: 31165774

62. Reinert K, Langmead B, Weese D, Evers DJ. Alignment of Next-Generation Sequencing Reads. Annu Rev Genomics Hum Genet. 2015; 16: 133–151. https://doi.org/10.1146/annurev-genom-090413-025358 PMID: 25939052

63. Evans C, Hardin J, Stoebel DM. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Brief Bioinform. 2018; 19: 776–792. https://doi.org/10.1093/bib/bbx008 PMID: 28334202

64. Brodin J, Mild M, Hedskog C, Sherwood E, Leitner T, Andersson B, et al. PCR-Induced Transitions Are the Major Source of Error in Cleaned Ultra-Deep Pyrosequencing Data. PLoS One. 2013; 8. https://doi.org/10.1371/journal.pone.0070388 PMID: 23894647

65. Ma X, Shao Y, Tian L, Flasch DA, Mulder HL, Edmonson MN, et al. Analysis of error profiles in deep next-generation sequencing data. Genome Biol. 2019; 20. https://doi.org/10.1186/s13059-019-1659-6 PMID: 30867008

66. Zhou YH, Xia K, Wright FA. A powerful and flexible approach to the analysis of RNA sequence count data. Bioinformatics. 2011. https://doi.org/10.1093/bioinformatics/btr449 PMID: 21810900

67. Wu H, Wang C, Wu Z. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. Biostatistics. 2013; 14: 232–243. https://doi.org/10.1093/biostatistics/kxs033 PMID: 23001152

68. Hardcastle TJ, Kelly KA. BaySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010; 11:422. https://doi.org/10.1186/1471-2105-11-422 PMID: 20698981

69. Van De Wiel MA, Leday GGR, Pardo L, Rue H, Van Der Vaart AW, Van Wieringen WN. Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. Biostatistics. 2013; 14: 113–128. https://doi.org/10.1093/biostatistics/kxs031 PMID: 22988280

70. Wirén A, Wright D, Jensen P. Domestication-related variation in social preferences in chickens is affected by genotype on a growth QTL. Genes, Brain Behav. 2013; 12: 330–337. https://doi.org/10.1111/gbb.12017 PMID: 23331324

71. Albert FW, Carlborg Ö, Plyusnina I, Besnier F, Hedwig D, Lautenschläger S, et al. Genetic architecture of tameness in a rat model of animal domestication. Genetics. 2009; 182: 541–554. https://doi.org/10.1534/genetics.109.102186 PMID: 19363126

72. Carneiro M, Rubin CJ, Di Palma F, Albert FW, Alföldi J, Barrio AM, et al. Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. Science (80-). 2014; 345: 1074–1079. https://doi.org/10.1126/science.1253714 PMID: 25170157

73. Freedman AH, Schweizer RM, Ortega-Del Vecchyo D, Han E, Davis BW, Gronau I, et al. Demographically-Based Evaluation of Genomic Regions under Selection in Domestic Dogs. PLoS Genet. 2016; 12: e100585. https://doi.org/10.1371/journal.pgen.1005851 PMID: 26943675

74. Kukekova A, Johnson J, Xiang X-Y, Feng S-H, Liu S, Rando H, et al. The red fox genome assembly identifies genomic regions associated with tame and aggressive behaviors. Nat Ecol Evol. 2018; 2: 1479–1491. https://doi.org/10.1038/s41559-018-0611-6 PMID: 30082739

75. Saetre P, Lindberg J, Ellegren H, Vilà C, Jazin E, Leonard JA, et al. From wild wolf to domestic dog: Gene expression changes in the brain. Mol Brain Res. 2004; 126: 198–206. https://doi.org/10.1016/j.molbrainres.2004.05.003 PMID: 15249144

76. Kukekova A, Johnson J, Teiling C, Li L, Oskina I, Kharlamova A, et al. Sequence comparison of prefrontal cortical brain transcriptome from a tame and an aggressive silver fox (Vulpes vulpes). BMC Genomics. 2011; 12: 482. https://doi.org/10.1186/1471-2164-12-482 PMID: 21967120

77. Heyne HO, Lautenschläger S, Nelson R, Besnier F, Rotival M, Cagan A, et al. Genetic influences on brain gene expression in rats selected for tameness and aggression. Genetics. 2014; 198: 1277–1290. https://doi.org/10.1534/genetics.114.168948 PMID: 25189874

78. Lilja J, Ivaska J. Integrin activity in neuronal connectivity. J Cell Sci. 2018; 131: jcs212. https://doi.org/10.1242/jcs.212803 PMID: 29907643

79. González-Amaro R, Sánchez-Madrid F. Cell adhesion molecules: selectins and integrins. Crit Rev Immunol. 1999; 19: 389–429. PMID: 10647744

80. Winterer G, Mittelstrass K, Giegling I, Lamina C, Fehr C, Brenner H, et al. Risk gene variants for nicotine dependence in the CHRNA5-CHRNA3-CHRNB4 cluster are associated with cognitive performance. Am J Med Genet Part B Neuropsychiatr Genet. 2010; 153: 1448–1458. https://doi.org/10.1002/ajmg.b.31126 PMID: 20886544

81. Zhang H, Kranzler HR, Poling J, Gruen JR, Gelernter J. Cognitive flexibility is associated with KIBRA variant and modulated by recent tobacco use. Neuropsychopharmacology. 2009; 34: 2508–2516. https://doi.org/10.1038/npp.2009.80 PMID: 19606085

82. Eyers PA, Keeshan K, Kannan N. Tribbles in the 21st Century: The Evolving Roles of Tribbles Pseudokinases in Biology and Disease. Trends Cell Biol. 2017; 27: 284–298. https://doi.org/10.1016/j.tcb.2016.11.002 PMID: 27908682

83. Miller KA, Williams LH, Rose E, Kulper M, Dahl HHM, Manji SSM. Inner Ear Morphology Is Perturbed in Two Novel Mouse Models of Recessive Deafness. PLoS One. 2012; 7(12):e512. https://doi.org/10.1371/journal.pone.0051284 PMID: 23251483

84. Martel G, Nishi A, Shumyatsky GP. Stathmin reveals dissociable roles of the basolateral amygdala in parental and social behaviors. Proc Natl Acad Sci U S A. 2008; 105: 14620–14625. https://doi.org/10.1073/pnas.0807507105 PMID: 18794533