U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Historical Demography and Differentiation of the Gray Wolf (*Canis lupus*)

Pedro Manuel Soares da Silva

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

Biologia

2016

Historical Demography and Differentiation of the Gray Wolf (*Canis lupus*)

Pedro Manuel Soares da Silva
Programa Doutoral em Biodiversidade, Genética e Evolução
Departamento de Biologia
2016

Orientadora
Raquel Godinho, Professora Auxiliar Convidada e Investigadora,
Departamento de Biologia, Faculdade de Ciências da Universidade do Porto

Coorientador
Robert Wayne, Professor,
Department of Ecology and Evolutionary Biology, University of California, Los Angeles

Nota Prévia

Na elaboração desta tese, e nos termos do número 2 do Artigo 4º do Regulamento Geral dos Terceiros Ciclos de Estudos Universidade do Porto e do Artigo 31º do Decreto-Lei 74/2006, de 24 de Março, com a nova redação introduzida pelo Decreto-Lei 230/2009, de 14 de Setembro, foi efetuado o aproveitamento total de um conjunto coerente de trabalhos de investigação já publicados ou submetidos para publicação em revistas internacionais indexadas e com arbitragem científica, os quais integram alguns dos capítulos da presente tese. Tendo em conta que os referidos trabalhos foram realizados com a colaboração de outros autores, o candidato esclarece que, em todos eles, participou ativamente na sua conceção, na obtenção, análise e discussão de resultados, bem como na elaboração da sua forma publicada.

A Faculdade de Ciências da Universidade do Porto foi a instituição de origem do candidato, tendo o trabalho sido realizado sob orientação da 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) e sob coorientação do Doutor Robert Wayne, Professor do Departamento de Ecologia and Biologia Evolutiva da Universidade da Califórnia, Los Angeles (EEB, UCLA). Os trabalhos foram realizados no CIBIO-InBio e no EEB, UCLA, que funcionaram como entidades de acolhimento.

Este trabalho foi apoiado pela Fundação para a Ciência e Tecnologia (FCT) através da bolsa de doutoramento SFRH/BD/60549/2009

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

POPH QUALIFICAR É CRESCER. | QREN QUADRO DE REFERÊNCIA ESTRATÉGICO NACIONAL PORTUGAL 2007.2013 | GOVERNO DA REPÚBLICA PORTUGUESA | UNIÃO EUROPEIA Fundo Social Europeu

CIBIO Centro de Investigação em Biodiversidade e Recursos Genéticos

U.PORTO FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

Agradecimentos / Acknowledgments

Sem o precioso apoio de várias pessoas este trabalho não teria sido possível. Quero aqui deixar-lhes os meus sinceros agradecimentos, desde já pedindo desculpa a quem eventualmente ficar esquecido.

/

Without the precious support of several people this work would not have been possible. I would like to leave them here my sincerest thanks, apologizing in advance in case someone is forgotten.

Aos meus pais devo tudo. Muito obrigado pelo amor e apoio incondicionais que sempre me deram durante toda a minha vida. O respeito e compreensão que sempre demonstraram pelas minhas escolhas foram essenciais para ultrapassar as fases mais difíceis destes últimos anos.

À minha orientadora Raquel Godinho tenho a agradeçer a oportunidade de ter desenvolvido este trabalho. Muito obrigado pela confiança depositada em mim e nas minhas capacidades para levar este projecto avante, mesmo nos momentos em que essa confiança me faltava a mim próprio. O teu contínuo apoio, encorajamento e paciência foram essenciais para me manter no rumo certo.

To my co-supervisor Bob Wayne I would like to thank the availability with which I was received at UCLA, and the opportunity to work on several interesting projects. The time spent in your lab was a period of intense work and learning, and I firmly believe it made me grow as a person and a researcher.

I am also greatly indebted to the excellent researchers with which I had the pleasure of collaborating at UCLA, namely Adam Freedman, John Novembre and John Pollinger. In your own different ways, each of you showed me by example what it means to be a great scientist today.

I would like to thank CIBIO-InBIO and the Ecology and Evolutionary Biology Department at UCLA for hosting me during during all these years. I would also like to acknowledge the finantial support of Fundação para a Ciência e Tecnologia through the PhD grant SFRH/BD/60549/2009.

This work would not have been possible, and would certainly have been much less fun, without my dear colleagues and friends of team 'Los Lobos': Marco Galaverni, Rena Schweizer, Péter Marx, Zhenxin Fan and Diego Ortega-Del Vecchyo (yes, you were never considerd a 'Lobo' by outsiders, but in my mind you unquestionably deserve the title of 'Honorary Lobo'). This work is also yours. I will forever remember fondly the intense coding sessions writing those '5 minute scripts', the late night movie sessions (because for what else would you need a gigantic TV screen in a lab?) and the trips to explore California and the US. It was a pleasure and an honor to share this journey with you.

I would also like to thank all other members of the Wayne lab for the welcoming and supporting environment that made work so much easier: Rachel Johnston, Jacqueline Robinson, Sarah Hendricks, Devaughn Fraser, Tiffany Armenta, Amanda Lea, Bridgett vonHoldt, Pauline Charruau, Katherine Pease, Laurel Klein and Wenping Zhang. A special thanks to Shauna Price and Graham Slater for receiving me in LA and having the kindness and patience to help with all the little annoying things that moving to a new city and a new country entails.

To my great friend Sergio Nigenda a very special thanks for letting my jet-lagged self crash on the floor on the first night I arrived in LA, and then again repeatedly over the years. Who would have thought we would end up as housemates a few years later? Your smile and companionship brightened the whole period I stayed in LA. Shine on you crazy diamond!

I will also always remember all other friends that made my time in LA an amazing period of my life, including Adriana Garmendia, Roxana Curiel, Serena Acri, and Saeed Hafeznezami and his parents.

Aos meus amigos e colegas do CIBIO quero agradecer a simpatia, disponibilidade e camaradagem que sempre demonstraram. À Susana Freitas, companheira de contentor, pela amizade, apoio e discussões científicas (e não só). Foste um importante ponto de apoio e âncora de sanidade nos últimos tempos deste doutoramento. Ao Filipe Vieira, pela amizade, perspectiva equilibrada e os sábios conselhos que me transmitiste, mesmo na tua breve passagem pelo CIBIO. Ao João Maia pelo suporte moral e ajuda com as burocracias da faculdade. Ao José Melo Ferreira, João Marques, Fredrik Oxelfelt, John Archer, António Muñoz e Orlando Rodrigues agradeço o apoio informático

e logístico que directamente contribuíram para a realização deste trabalho. Ao Sr. Bernardino pela não menos importante tarefa de tornar o CIBIO mais acessível.

Quero estender também um agradecimento especial aos membros do CTM, no qual continuei a sentir-me em casa, mesmo já não tendo sido o meu local de trabalho. À Susana Lopes, Diana Castro, Sara João, Sofia Mourão e restantes membros do CTM obrigado por me fazerem sentir sempre bem-vindo de volta e pelo encorajamento dado ao longo dos anos. Além disso, muitos dos dados utilizados neste trabalho foram gerados pelo vosso árduo trabalho no laboratório, pelo que sem vocês este trabalho também não teria sido possível.

Muito obrigado a todos os meus amigos que me apoiaram nestes longos e tortuosos anos. Um agradecimento muito especial ao Carlos Vieira pela amizade, pelo incansável encorajamento e pelo exemplo de coragem e tenacidade face às adversidades.

Por último, quero agradecer a todas as pessoas que, mesmo indirectamente, tornaram possível este trabalho, muitas das quais não conheço pessoalmente, nem me conhecem a mim. Quero agradecer, nomeadamente, a todas as pessoas e instituições envolvidas na recolha e disponibilização de amostras de lobo ao longo dos anos e aos técnicos envolvidos na produção dos dados.

Abstract

Gray wolves (*Canis lupus*) are one of the most widespread and charismatic mammalian predators, however relatively little is known about their evolutionary history. Morphologically distinct regional forms, some of them considered subspecies, have been identified worldwide, but the factors that contributed to this differentiation are not always clear. Additionally, several recent studies have challenged our knowledge about the ecology and genetic diversity patterns of wolves and other mobile carnivores: despite the species' remarkable capability for dispersal, unexpectedly high levels of genetic population structure on a regional scale have been found. The wolf is also the only ancestor of the dog, the first animal species to be domesticated and an important companion to humans. Identifying the geographical location and date of dog origins has been subject of much debate due to the great morphological and genetic similarities as well as the complex history of dog translocation and admixture with wild wolves.

Next-generation sequencing technology and recently developed demographic methods make it possible to investigate the evolutionary history of species and populations with recent divergence times and admixture, such as worldwide wolf populations and dogs. In addition, traditional genetic markers, such as microsatellites, are also useful for the exploration of recent population history and genetic structure. In this thesis, the past and present factors determining the current patterns of genetic variability of wolf populations are explored using both genomic and microsatellite data in combination with phylogenomic and population genetic approaches. Specifically, the general objectives of this thesis were to 1) investigate the genetic structure and divergence of European wolf populations, with a special focus on their historical divergence and the genetic population structure within the Iberian wolf population; and 2) to infer the more general demographic history of worldwide wolf populations, and how it relates to dog domestication.

The generation of full genome sequences from individual wolves across the world has allowed an unprecedented view into their evolutionary history. Wolves were found to have experienced a dramatic population reduction ca. 30-50 thousand years ago, implying that current wolves descend from populations that expanded after the end of the Pleistocene. Several worldwide wolf populations were found to have an ancient divergence that predate the current isolation due to human persecution and anthropogenic environmental impacts, and whose genetic and morphological distinctiveness might have been exacerbated by inbreeding and adaptations to local conditions. In particular, wolves from Italy and Iberia were found to share a similar

demographic history and old timing of divergence (ca. 2.4-7.4 thousand years ago) without significant post-divergence gene flow. This result supports their long-term isolation, but also the influence of yet unknown environmental factors that may not be directly related to the end of the Pleistocene. Regarding dog domestication, none of the sampled wolf populations was found to be more related to dogs, supporting the hypothesis that the ancient wolf lineage from which dogs descend is extinct. However, several instances of post-divergence gene flow between dogs and local wolf populations have been found. Wolf-dog divergence dates are estimated between 10-30 thousand years ago, in accordance with an increasing number of other recent genomic and ancient DNA studies.

The use of microsatellite markers from an extensive sample of wolves in the Iberian Peninsula revealed remarkable levels of genetic population structure, in an unexpectedly reticulated pattern for such a relatively small area. The current population dynamic of Iberian wolves appears to resemble a meta-population characterized by relatively high genetic differentiation and low levels of gene flow. The described population structure and dispersal patterns of Iberian wolves might reflect their recent decline and fragmentation, caused by human persecution and habitat changes, as well as adaptive responses to specific ecological conditions. The identified population structure also has consequences regarding demographic inferences using widely-used methods for estimating effective population sizes, population size changes and more complex demographic histories. The effective size of the Iberian wolf population was found to vary depending how population structure is taken into account. Significant signs of bottlenecks were only found in some of the identified subpopulations, which might be a consequence of the low power of these tests in structured populations. Demographic inferences based on a likelihood method suggest that the Iberian wolf population might have suffered a decline starting earlier than expected (ca. 500 years ago), although further investigation is needed to confirm if this value is not inflated due to the violation of model assumptions.

The present work demonstrates the utility of both traditional and emerging molecular genetic variation markers in the inference of evolutionary history. Powerful demographic inference methods allowed the use of such data to uncover the past and present determinants of genetic patterns of diversity of wolves. These inferences are of importance from a historical perspective and to inform effective conservation and management decisions.

keywords: wolf, dog, demography, genomics, population genetics, next generation sequencing (NGS), microsatellites, population structure, gene flow, domestication, bottleneck, *Canis lupus*, Iberian Peninsula

Resumo

O lobo (*Canis lupus*) é um dos mamíferos predadores mais extensamente distribuídos e também mais carismáticos, e no entanto relativamente pouco se sabe da sua história evolutiva. A nível mundial identificam-se várias formas regionais morfologicamente distintas, sendo algumas consideradas subespécies, mas nem sempre é claro que factores contribuíram para esta diferenciação. Adicionalmente, vários estudos recentes abalaram o nosso conhecimento da ecologia e dos padrões de diversidade genética dos lobos e de outros carnívoros com alta mobilidade: não obstante a extraordinária capacidade de dispersão desta espécie, têm sido encontrados níveis altos e inesperados de estruturação genética populacional à escala regional. O lobo é também o único antepassado do cão, a primeira espécie animal a ser domesticada e um importante companheiro para os seres humanos. A identificação do local geográfico e a datação da origem do cão tem sido um tema largamente debatido devido às dificuldades de interpretação resultantes das grandes semelhanças morfológicas e genéticas entre cães e lobos, bem como à complexa história de translocação do cão e a eventos posteriores de miscigenação entre as duas formas.

As tecnologias de sequenciação de alto rendimento ('next-generation sequencing') e métodos demográficos desenvolvidos recentemente tornam possível a investigação da história evolutiva de espécies e populações com divergências recentes e miscigenação, tais como as populações mundiais de lobo e cães. Além disso, marcadores genéticos tradicionais, tais como microssatélites, são também úteis para a exploração da história populacional recente e estruturação genética. Nesta tese foram explorados os factores presentes e passados que determinam os padrões contemporâneos de variabilidade genética das populações de lobo, através do uso de dados genómicos e de microssatélites em conjunto com abordagens filogenómicas e de genética populacional. Especificamente, os objectivos gerais foram: 1) investigar a estrutura genética e a divergência das populações europeias de lobo, com especial ênfase na sua divergência histórica e na estrutura genética populacional da população ibérica; e 2) inferir a história demográfica geral das populações mundiais de lobo, e como esta se relaciona com a domesticação do cão.

A sequenciação de genomas completos de lobos de todo o mundo permitiu uma visão inédita da sua história evolutiva. Verificou-se que os lobos sofreram uma dramática redução populacional ca. 30-50 mil anos atrás, significando isto que os lobos actuais descendem de populações que expandiram após o fim do Pleistoceno. Várias populações mundiais aparentam uma divergência antiga que precede o seu isolamento

actual devido à perseguição humana e impactos ambientais antropogénicos, e cuja diferenciação genética e morfológica possa ter sido exacerbada pela consanguinidade e adaptações a condições locais. Em particular, os lobos italianos e ibéricos aparentam ter histórias demográficas e tempos de divergência (ca. 2.4-7.4 mil anos atrás) similares, e uma ausência de fluxo génico após a separação. Este resultado apoia a hipótese do seu isolamento antigo mas também a influência de factores ambientais ainda desconhecidos que poderão não estar directamente relacionados com o final do Pleistoceno. Em relação à domesticação do cão, nenhuma das populações lupinas amostradas é mais próxima dos cães, apoiando a hipótese de que a linhagem ancestral que originou o cão se encontra extinta. No entanto, foram encontrados vários casos de fluxo génico entre cães e populações locais de lobos após a sua divergência. A divergência entre lobo e cão é estimada em cerca de 10 a 30 mil anos atrás, de acordo com um crescente número de outros estudos genómicos de DNA antigo publicados recentemente.

O uso de microssatélites de uma extensa amostragem de lobos da Península Ibérica revelou a existência de níveis extraordinários de estruturação genética populacional, sob a forma de um padrão inesperadamente reticulado para uma área relativamente pequena. A dinâmica populacional da população ibérica actual parece-se com uma meta-população caracterizada por diferenciação relativamente elevada e baixos níveis de fluxo génico. A estruturação descrita e os padrões de dispersão dos lobos ibéricos poderão reflectir a sua história recente de declínio e fragmentação, causada por perseguição humana e alterações no habitat, bem como de respostas adaptativas a condições ecológicas específicas. A estruturação populacional encontrada tem também consequências para as inferências demográficas a partir de métodos largamente utilizados para a estimativa de tamanhos efectivos populacionais, alterações de tamanho populacional e de histórias demográficas mais complexas. Verificou-se que o tamanho efectivo populacional da população ibérica de lobo varia consoante a estruturação populacional é tida ou não em conta. Sinais significativos de reduções populacionais (*bottlenecks*) foram encontrados apenas em algumas das subpopulações descritas, o que poderá ser uma consequência do baixo poder destes testes em populações estruturadas. As inferências demográficas baseadas num método de verosimilhança (*likelihood*) sugerem que a população ibérica de lobo sofreu um declínio que começou mais cedo do que o previsto (ca. de 500 anos atrás), embora seja necessário efectuar trabalho suplementar para rejeitar um possível inflacionamento deste valor devido a desvios às assumpções do modelo.

O presente trabalho demonstra a utilidade de marcadores moleculares de variação genética tradicionais e emergentes na inferência de histórias evolutivas. A existência de poderosos métodos de análise demográfica permitiram o uso desse tipo de dados para desvendar factores passados e presentes que determinam os padrões de variabilidade genética dos lobos. Estas inferências têm importância do ponto de vista histórico e na implementação de medidas efectivas de gestão e conservação desta espécie.

palavras-chave: lobo, cão, demografia, genómica, genética populacional, sequenciação de alto rendimento, microssatélites, estruturação populacional, fluxo génico, domesticação, reduções populacionais, *Canis lupus*, Península Ibérica

Table of Contents

List of Tables

List of Figures

Abbreviations

bp - base pairs
CI - confidence interval
CNE - conserved non-coding element
CpG - citosine-phosphate-guanine
CPS - cryptic population structure
DNA - deoxyribonucleic acid
G-PhoCS - Generalized Phylogenetic Coalescent Sampler
GPS - Global Positioning System
GWAS - genome-wide association study
HWE - Hardy-Weinberg equilibrium
indel - insertion/deletion
kya - thousand years ago
LGM - last glacial maximum
MCP - minimum convex polygon
ML - maximum likelihood
mtDNA - mitochondrial DNA
Mya - million years ago
N_e - effective population size
NGS - Next-Generation Sequencing
NJ - neighbor joining
PCA - principal components analysis
PCR - polymerase chain reaction
PE - paired end
PSMC - pairwise sequentially Markovian coalescent
qPCT- quantitative PCR
RNA - ribonucleic acid
SNP - single nucleotide polymorphism
SNV - single nucleotide variant
VCF - variant call format
ya - years ago

Chapter 1 - GENERAL INTRODUCTION

1.1 - The Gray Wolf (*Canis lupus*)

1.1.1 - Taxonomy and evolution

The gray wolf (*Canis lupus* Linnaeus 1785) is the largest of the wild canids (family *Canidea*, of the order Carnivora), a lineage that also includes coyotes, jackals, foxes and other dog-like mammals (Wilson and Reeder 2005). Canids originated in the late Miocene (ca. 6 Mya) in North America and reached Asia across Beringia during a global warm period in the early Pliocene, ca. 5-4 Mya (Wang and Tedford 2008). Several species of the *Canis* genus appeared in North America and Eurasia during and after this period. Morphological (Nowak 1995) and genetic (Wayne et al. 1997) studies support that wolves and coyotes descend from the same lineage, represented in the fossil record by the ancestral *Canis lepophagus* (or *C. arnensis*). This species originated at the end of the Pliocene (ca. 2.6 Mya) in North America and then expanded into Eurasia. The coyote and wolf lineages appear to have been separated by 1.5 Mya (Nowak 2003; Wang and Tedford 2008). *C. lepophagus* was closer in size to extant coyotes, but appears to have originated several larger, wolf-sized species on both continents: *C. armbrusteri* and *C. dirus* in North and South America, *C. gezi* and *C. nehringi* in South America, and *C. etruscus* in Eurasia. The rapid diversification of these and other *Canis* species near the Pliocene-Pleistocene boundary (ca. 1.8 Mya) is known as the 'Wolf Event' (Wang and Tedford 2008). This period is also associated with the origin of the mammoth steppe biome following intense glaciations, and it is possible that early humans (*Homo erectus* and *H. sapiens*) competed with wolves for similar types of prey (Wang and Tedford 2008).

C. lupus seems to have originated from *C. etruscus* in the early to middle Pleistocene through an intermediate stage denominated *C. mosbachensis*, probably in the arctic regions of Eurasia. The fossil record suggests gray wolves reached Europe ca. 800 kya. They also expanded into North America at least 500 kya, but were restricted above the Arctic Circle for a considerable time before reaching the midcontinent in the last glacial cycle (ca. 100 kya) (Wang and Tedford 2008). Recent genetic studies identify the Old / New World split as the most salient genetic division of contemporary wolf populations (VonHoldt et al. 2010; Larson et al. 2012).

Wolves therefore evolved in the Arctic environment of northern Eurasia, Beringia and northernmost North America (Wang and Tedford 2008). Possibly as a consequence

of this origin, wolves appear to have had an abundant and widespread distribution in Europe during the Pleistocene glacial periods (Sommer and Benecke 2005), contrary to many other animal and plant species that experienced drastic range contractions (Taberlet et al. 1998; Hewitt 2000). The study of extant mtDNA variability in wolves has found relatively recent coalescent times and an absence of any large-scale geographical structure (Vilà et al. 1999), which has been hypothesized to be the result of their remarkable adaptability and dispersal capabilities.

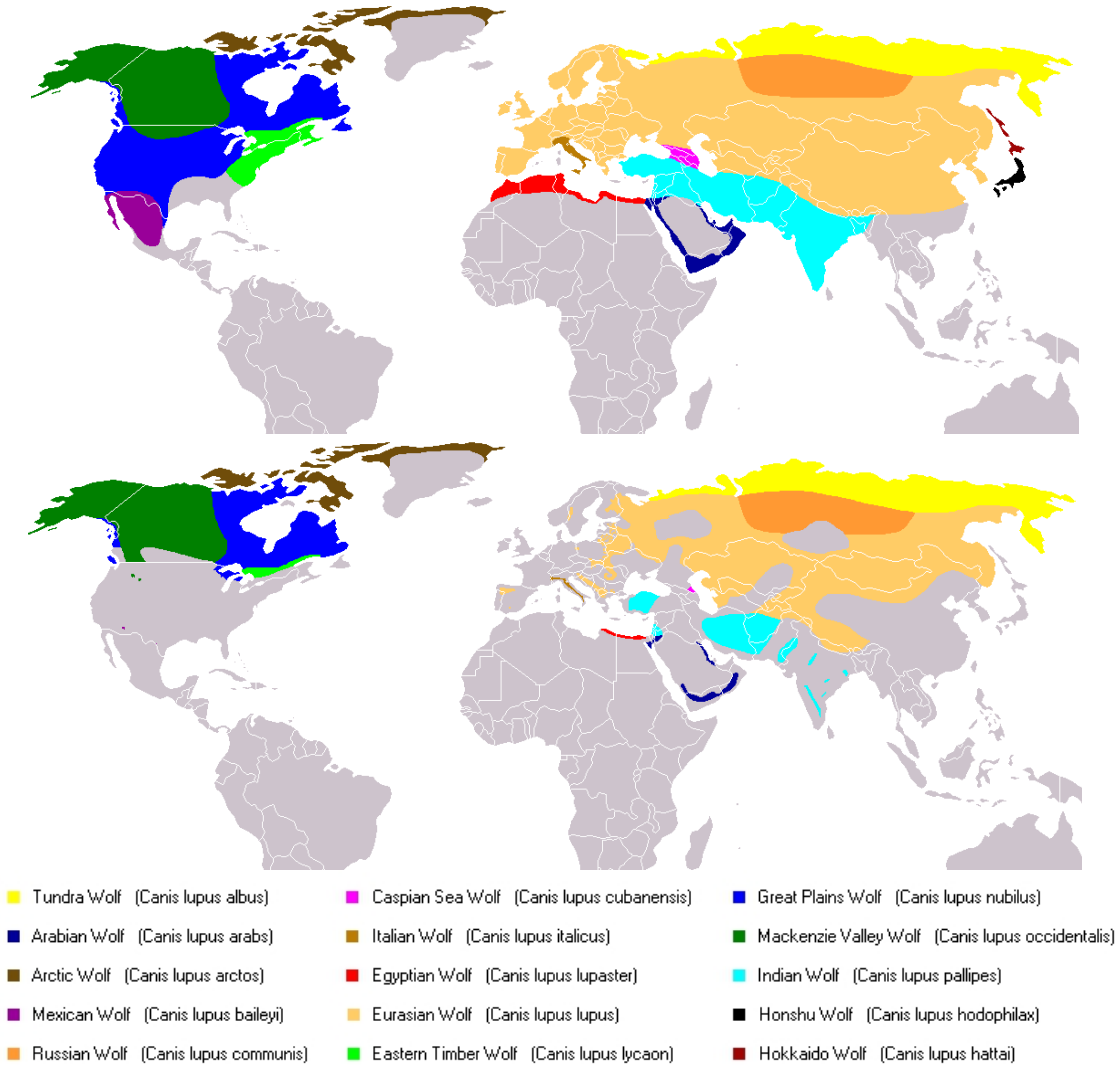


Fig. 1-1: Original (above) and present (below) distributions of grey wolf subspecies, according to Nowak, 2003. Image: Wikimedia Commons

Extant wolves possess a considerable variability of body sizes, weight and coat colors. Adult male wolves weigh from 20 to 80 kg, while females are usually smaller (15-55 kg). The total body length is 110-148 cm, with the tail usually representing up to a third of the total length, and height at the shoulders averages 50-70 cm (Peterson and Ciucci 2003). Larger animals are usually found in northern latitudes, while Mediterranean

wolves for instance are much smaller, rarely exceeding 35 kg (Boitani 2000). Wolf coat color is very variable, ranging from white (in arctic regions) to brown, reddish or gray. Additionally, there is substantial individual variation affected by age, sex, season of the year or health condition (Mech and Boitani 2003).

Given the wide distribution of *C. lupus*, and its remarkably high phenotypic variation regarding size, weight and coat color, several extant morphologically distinct groups have been identified. These groups have traditionally been designated with subspecific names, but given the lack of formal taxonomic rules, many disparities exist between authors (Nowak 2003; Sillero-Zubiri et al. 2004; Wilson and Reeder 2005). A statistical analysis of the skull morphology of worldwide wolf populations led to the proposal of extant eleven subspecies (Fig. 1-1) (Nowak 1995; Nowak and Federoff 2002; Nowak 2003). In North America, five subspecies were recognized: *C. l. arctos* (arctic wolf), *C. l. occidentalis* (northwestern wolf), *C. l. nubilus* (plains wolf), *C. l. baileyi* (Mexican wolf), and *C. l. lycaon* (eastern or Great Lakes wolf). The taxonomic status of the eastern wolf, sometimes considered a separate species (*C. lycaon*), as well as that of a possibly related species, the red wolf (*C. rufus*), is however disputed given their possible origin by hybridization with coyotes (Wayne and Jenks 1991; Wilson et al. 2000; Nowak 2003; Chambers et al. 2012; Rutledge et al. 2015). In Eurasia, the proposed wolf subspecies include *C. l. albus* (an Eurasian arctic wolf), *C. l. communis*, *C. l. lupus* (the moderate sized form distributed over most of Asia and Europe), *C. l. cubanensis* (that occurs only in the Caucasus), *C. l. pallipes* (a form adapted to desert conditions and distributed over most of southwestern Asia, from the Indian subcontinent to the Arabian Peninsula), and *C. l. italicus* (Italian wolf).

In 1907, Spanish zoologist Ángel Cabrera proposed that wolves of the Iberian Peninsula represented a distinct subspecies (*C. l. signatus* Cabrera 1907). It is characterized by distinctive dark marks on its front legs and on the tail (which are responsible for its designation: 'signatus' means 'marked' or 'signed') while the remaining fur is generally brown and grey (Fig. 1-2). However this classification has not gained wide acceptance (e.g. Nowak 1995; Nowak 2003; Sillero-Zubiri et al. 2004; Wilson and Reeder 2005), mainly due to the lack of supporting studies (but see Petrucci-Fonseca 1990; Vilà 1993, which support the divergence based on dental morphology and cranial measurements). Genetic studies, using microsatellite markers and SNPs have also demonstrated that the genetic differentiation between Iberian wolves and their closest population, Italian wolves, are of a similar magnitude as the differentiation between Italian wolves and remaining eastern European populations (Lucchini et al. 2004; Pilot et al. 2014). Italian wolves had also been proposed to constitute a different subspecies

based on morphological (Nowak and Federoff 2002) and genetic data (Lucchini et al. 2004).

Fig. 1-2: Iberian wolf with its distinctive dark markings on the front legs and tail. Image: R Godinho

1.1.2 - Ecology and social structure

Wolves are social animals that typically live in territorial groups termed packs, which usually form around a mated dominant pair ('alpha pair'). Among pack members, a linear hierarchy exists, maintained through ritualized aggressive behavior, in which dominant individuals take most of the initiative and have most of the privileges in feeding and reproducing (Mech and Boitani 2003). In ideal conditions, the breeding pair produces offspring every year, which can remain with the pack for more than 4 years. Wolf packs are therefore usually comprised of related individuals, born during several years (Mech 1999), although packs with a different structure (e.g. including unrelated adoptees) have been recorded (Mech and Boitani 2003). The size and composition of wolf packs is variable, and can be affected by several variables, such as food availability, size and type of prey, or intensity of human disturbance (Fuller et al. 2003). Packs are typically constituted by 3-11 individuals (Boitani 2000; Fuller et al. 2003).

Wolves reach sexual maturity at two years old, but can defer their reproduction if they remain in their natal pack (Mech and Boitani 2003). Due to the tight social and territorial bonds of wolf packs, individuals must usually disperse from their natal pack, find a mate, and establish a territory with adequate resources. Consequently, individual

breeding strategies can involve either close- or long-distance dispersal. In the first case, wolves can wait for a breeding position to open, either in their natal or a neighboring pack, become an extra breeder in the pack, carve out a new territory, or usurp an active breeder (Mech and Boitani 2003). In large packs, 'splitting' has been observed, in which a group of wolves splits off and assumes a new territory. Mating pairs that form within the natal territories of one of the elements can also attempt to establish a new pack in a territory adjacent or partially overlapping with the natal pack (a process known as 'budding') (Mech and Boitani 2003). On the other hand, long-distance dispersal can take wolves into new populations or to the edge of the species' local distribution. Movement usually happens in a more or less single direction during the course of several weeks. Individuals of both sexes have been observed to travel distances over 1000 km, and circumvent large topographical accidents or human-made structures (Merrill and Mech 2000; Wabakken et al. 2001; Ciucci et al. 2009; Andersen et al. 2015; Ražen et al. 2016). The tendency to disperse larger distances seems to be more common in situations of intense resource competition and instability. In fluctuating populations, a greater proportion of mature individuals dispersed during population declines or increases than during stable periods (Fuller et al. 2003). Long-distance dispersal also occurs in very low-density populations, where the chance of finding a mate is lower (Wabakken et al. 2001; Ražen et al. 2016).

Wolves are highly territorial, and their territories are often very large (tens to thousands of km^2) (Mech and Boitani 2003). Individuals travel regularly to hunt and maintain the boundaries of their territory, which are advertised through howling and markings with urine and faeces. Boundaries are rarely trespassed because it may lead to violent or even fatal aggressions. Within territories, different zones can be recognized, including activity centers, that are usually locations associated with reproduction (Fuller et al. 2003; Theuerkauf et al. 2003), and buffer zones on the territory peripheries, that are less frequented and therefore minimize inter-pack aggressions (Mech 1977; Mech and Harper 2002). In a well-established population, a mosaic of territories develops, wherein packs compete for space and resources. Territory sizes are very variable, from <100 km^2 to $>$ 5000 km^2 (Fuller et al. 2003; Mech and Boitani 2003), and depend on several factors, mainly on the size and distribution of prey, but also on the size of the pack or specific characteristics of the habitat (Fuller et al. 2003; Jędrzejewski et al. 2007).

Wolves are flexible and opportunistic carnivores, and their diet varies a lot depending on locally available food sources (Mech and Boitani 2003; Imbert et al. 2016; Newsome et al. 2016). They usually prey on large ungulates, but are capable of hunting prey of different sizes. In Europe, much of the natural wolf habitat has been altered and

fragmented by human activities and, as such, domestic animals can represent a substantial part of the wolf diet in these regions. In southern Europe, such as the Iberian Peninsula and Italy, wolves commonly prey on livestock, such as goats, sheep and horses (Peterson and Ciucci 2003; Álvares 2011; López-Bao et al. 2013), which often results in conflict with local human populations (Boitani 2000; Chapron et al. 2014; López-Bao et al. 2015). Where both wild and domestic prey are found, wolf diet seems to depend on their relative availability and on the influence of climatic or demographic factors. In general, predation on livestock increases during the grazing period, while wild prey is more important during the rest of the year (Peterson and Ciucci 2003). Like other big carnivores, wolves are 'keystone species' in many ecosystems, exerting impacts on inferior trophic levels (Estes et al. 2011). For instance, they regulate herbivore populations that constitute its main prey and influence the distribution of mesopredators and herbivores, which in turn can have great impacts on plant communities (Estes et al. 2011; Ripple and Beschta 2012).

1.1.3 - Current distribution, population trends and conservation issues

Wolves are highly adaptable and occupy a wide range of habitats, from warm deserts to cold tundra. The original distribution of the gray wolf included most of the Northern Hemisphere above 15ºN in North America and 12ºN in India and the Arabian Peninsula (Fig. 1-1) (Sillero-Zubiri et al. 2004). Today, wolf populations still occupy most of their former range in Asia, Alaska and Canada, but have been greatly reduced and fragmented everywhere else. They have been extirpated from Japan, Mexico, most of the USA and Central and Western Europe (Ripple et al. 2014). This severe decline is the result of habitat modifications and reduction of natural prey as human settlements developed over the last centuries, and also of direct persecution, often sponsored by officially sanctioned eradication plans in the 19th and 20th centuries (Mech and Boitani 2003). In Europe, significantly sized wolf populations remain in the Iberian Peninsula, Italy, the Balkans and Scandinavia. More numerous populations exist in Eastern Europe, which are in contact with Asian populations, where wolf abundance is also higher (Mech and Boitani 2003; Chapron et al. 2014).

Changes in public opinion and the implementation of protection measures have led to the recovery and expansion of several of the wolf populations in Western Europe and North America in recent decades. Wolves were reintroduced in the Western USA, and in Europe small populations have established themselves in areas where wolves have been extinct for more than 50 years, such as Switzerland, France, Germany,

Norway and Sweden (Boitani 2000; Mech and Boitani 2003). This recovery parallels the trend of other large carnivores, whose populations have been recovering and expanding, increasingly spreading back also into more human-dominated landscapes (Chapron et al. 2014). While usually associated with the wilderness and remote areas, wolves also display a great capacity to survive in regions of high human density (Fritts et al. 2003; Llaneza et al. 2012). It is estimated that the total worldwide wolf population is over 300,000, with the largest populations in Canada, Russia, Kazakhstan, Mongolia, China and Alaska (Mech and Boitani 2003; Sillero-Zubiri et al. 2004).

For conservation purposes, 10 wolf populations have been delineated in Europe, which are considered as management units independently of political and administrative borders (Boitani and Ciucci 2009; Kaczensky et al. 2012). Population trends in these units are assumed to only result from reproduction and mortality and not migration. From West to East, these comprise the Iberian Peninsula, with two distinct populations (Northwestern Iberia and Sierra Morena), the Western Alps, the Italian Peninsula, the Carpathians, the Balkans (Dinaric-Balkan population), the Baltic, Finland and adjacent Russian regions (Karelian population), Scandinavia, and Germany and Western Poland (Central European population) (fig 1-3).

In areas without significant human influence, natural mortality be caused by intra-specific aggression, illness, wounds from hunting accidents, and starvation and malnutrition (Boitani 2000). For wolves living in more densely populated areas however, human actions are usually the main cause of mortality. These usually include direct killing, often as an accidental result of hunting or poaching of other species, either through poisoning, shooting or trapping; and disturbance of the habitat, namely of reproduction areas or elimination of food sources (Boitani 2000; Chapron et al. 2014). Large carnivores such as the wolf are particularly susceptible to these threats due to intrinsic biological characteristics such as relatively long gestation periods, low natural population densities and high trophic level (Cardillo et al. 2004).

1.1.4 - Domestication

Pleistocene wolves were the first animal species to be domesticated by humans, in a process whose timing and location is still subject to a lot of debate (Larson et al. 2012). Genetic studies have confirmed that gray wolves are the only ancestors of domestic dogs (Vilà et al. 1997; Lindblad-Toh et al. 2005), since it was sometimes proposed that other canids, namely coyotes and jackals, could also have contributed to the domestic dog's ancestry (Wayne and Ostrander 2005).

Fig. 1-3: Wolf distribution in Europe. Dark and light gray areas represent permanently and sporadically populated areas, respectively. Red boundaries represent countries for which information is available. Population units referenced in the main text are in blue. Image: Large Carnivore Initiative for Europe.

Archaeological evidence pertaining to the origin of dogs has mostly been contentious due to difficulties in discriminating between ancient remains of small wolves and early domesticated dogs. Morphological differentiating characters in these specimens may not have been fixed during the early stages of domestication, or may be indistinguishable from local, now extinct wolf populations (Larson et al. 2012). Additionally, the widespread distribution of wolves and of several other similar canids greatly hinder any inferences from the geographic distribution of fossil forms. Specimens claimed to be ancient dogs found in Europe and Asia date as far back as 36 kya (Germonpré et al. 2009; Ovodov et al. 2011; Germonpré et al. 2012; Druzhkova et al. 2013).

Attempts to identify the location and date of dog origins have mainly come from genetic studies. Initial estimated dates, based on mtDNA divergence of modern dogs and wolves, suggested dogs originated more than 100 kya (Vilà et al. 1997). These estimates have however been found to be severely overestimated by subsequent studies. Based on regional patterns of modern dog diversity, dog domestication was

proposed to have happened in East Asia ca. 16 kya (Savolainen et al. 2002; Pang et al. 2009). The methodology of these studies has, however, been questioned when it was demonstrated that similar genetic diversity patterns could be observed in regions that have been excluded as origins of domestication, probably due to exceptional levels of translocation and admixture (Boyko et al. 2009). Studies of SNP and nuclear DNA variation concluded that both East Asian and Near Eastern wolf populations contributed significantly to the modern dog gene pool, arguing for a Middle Eastern origin of dogs (Gray et al. 2010; VonHoldt et al. 2010). Analyses of mitochondrial genomes of prehistoric canids from Eurasia and the New World, and of modern dogs and wolves, have shown that all modern dogs are phylogenetically most closely related to either ancient or modern canids of Europe, and suggesting a domestication date of 19-32 kya (Thalmann et al. 2013). The sequencing of an ancient Siberian wolf genome suggested a domestication around 27 kya, and that ancient wolves of this region contributed to the ancestry of high-latitude dog breeds (Skoglund et al. 2015) A recent genomic study suggests an origin in southern East Asia, ca. 33 kya based on the higher genetic diversity of dogs in that region and their closeness to wolves (Wang et al. 2016).

Early stages of domestication were probably not deliberate nor directed, and possibly resulted from the opportunistic scavenging behavior of wolves attracted to human camps. Domestication traits therefore probably evolved by natural, rather than artificial selection (Larson et al. 2012). Following domestication, dogs rapidly became ubiquitous companions of human populations, being used as sentinels, hunting companions, pets, in transport and herding, and as a food source (Wayne and Ostrander 2005). From Eurasia, dogs accompanied humans into the New World ca 12-14 kya (Leonard et al. 2002), and through the Pacific islands to Australia, where they originated the wild dingo (Savolainen et al. 2004).

A variety of dog morphologies have existed for several thousand years but most of modern dog breeds have existed only since the 19th century (Parker et al. 2004). These groups are governed by strict breeding rules, forming closed gene pools to maintain a set of characteristic traits. Genetic studies of modern dog breeds found evidence for strong genetic isolation between dogs of different breeds and a high genetic homogeneity within individual breeds (Parker et al. 2004). These studies have also identified a subset of breeds, mainly of Asian and Nordic origin, that are consistently placed in basal phylogenetic positions, and have been designated as 'ancient' or 'basal' (Fig. 1-4) (Parker et al. 2004; VonHoldt et al. 2010; Larson et al. 2012). Dogs from these breeds show the closest relationship with wolves, and have been proposed to represent the best surviving representatives of the ancestral dog gene pool (VonHoldt et al. 2010),

although their basal position could also be explained by their long-term isolation from other breeds (Larson et al. 2012). Today, more than 350 dog breeds exist, that encompass a great morphological variety: dogs can differ more than 40x in size (more than any other mammal), have diverse behavioral dispositions, and exhibit a great range of body shapes (skull outline, body constitution, tail conformation, etc.) and coat characteristics (color, length, texture and curl) that are generally not present in wild wolves (Wayne and Ostrander 2005). Most dogs today, however, still live as semi-feral human commensals known as 'village dogs', descending from their own ancient village-dog populations and are not significantly admixed with modern breeds (Boyko et al. 2009).

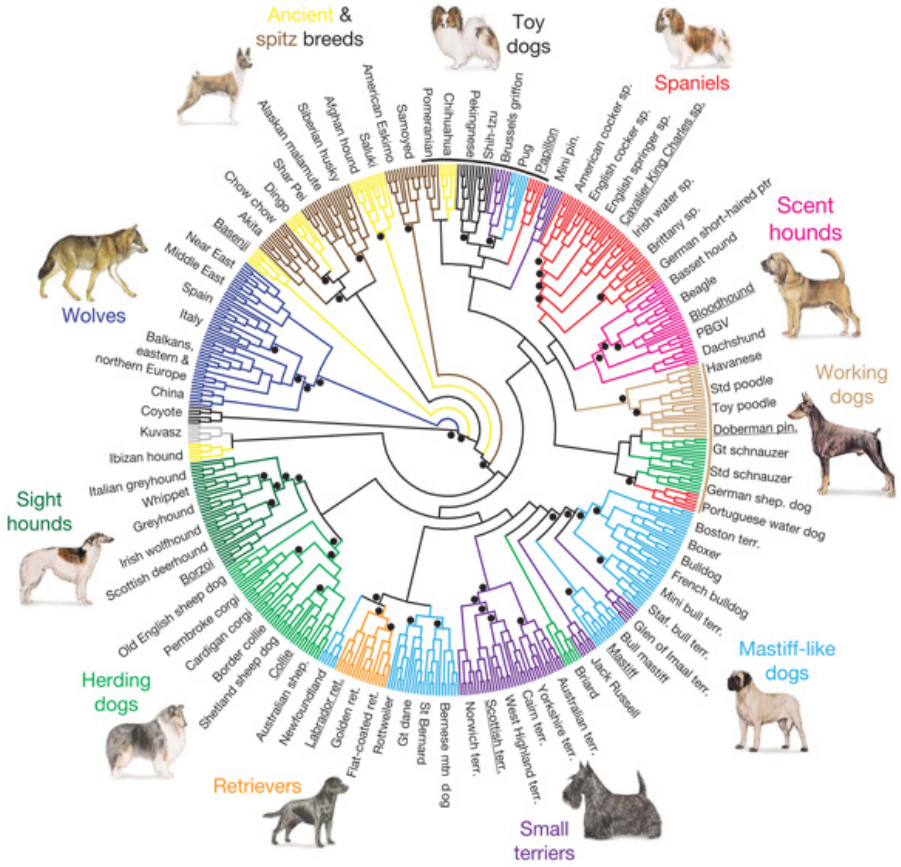


Fig. 1-4: Neighbor-joining tree of wolves and dogs of different breeds based on haplotype-sharing, using 48k SNPs, by vonHoldt et al. 2010. Branch color indicates the phenotypic/functional designation used by dog breeders. Image: vonHoldt et al. 2010

Dogs, wolves and coyotes can interbreed and produce fertile offspring, given their recent evolutionary divergence and lack of effective reproductive isolation. This can be a conservation problem in areas where they coexist (Vilà and Wayne 1999). Interbreeding between domestic animals and their wild counterparts, with or without significant introgression, is generally considered undesirable, mainly in endangered populations, as it promotes genetic homogenization and possibly leads to the disintegration of traits that have arisen as local adaptations (Allendorf et al. 2001; Randi 2008). Hybridization between gray wolves and dogs has been described worldwide. It is thought that hybridization is more frequent near human settlements in disrupted wolf habitats, where feral and domestic dogs are common; therefore wolf populations of Mediterranean countries have generally been considered to be at higher risk because of their small size and extensive contact with dogs (Vilà and Wayne 1999). Indeed, hybrid individuals were identified in Portugal and Spain (Godinho et al. 2011, 2015), Italy (e.g. Randi and Lucchini 2002; Caniglia et al. 2013) and Israel (Vilà and Wayne 1999) but also in Sweden (Vilà et al. 2003), Estonia and Latvia (Andersone 2002; Hindrikson et al. 2013) and several eastern European countries (Vilà et al. 1997; Vilà and Wayne 1999; Randi et al. 2000). Population genetic studies have, however, shown that these events are relatively rare and there is no evidence for significant introgression of dog genes into wild wolf populations (Vilà and Wayne 1999; Randi 2008; Randi 2011).

1.2 - Inference of Population History from Genetic Data

1.2.1 - Recent advances in population inferences from molecular data

Since molecular genetic variation started being surveyed directly in biological populations, there has been a dramatic progress in fields of study trying to understand the evolutionary forces that shape the patterns of genetic diversity within and between species. This has been a result of both technical advances that make obtaining genetic data easier and cheaper, and of the development of more sophisticated and realistic models of evolution.

While the direct analysis of DNA sequence variants has been possible for many years, other genetic markers, such as electrophoretic variants, restriction fragment length polymorphisms or microsatellites have been mostly preferred, and technological limitations prevented the genotyping of a large number of these markers on a truly genomic scale (Luikart et al. 2003). In recent years, the development of fast and massive sequencing and genotyping technologies ('next-generation sequencing' - NGS) has made the collection of large datasets of genetic variation a reality in several organisms, and led to a proliferation of comparative genetic variation databases (Metzker 2009; Davey et al. 2011). While many of these genomic methods have been pioneered in humans, due to easier access to samples and an increased interest of the scientific community, they are increasingly being applied to other species as well (Ekblom and Galindo 2010; Davey et al. 2011). Parallel to the increasing data generation, there has also been a rapid growth in computational storage and processing power, which allows for both easier storage and access to these types of data, as well as the implementation of more complex, and therefore computationally demanding, statistical analyses. As more and more genome-scale data sets are being generated at an increasing pace and lower cost, challenges become more related to the interpretation of the data than to its acquisition (Schraiber and Akey 2015).

A big focus in the fields of molecular phylogenetics, phylogeography, population genetics and others has always been the development of models that describe the effects of evolutionary processes on the genetic variation of organisms. In this sense, a model is a simplified mathematical formulation of the biological processes that produce the data, incorporating parameters of interest such as mutation and recombination rates, population sizes, divergence times, etc. and that predict how forces such as genetic drift, selection or migration affect patterns of genetic variability. These models can be used to

understand the past, and otherwise unknown, evolutionary history of extant populations because past events leave specific genetic imprints that can be detected today (Schraiber and Akey 2015). They can also be applied to a variety of questions, from the divergence of genetic lineages, including speciation, the environmental or historical determinants of their geographical distribution, or the demographic or ecological constraints that gave rise to the observed patterns of population genetic variability. The fields mentioned above traditionally differ on which type of events they emphasize, but have been brought closer together in recent years (Cutter 2013). This integration has been facilitated by the common reliance of these disciplines on molecular sequence variation data, which means that they all can leverage the large multilocus datasets resulting from NGS that are now available (McCormack et al. 2013). Theoretical advances in the field of coalescent theory further facilitated this integration (Rosenberg and Nordborg 2002). This has led to a better understanding on how diverse processes such as demography, selection and speciation interact in shaping genomes, improving our understanding of biodiversity from micro- to macro-evolutionary scales (Edwards 2009; Cutter 2013).

1.2.2 - Inferences of demographic history

The inference of demographic history from genetic data is based on the fact that evolutionary forces such as genetic drift, migration, mutation and natural selection change the frequencies of alleles in a population trough time. While natural selection can have profound impacts, leading to the increase or decrease of the frequency of certain alleles depending on their contribution to reproductive success, many of the genetic variation patterns seen today are the result of so-called 'neutral evolution' resulting from demographic events such as population size changes or fragmentation (Emerson 2001). The size of a population is an important parameter because it determines the strength of genetic drift, i.e. the amount of genetic variability lost between generations due to chance. In population genetics terms, the expression 'effective population size' (N_e) has been used to signify the size of a theoretical population with the same rate of genetic drift as the population being studied (Wright 1931). Changes in effective population size, namely severe reductions, can lead to the fixation or loss of alleles due to genetic drift. In populations with large effective sizes, the effect of genetic drift can be negligible; in the absence of migration or selection, mutation-drift equilibrium can be sustained, where the loss of diversity through genetic drift is compensated by the introduction of diversity

through mutation. Migration between populations, by the dispersal of the organisms themselves or their gametes, leads to gene flow, i.e. the movement of alleles between populations. The differentiation between populations will then be a consequence of the levels of gene flow between them.

Many demographic inference methods approach the task of inferring demographic history by comparing a statistic that summarizes an important aspect of the data, calculated from the sampled population, to their expected distribution assuming neutrality. For example, early studies focused on how changes in population size affect the distribution of pairwise differences between individual DNA sequences and the number of segregating sites within a population (Tajima 1989; Rogers and Harpending 1992). Strong population reductions can also originate distinctive signatures of expected heterozygosity and, in the case of microsatellites, allele size distribution (Cornuet and Luikart 1996; Luikart and Cornuet 1998; Garza and Williamson 2001). The expected distribution of these quantities at mutation-drift equilibrium can be derived mathematically using appropriate models, and comparison of the actual values calculated from the sample can be used to make inferences about population size history. Methods based on summary statistics do not make use of the full information in the data however, and often have limited statistical power compared to more complex methods (Marjoram and Tavaré 2006).

More advanced and powerful methods to characterize demographic histories and infer related parameters have been developed by employing mathematical approaches such as maximum likelihood, and Bayesian and Approximate Bayesian Computations (Kuhner 2009; Bertorelle et al. 2010). These methods can be very detailed in the histories they attempt to infer. Many of these approaches assume a basic model of sequential demographic events (such as population splits, growth or decline), associated with a set of parameters such as effective population sizes, divergence times, migration rates, etc. which they attempt to estimate (Fig. 1-5). In likelihood-based analyses the mathematical probability of obtaining the data (the observed genetic variants and their frequencies) is calculated conditional on the different parameters of interest (migration rates, effective population sizes, population growth rates, etc.) using a stochastic evolutionary model such as coalescent theory; estimates of demographic parameters can then be obtained by maximum likelihood, i.e. by inferring the parameter values that maximize the likelihood of the data (Marjoram and Tavaré 2006). This calculation is computationally very demanding, which represents one of the most important limitations of these methods (Marjoram and Tavaré 2006; Schraiber and Akey 2015). Many methods use a Bayesian approach for parameter inference, wherein prior knowledge of parameters of

interest is taken into account. This is expressed as a prior distribution (parameter values that are considered likely before the data is examined), which is modified by the observed data to produce a posterior distribution (proportional to the likelihood and the prior distribution) that summarizes the updated knowledge about parameters conditional on the observed data (Beaumont and Rannala 2004). In an approximate Bayesian framework, calculation of the likelihood is further approximated by performing simulations and calculations on a set of summary statistics instead of the full data itself (Lopes and Beaumont 2010; Bertorelle et al. 2010).

Coalescent theory (Kingman 1982) underlies many of these more complex methods. This theory provides a theoretical framework for describing the genealogical relationships between a sample of chromosomal segments taken from a population, such as large multi-locus datasets generated by NGS. Methods based on the coalescent can consequently take into account the information contained within the relative positions of the internal nodes of the genealogies of those chromosomal segments that single-statistic methods cannot. These ancestral genealogies take the form of trees whose topology will depend on different demographic factors. Coalescent theory underlies genetic methods in many areas because it is applicable across different timescales, and can be used as a mathematical modelling tool to derive estimators of population parameters and to devise statistical tests of models of evolution (Rosenberg and Nordborg 2002). It is also used as a simulation tool, since it allows to simulate samples from a variety of models in a more easy and efficient manner compared to classical population-genetics simulations (Rosenberg and Nordborg 2002).

Estimates of demographic history are fulcral for addressing a variety of questions. From a historical point of view, they offer a window into the past of a species or population, helping to explain their current distribution, genetic structure or diversity. They help to understand the impact of past environmental changes, from which inferences for the future might also be drawn (Knowles 2009; Hickerson et al. 2010). From a conservation perspective, past bottlenecks and small founder populations can lead to higher frequency of deleterious mutations, which has to be taken into account in management and conservation strategies (Frankham 2005). Genome-wide association studies (GWAS) for inferring the genetic basis of complex heritable traits and diseases can produce false positive associations if the demography of the population is not taken into account (Lohmueller 2014). Understanding of historical population demography is also important for calibrating correct null models of neutral genome evolution for the detection of regions subject to evolutionary pressure from natural selection (Nielsen 2005; Campbell and Tishkoff 2008; Lohmueller 2014).

Even with substantial progress having been made, there are still difficulties in demographic inference. Genetic data is inherently limited because of the stochasticity of coalescent events, which can never be directly observed (Rosenberg and Nordborg 2002). To extract meaningful results from demographic inference methods it is almost always necessary to rely on external calibrations, such as mutation rates or absolute divergence times estimated from the fossil record, which can add further uncertainty (Schraiber and Akey 2015). Genetic demographic inferences can also be confounded by other evolutionary phenomena, such as selection or unaccounted population structure. Methods to account for the joint effects of selection and demography are not yet very robust (Schraiber and Akey 2015), and therefore the common approach followed by many studies has been to restrict analyses to genomic regions least likely to be influenced by selection, such as non-coding regions (Gronau et al. 2011; Cutter 2013). Regarding population structure, many parametric models of demographic history assume single, randomly mating populations, which is not always the case of natural populations, and often produce erroneous inferences when this assumption is not met (Wakeley 2000; Nielsen and Beaumont 2009). The investigation of patterns of genetic population structure is therefore an important early step in the inference of demographic parameters from genetic data (Schraiber and Akey 2015). These patterns can also be of direct interest in understanding a species' or populations' evolutionary history or ecology.

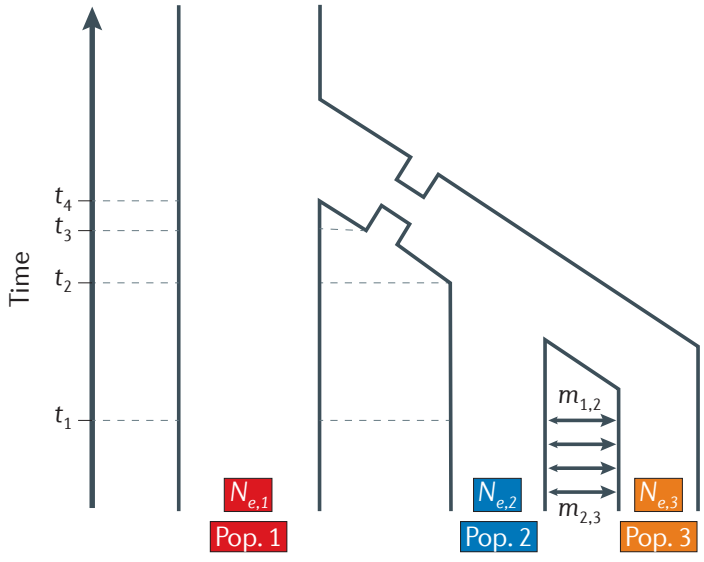


Fig. 1-5: A simple three-population demographic model with population divergence, changes in population size (at time t_3) and assymetric gene flow (m). N_e represents effective population sizes. Image: Schraiber and Akey, 2015

1.3 - Objectives and Thesis Structure

The general goal of this thesis is to explore past and present factors determining the current patterns of genetic variability of wolf populations. For this, a variety of phylogenomic and population genetics approaches were used to explore genetic variation data of diverse wolf populations. More specifically, the proposed objectives were:

i) to investigate the genetic structure and divergence of European wolf populations, with a special focus on their historical divergence and the genetic population structure within the Iberian wolf population;

ii) to infer the more general demographic history of Old and New World wolves, and how it relates to the domestication of the dog.

The research contained in this thesis is presented in the form of five scientific articles which either have already been published in international peer-reviewed journals or are currently in preparation for submission. In accordance with the two objectives described above, these papers constitute Chapters II and III of the present thesis. Due to the variation in format and graphical presentation of the papers published in different journals, the text, tables and figures have been formatted in an uniform way without changing their content.

The following chapters are structured as follows:

Chapter 2 - Wolf Population Structure and Differentiation in Europe

Paper I: **Silva P** et al. (in prep) Cryptic Population Structure and Evidence of Low Dispersal in the Iberian Wolf

This paper explores the genetic population structure of the Iberian wolf population by integrating genetic and ecological data from a large sample of individuals from this population. Geographical and genetically meaningful subpopulations were identified, their organization and connectivity was assessed and the resulting patterns were interpreted according to the ecological features and recent history of this population.

Paper II: **Silva P** et al. (in prep) Historic Demography and Divergence of European Wolf Populations

This paper makes use of full genome data of individuals from European wolf populations to investigate their evolutionary history, focusing on the timings of their divergence, levels of gene flow and long-term effective population sizes.

Chapter 3 - Historical Demography of Wolves and the Domestication of the Dog

Paper III: Freedman AH, Gronau I, Schweizer RM, Ortega Del-Vecchyo D, Han E, **Silva PM**, Galaverni M, Fan Z, Marx P, Lorente-Galdos B, Beale H, Ramirez O, Hormozdiari F, Alkan C, Vilà C, Squire K, Geffen E, Kusak J, Boyko AR, Parker HG, Lee C, Tadigotla V, Siepel A, Bustamante CD, Harkins TT, Nelson SF, Ostrander EA, Marques-Bonet T, Wayne RK & Novembre J (2014) Genome Sequencing Highlights the Dynamic Early History of Dogs. PLoS Genetics 10(1): e1004016. doi:10.1371/journal.pgen.1004016

This paper presents an analysis of several canid genomes with the goal of clarifying the process of dog domestication. Whole genomes from wolves of putative regions of dog origin, dogs of basal breeds and a golden jackal as an outgroup are used to gain an more detailed view of the shared history of wolves and dogs, including ancestral population sizes, population divergence times and rates of gene flow.

Paper IV: Fan Z*, **Silva P***, Gronau I, Wang S, Armero AS, Schweizer RM, Ramirez O, Pollinger J, Galaverni M, Ortega-Del Vecchyo D, Lianming D, Zhang W, Zhang Z, Xing J, Vilà C, Marques-Bonet T, Godinho R, Yue B & Wayne RK (2016) Worldwide patterns of genomic variation and admixture in gray wolves Genome Research 26:163-173 doi:10.1101/gr.197517.115
* equal contribution

This paper uses a wide sample of 34 canine genomes to investigate several questions regarding the evolutionary history of worldwide wolf populations and their relationship with dogs. The extensive sampling is used to assess the patterns of genomic variability across the entire geographic range of wolves, investigate their demographic history and admixture with dogs, and to explore questions related to dog domestication.

Paper V: **Silva P** et al. (in prep) The Effects of Population Structure and Sampling Scheme on Demographic Inferences from Microsatellite Data: an Empirical Test on the Iberian Wolf Population

This paper uses the Iberian wolf population as a test case to assess the limitations of widely-used demography inference methods based on microsatellite data. The confounding effects of population sub-division, gene flow among subpopulations and sampling scheme are evaluated on the performance of linkage disequilibrium effective population size estimation methods (LD-N_e) and summary-statistics (heterozygosity excess and M ratio) and likelihood-based methods (MSVAR) to infer past population size changes.

Chapter 4 - General Discussion

This chapter presents a general discussion of the results obtained by the research described in the previous chapters.

The supplementary materials for all papers in this thesis are aggregated in the **Appendix.**

1.4 - References

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

Álvares F. 2011. Ecologia e Conservação do Lobo (Canis lupus, L.) no Noroeste de Portugal. PhD Thesis, Universidade de Lisboa, Lisboa.

Andersen LW, Harms V, Caniglia R, Czarnomska SD, Fabbri E, Jędrzejewska B, Kluth G, Madsen AB, Nowak C, Pertoldi C, et al. 2015. Long-distance dispersal of a wolf, Canis lupus, in northwestern Europe. Mammal Research:163–168.

Andersone Z. 2002. Hybridisation between wolves and dogs in Latvia as documented using mitochondrial and microsatellite DNA markers. Mammalian Biology - Zeitschrift fur Saugetierkunde 67:79–90.

Beaumont MA, Rannala B. 2004. The Bayesian revolution in genetics. Nature reviews. Genetics 5:251–261.

Bertorelle G, Benazzo A, Mona S. 2010. ABC as a flexible framework to estimate demography over space and time: some cons, many pros. Molecular ecology:2609–2625.

Boitani L. 2000. Action plan for the conservation of the wolves (Canis lupus) in Europe. Council of Europe.

Boitani L, Ciucci P. 2009. Wolf management across Europe: Species Conservation without Boundaries. A new era for wolves and people: wolf recovery, human attitudes and policy:15–39.

Boyko AR, Boyko RH, Boyko CM, Parker HG, Castelhano M, Corey L, Degenhardt JD, Auton A, Hedimbi M, Kityo R, et al. 2009. Complex population structure in African village dogs and its implications for inferring dog domestication history. Proceedings of the National Academy of Sciences of the United States of America 106:13903–13908.

Campbell MMC, Tishkoff S a. 2008. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annual review of genomics and human genetics 9:403–433.

Caniglia R, Fabbri E, Greco C, Galaverni M, Manghi L, Boitani L, Sforzi A, Randi E. 2013. Black coats in an admixed wolf x dog pack is melanism an indicator of hybridization in wolves? European Journal of Wildlife Research 59:543–555.

Cardillo M, Purvis A, Sechrest W, Gittleman JL, Bielby J, Mace GM. 2004. Human population density and extinction risk in the world's carnivores. PLoS Biology 2:909–914.

Chambers SM, Fain SR, Fazio B, Amaral M. 2012. An account of the taxonomy of North American wolves from morphological and genetic analyses. North American Fauna 77:1–67.

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

Ciucci P, Reggioni W, Maiorano L, Boitani L. 2009. Long-Distance Dispersal of a Rescued Wolf From the Northern Apennines to the Western Alps. Journal of Wildlife Management 73:1300–1306.

Cornuet J, Luikart G. 1996. Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. Genetics 144:2001–2014.

Cutter AD. 2013. Integrating phylogenetics, phylogeography and population genetics through genomes and evolutionary theory. Molecular Phylogenetics and Evolution 69:1172–1185.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics 12:499–510.

Druzhkova AS, Thalmann O, Trifonov V a., Leonard J a., Vorobieva N V., Ovodov ND, Graphodatsky AS, Wayne RK. 2013. Ancient DNA Analysis Affirms the Canid from Altai as a Primitive Dog. PLoS ONE 8:0–5.

Edwards S V. 2009. Is a new and general theory of molecular systematics emerging? Evolution 63:1–19.

Ekblom R, Galindo J. 2010. Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity 107:1–15.

Emerson B. 2001. Revealing the demographic histories of species using DNA sequences. Trends in Ecology & Evolution 16:707–716.

Estes J a, Terborgh J, Brashares JS, Power ME, Berger J, Bond WJ, Carpenter SR, Essington TE, Holt RD, Jackson JBC, et al. 2011. Trophic downgrading of planet Earth. Science (New York, N.Y.) 333:301–306.

Frankham R. 2005. Genetics and extinction. Biological Conservation 126:131–140.

Fritts SH, Stephenson RO, Hayes RD, Boitani L. 2003. Wolves and humans. In: Mech LD, Boitani L, editors. Wolves: Behavior, Ecology, and Conservation. p. 289–316.

Fuller TK, Mech LD, Cochrane JF. 2003. Wolf population dynamics. In: Mech LD, Boitani L, editors. Wolves: Behavior, Ecology, and Conservation. The University of Chicago Press. p. 161–191.

Garza JC, Williamson EG. 2001. Detection of reduction in population size using data from microsatellite loci. Molecular ecology 10:305–18.

Germonpré M, Lázničková-Galetová M, Sablin M V. 2012. Palaeolithic dog skulls at the Gravettian Předmostí site, the Czech Republic. Journal of Archaeological Science 39:184–202.

Germonpré M, Sablin M V., Stevens R, Hedges R, Hofreiter M, Stiller M, Despres V. 2009. Fossil dogs and wolves from Palaeolithic sites in Belgium, the Ukraine and Russia: osteometry, ancient DNA and stable isotopes. Journal of Archaeological Science 36:473–490.

Godinho R, Llaneza L, Blanco JC, Lopes S, Alvares F, García EJ, Palacios V, Cortés Y, Talegón J, Ferrand N, et al. 2011. Genetic evidence for multiple events of hybridization between wolves and domestic dogs in the Iberian Peninsula. Molecular ecology 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 non-invasive samples with ancestry informative markers. Molecular Ecology Resources:317–328.

Gray MM, Sutter N, Ostrander EA, Wayne R. 2010. The IGF1 small dog haplotype is derived from Middle Eastern gray wolves. BMC Biology 8:16.

Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A. 2011. Bayesian inference of ancient human demography from individual genome sequences. Nature Genetics 43:1031–1034.

Hewitt G. 2000. The genetic legacy of the Quaternary ice ages. Nature 405:907–13.

Hickerson MJ, Carstens BC, Cavender-Bares J, Crandall KA, Graham CH, Johnson JB, Rissler L, Victoriano PF, Yoder AD. 2010. Phylogeography's past, present, and future: 10 years after Avise, 2000. Molecular Phylogenetics and Evolution 54:291–301.

Hindrikson M, Remm J, Männil P, Ozolins J, Tammeleht E, Saarma U. 2013. Spatial Genetic Analyses Reveal Cryptic Population Structure and Migration Patterns in a Continuously Harvested Grey Wolf (Canis lupus) Population in North-Eastern Europe. PLoS ONE 8:e75765.

Imbert C, Caniglia R, Fabbri E, Milanesi P, Randi E, Serafini M, Torretta E, Meriggi A. 2016. Why do wolves eat livestock?: Factors influencing wolf diet in northern Italy. Biological Conservation 195:156–168.

Jędrzejewski W, Schmidt K, Theuerkauf J, Jędrzejewska B, Kowalczyk R. 2007. Territory size of wolves Canis lupus: Linking local (Białowieża Primeval Forest, Poland) and Holarctic-scale patterns. Ecography 30:66–76.

Kaczensky P, Chapron G, von Arx M, Huber D, Andrén H, Linnell JDC. 2012. Status, Management and Distribution of Large Canivores - Bear, Lynx, Wolf & Wolverine - in Europe.

Kingman JFC. 1982. The coalescent. Stochastic Processes and their Applications 13:235–248.

Knowles LL. 2009. Statistical Phylogeography. Annual Review of Ecology, Evolution, and Systematics 40:593–612.

Kuhner MK. 2009. Coalescent genealogy samplers: windows into population history. Trends in Ecology & Evolution 24:86–93.

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. Proceedings of the National Academy of Sciences of the United States of America 109:8878–83.

Leonard J a, Wayne RK, Wheeler J, Valadez R, Guillén S, Vilà C. 2002. Ancient DNA evidence for Old World origin of New World dogs. Science 298:1613–1616.

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

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

Lohmueller KE. 2014. The Impact of Population Demography and Selection on the Genetic Architecture of Complex Traits. PLoS Genetics 10.

Lopes JS, Beaumont M a. 2010. ABC: A useful Bayesian tool for the analysis of population data. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases 10:825–832.

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. Biodiversity and Conservation 24:2105–2108.

López-Bao JV, Sazatornil V, Llaneza L, Rodríguez A. 2013. Indirect Effects on Heathland Conservation and Wolf Persistence of Contradictory Policies that Threaten Traditional Free-Ranging Horse Husbandry. Conservation Letters 6:448–455.

Lucchini V, Galov A, Randi E. 2004. Evidence of genetic distinction and long-term population decline in wolves (Canis lupus) in the Italian Apennines. Molecular Ecology 13:523 – 536.

Luikart G, Cornuet J-M. 1998. Empirical Evaluation of a Test for Identifying Recently Bottlenecked Populations from Allele Frequency Data. Conservation Biology 12:228–237.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P. 2003. The power and promise of population genomics: from genotyping to genome typing. Nature reviews. Genetics 4:981–994.

Marjoram P, Tavaré S. 2006. Modern computational approaches for analysing molecular genetic variation data. Nature Reviews Genetics 7:759–770.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. Molecular Phylogenetics and Evolution 66:526–538.

Mech LD. 1977. Wolf-pack buffer zones as prey reservoirs. Science 198:320–321.

Mech LD. 1999. Alpha status, dominance, and division of labor in wolf packs. Canadian Journal of Zoology 77:1196–1203.

Mech LD, Boitani L. 2003. Wolf Social Ecology. In: Mech LD, Boitani L, editors. Wolves: Behavior, Ecology, and Conservation. p. 1–34.

Mech LD, Harper EK. 2002. Differential use of a wolf, Canis lupus, pack territory edge and core. Canadian Field-Naturalist 116:315–316.

Merrill SB, Mech LD. 2000. Details of Extensive Movements by Minnesota Wolves (Canis lupus). The American Midland Naturalist 144:428–433.

Metzker ML. 2009. Sequencing technologies — the next generation. Nature Reviews Genetics 11:31–46.

Newsome TM, Boitani L, Chapron G, Ciucci P, Dickman CR, Dellinger JA, L??pez-Bao J V., Peterson RO, Shores CR, Wirsing AJ, et al. 2016. Food habits of the world's grey wolves. Mammal Review:1–15.

Nielsen R. 2005. Molecular signatures of natural selection. Annual Review of Genetics 39:197–218.

Nielsen R, Beaumont MA. 2009. Statistical inferences in phylogeography. Molecular Ecology 18:1034–1047.

Nowak R, Federoff N. 2002. The systematic status of the Italian wolf Canis lupus. Acta theriologica 47:333–338.

Nowak RM. 1995. Another look at wolf taxonomy. Ecology and conservation of wolves in a changing world. Canadian Circumpolar Institute, Edmonton 375.

Nowak RM. 2003. Wolf evolution and taxonomy. In: Wolves: Behavior, ecology, and conservation. University of Chicago Press: Chicago, IL, USA. p. 239–258.

Ovodov ND, Crockford SJ, Kuzmin Y V., Higham TFG, Hodgins GWL, van der Plicht J. 2011. A 33,000-Year-Old incipient dog from the Altai Mountains of Siberia: Evidence of the earliest domestication disrupted by the last Glacial Maximum. PLoS ONE 6:4–10.

Pang J-F, Kluetsch C, Zou X-J, Zhang A -b., Luo L-Y, Angleby H, Ardalan A, Ekström C, Sköllermo A, Lundeberg J, et al. 2009. mtDNA Data Indicates a Single Origin for Dogs South of Yangtze River, less than 16,300 Years Ago, from Numerous Wolves. Molecular Biology and Evolution 26:2849–2864.

Parker HG, Kim L V, Sutter NB, Carlson S, Lorentzen TD, Malek TB, Johnson GS, DeFrance HB, Ostrander E a, Kruglyak L. 2004. Genetic structure of the purebred domestic dog. Science 304:1160–4.

Peterson RO, Ciucci P. 2003. The wolf as a carnivore. In: Mech LD, Boitani L, editors. Wolves: Behavior, Ecology, and Conservation. The University of Chicago Press. p. 104–130.

Petrucci-Fonseca F. 1990. O lobo ibérico (Canis lupus signatus Cabrera, 1907) em Portugal. PhD Thesis, Faculdade de Ciencias da Universidade de Lisboa, Lisbon, Portugal.

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

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

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

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. Conservation Genetics 3:31–45.

Randi E, Lucchini V, Christensen MF, Mucci N, Funk SM, Dolf G, Loeschcke V. 2000. Mitochondrial DNA Variability in Italian and East European Wolves: Detecting the Consequences of Small Population Size and Hybridization. Conservation Biology 14:464–473.

Ražen N, Brugnoli A, Castagna C, Groff C, Kaczensky P, Kljun F, Knauer F, Kos I, Krofel M, Luštrik R, et al. 2016. Long-distance dispersal connects Dinaric-Balkan and

Alpine grey wolf (Canis lupus) populations. European Journal of Wildlife Research:137–142.

Ripple WJ, Beschta RL. 2012. Trophic cascades in Yellowstone: The first 15years after wolf reintroduction. Biological Conservation 145:205–213.

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 343:1241484.

Rogers AR, Harpending H. 1992. Population growth makes waves in the distribution of pairwise genetic differences. Molecular Biology and Evolution 9:552–569.

Rosenberg NA, Nordborg M. 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nature Reviews. Genetics 3:380–90.

Rutledge LY, Devillard S, Boone JQ, Hohenlohe PA, White BN, Drive EB, Canada KJ, Biome L De. 2015. RAD sequencing and genomic simulations resolve hybrid origins within North American Canis. Biology Letters 11:20150303.

Savolainen P, Leitner T, Wilton AN, Matisoo-Smith E, Lundeberg J. 2004. A detailed picture of the origin of the Australian dingo, obtained from the study of mitochondrial DNA. Proceedings of the National Academy of Sciences of the United States of America 101:12387–90.

Savolainen P, Zhang Y, Luo J, Lundeberg J, Leitner T. 2002. Genetic evidence for an East Asian origin of domestic dogs. Science 298:1610–3.

Schraiber JG, Akey JM. 2015. Methods and models for unravelling human evolutionary history. Nature Reviews Genetics 16:727–740.

Sillero-Zubiri C, Hoffmann M, Macdonald D. 2004. Canids: Foxes, Wolves, Jackals and Dogs. Status Survey and Action Plan. Sillero-Zubiri C, Hoffmann M, Macdonald DW, editors. Gland, Switzerland and Cambridge, UK: IUCN/SSC Canid Specialist Group.

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. Current Biology:1–5.

Sommer R, Benecke N. 2005. Late-Pleistocene and early Holocene history of the canid fauna of Europe (Canidae). Mammalian Biology - Zeitschrift für Säugetierkunde 70:227–241.

Taberlet P, Fumagalli L, Wust-Saucy A-G, Cosson JF. 1998. Comparative phylogeography and postglacial colonization routes in Europe. Molecular ecology 7:453–64.

Tajima F. 1989. The effect of change in population size on DNA polymorphism. Genetics 123:597–601.

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 (New York, N.Y.) 342:871–4.

Theuerkauf J, Rouys S, Jędrzejewski W. 2003. Selection of den, rendezvous, and resting sites by wolves in the Bialowieza Forest, Poland. Canadian Journal of Zoology 81:163–167.

Vilà C. 1993. Aspectos morfológicos y ecológicos del lobo ibérico (Canis lupus L.). PhD Thesis. Universidad de Sevilla, Sevilla, Spain.

Vilà C, Amorim IR, Leonard JA, Petrucci-Fonseca F, Posada D, Crandall KA, Castroviejo J, Ellegren H, Wayne RK. 1999. Mitochondrial DNA phylogeography and population history of the grey wolf canis lupus. Molecular Ecology 8:2089–103.

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 CW, Sundqvist A-K, Flagstad Ø, Casulli A, Andersone Z, Ellegren H, Kojola I, Valdmann H, Halverson J. 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. Conservation Biology 13:195–198.

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.

Wabakken P, Sand H, Liberg O, Bjärvall A. 2001. The recovery, distribution, and population dynamics of wolves on the Scandinavian peninsula, 1978-1998. Canadian Journal of Zoology 79:710–725.

Wakeley J. 2000. The effects of subdivision on the genetic divergence of populations and species. Evolution 54:1092–1101.

Wang G-D, Zhai W, Yang H-C, Wang L, Zhong L, Liu Y-H, Fan R-X, Yin T-T, Zhu C-L, Poyarkov AD, et al. 2016. Out of southern East Asia: the natural history of domestic dogs across the world. Cell Research 26:21–33.

Wang X, Tedford RH. 2008. Dogs: Their Fossil Relatives and Evolutionary History. Columbia University Press.

Wayne RK, Girman DJ, Koepfli KP, Lau LM, Geffen E, Marshall CR. 1997. Molecular systematics of the Canidae. Systematic Biology 46:622–646.

Wayne RK, Jenks SM. 1991. Mitochondrial DNA analysis implying extensive hybridization of the endangered red wolf Canis rufus. Nature 351:565–568.

Wayne RK, Ostrander EA. 2005. The canine genome. Genome research 15:1706–1716.

Wilson DE, Reeder DM. 2005. Mammal species of the world: a taxonomic and geographic reference. Wilson DE, Reeder DM, editors. JHU Press.

Wilson PJ, Grewal S, Lawford ID, Heal JN, Granacki AG, Pennock D, Theberge JB, Theberge MT, Voigt DR, Waddell W, et al. 2000. DNA profiles of the eastern Canadian wolf and the red wolf provide evidence for a common evolutionary history independent of the gray wolf. Canadian Journal of Zoology 78:2156–2166.

Wright S. 1931. Evolution in Mendelian populations. Genetics 16:97–159.

Chapter 2 - WOLF POPULATION STRUCTURE AND DIFFERENTIATION IN EUROPE

Paper I: Silva P et al. (in prep) Cryptic Population Structure and Evidence of Low Dispersal in the Iberian Wolf

Paper II: Silva P et al. (in prep) Historic Demography and Divergence of European Wolf Populations.

Paper I - Cryptic Population Structure and Evidence of Low Dispersal in the Iberian Wolf

Silva P et al.

(manuscript in preparation for submission)

Abstract

While highly mobile carnivore mammals like the gray wolf have the capability of maintaining high levels of gene flow across large geographic distances, many recent studies have found surprising levels of genetic population structure in these populations. In this study, we examine the spatial genetic population structure of the Iberian wolf population, a currently isolated population with a recent history of human-induced decline and fragmentation. We use Bayesian clustering methods to identify geographically and genetically meaningful subpopulations, and combine this information with the location of sampled individuals and spatial behavior of wolves to investigate the organization of these groups and the more general patterns of gene flow. We find an exceptionally reticulated pattern of population structure in the Iberian Peninsula with low levels of gene flow between subpopulations. This structure can be described at two hierarchical levels, with 4 and 11 geographically meaningful genetic clusters that can be discerned. The identified subpopulation are characterized by moderate to high levels of differentiation (average pairwise F_{ST}=0.09-0.19), low levels of admixture, a very reduced number of dispersant individuals and varying degrees of genetic diversity that probably reflect local histories. Our results suggest that both the recent demographic history of the Iberian wolf population, as well as yet unknown obstacles to dispersal influence the high levels of genetic structure in this population.

Introduction

The rates at which populations exchange genes (i.e., gene flow) is one of the main driving forces determining the scale and magnitude of population genetic differentiation. It is expected that high rates of gene flow will lead to little spatial genetic structuration at small scales, although isolation-by-distance patterns can emerge with increasing geographical distances, even in the absence of barriers to dispersal (Manel et al. 2003). Until recently, little attention has been paid to the genetic structure of continuously distributed vagile species, such as large mammals. For example, many mammalian carnivores possess a high mobility, can disperse over very large distances and can occupy a great variety of habitats. It is expected therefore that they will have a strong potential for maintaining high rates of gene flow, and consequently reduced genetic differentiation across large parts of their ranges (e.g. Dalén et al. 2005; Carmichael et al. 2007; Tammeleht et al. 2010; Teacher et al. 2011; Row et al. 2012). However, recent studies have shown that continuous populations of vagile species can also exhibit notable levels of population genetic structure at different spatial scales, which cannot be explained by past or current barriers to dispersal alone (e.g. Rueness et al. 2003; Sacks et al. 2004; McRae et al. 2005; Pilot et al. 2006; Hindrikson et al. 2013).

Traditionally, population genetic structure has been considered the outcome of well-defined behavioral traits (e.g., colonies) (Surridge et al. 1999), spatial constraints, including geographical distance and topographical or human-made barriers (Broquet et al. 2006; Coulon et al. 2006), or historical factors, such as past range restrictions (Taberlet et al. 1998). Human-related factors, such as hunting pressure can also be influence the observed structuration patterns (Andreasen et al. 2012). However, recent studies have brought attention to less well-understood mechanisms promoting genetic population division by limiting dispersal, such as natal habitat-biased dispersal (Davis and Stamps 2004; Sacks et al. 2004; Pilot et al. 2006) and territoriality (Waters et al. 2013). Wolf (*Canis lupus*) dispersal distance, for example, is thought to be affected by population density and the probability of finding a mate, with longest dispersal distances having been reported from low-density populations (Wabakken et al. 2001; Ražen et al. 2016). New wolf packs can also be established on the edges of the natal territory of one of the founding individuals, leading to familiar ties between geographically close groups (Mech and Boitani 2003). As a consequence, genetic clusters consisting of related individuals can emerge in wolf populations, although the level of genetic differentiation will depend on the rates of inter-pack gene flow (Jansson et al. 2012). Despite their high dispersal capabilities, wolf populations can nonetheless present low levels of gene flow

and short dispersal distances, leading to the emergence of genetic structuration at small spatial scales and of genetically differentiated groups that behave like a meta-population (Carmichael et al. 2007; vonHoldt et al. 2010; Scandura et al. 2011; Jansson et al. 2012; Stronen et al. 2012).

The term 'cryptic population structure' (CPS) has been applied to discrete genetic clusters corresponding to spatial units in the absence of gaps in the local distribution and of physical barriers to movement, and often without evident phenotypic distinction (Sacks et al. 2005). The existence of cryptic genetic clusters is of great interest in terms of the behavioral and social processes they can reflect, their potential role in ecological and evolutionary processes, and ultimately their consequences for conservation and population management. Demographic connectivity among clusters will depend on the relative contributions of migrants to subdivision growth rates and, consequently, a source-sink dynamic among cryptic clusters may be hypothesized (Andreasen et al. 2012). On the other hand, the inclusion of behavioral data, such as spatial information from tracked animals, in addition to assignment procedures and estimates of differentiation from genetic data, has allowed the assessment of CPS patterns on a finer spatial scale and how individual clusters are organized (e.g. Sacks et al. 2005; Broquet et al. 2006; Boulet et al. 2007).

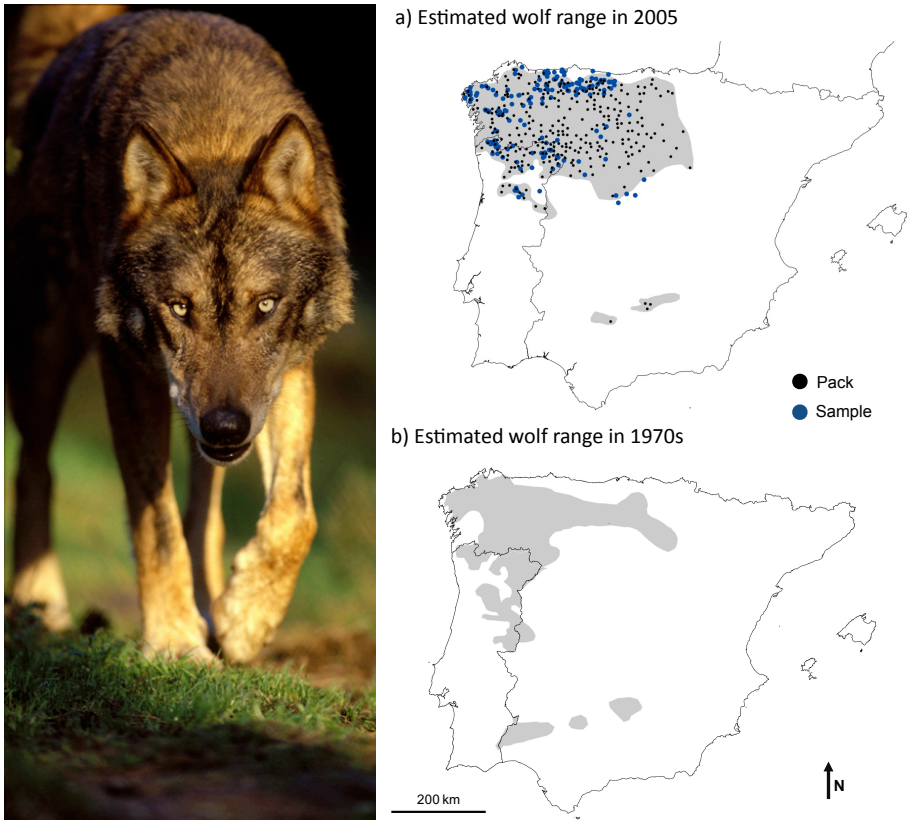


Fig. 2-1: a) Current wolf distribution in the Iberian Peninsula with pack and sample locations; b) areas of common and uncommon wolf presence in the 1970s according to Valverde, 1971.

Iberian wolves represent the largest wolf population in Western Europe, currently being isolated from the remaining European wolf populations (Chapron et al. 2014). This population suffered a severe decline since the beginning of the 20th century until the 1970s due to intense persecution (Valverde 1971). In recent decades, the population has been expanding, and currently numbers >2000 individuals in >300 packs, distributed mainly in the northwestern Iberian Peninsula, over ca. 140,000 km^2 (Blanco and Cortés 2012; Chapron et al. 2014). This population occurs in a human-dominated landscape (Blanco and Cortés 2007; Eggermann et al. 2011; Llaneza et al. 2012) and is remarkable for feeding mainly on anthropogenic sources of food (López-Bao et al. 2013; Llaneza and López-Bao 2015; Newsome et al. 2016). Additionally, a small and isolated wolf population on the brink of extinction remains in Sierra Morena (Southern Spain)(López-Bao et al. 2015). Here, we examine the spatial genetic population structure of the Iberian wolf population in light of the expected mobility capabilities of this species and the expansion processes occurred in this population in the last decades. We use Bayesian clustering methods to identify geographically and genetically meaningful groups, and combine this information with the location of sampled individuals as well as information about the spatial behavior of wolves to investigate the spatial organization of these groups. In particular, we intend to explore what distinct genetic groups can be identified in the Iberian wolf population, how they are spatially organized and connected, how the genetic diversity is partitioned among groups, and if the combination of spatial, behavioral and genetic data can help to elucidate how these groups are organized.

Materials and Methods

Sample collection and selection

Between 1995 and 2014, we collected 289 wolf tissue samples in Portugal and Spain (Fig. 2-1) from dead animals (mainly road kills, poached and legally culled or hunted wolves). Animals were never specifically killed for this study. For each tissue sample we recorded the GPS coordinates where the sample was collected, the sex and, when possible, the approximate age of the animal. Age was estimated by dental pattern and wear, according to Gipson et al. 2000. Our sampling encompasses ca. 80% of the wolf estimated range in 2005 in the Iberian Peninsula (Blanco and Cortés 2001; Álvares et al. 2005; Pimenta et al. 2005; Blanco and Cortés 2012).

Because several individuals (samples) could belong to the same pack, we used the following criteria to avoid close familiar relationships in our dataset: i) firstly, we calculated the Euclidean distance among all tissue sample locations, and plotted all

samples in relation to the distribution of the wolf packs detected between 1993 and 2005 (Fig. 2-1a); ii) secondly, for pairs or groups of samples separated less than 10 km, and/or within the same buffer area (100 km^2) of wolf packs (centered on the estimated rendezvous sites in that period), we noted the date in which every sample was collected; iii) considering the annual cycle of wolves (from May of a given year to the next May), we focused on those samples separated ≥5 yrs. This criteria ensured that more than 1 generation spanned between samples (mean generation time in wolves estimated in 3-4 yrs (Aspi et al. 2006; Mech et al. 2016)); iv) we selected a maximum of two samples for each 5-yr temporal window. To do this, we took into account preliminary information of missing data from microsatellite analyses (selecting those samples with the best performance), and the age of individuals, prioritizing adult individuals (> 2 yrs.). Samples with >20% missing data were also excluded for subsequent analyses. The final dataset used for this study contained 218 wolf tissue samples.

DNA extraction, markers and genotyping

Total genomic DNA was extracted using the QIAGEN DNeasy Blood & Tissue Kit (Qiagen, Valencia, CA, USA) according to the manufacturer's instructions. DNA quality and concentration was assessed using agarose gel electrophoresis and quantified in a Qubit fluorometer (Thermo Fisher Scientific, Wilmington, DE, USA). Samples were amplified for a set of 46 microsatellite loci in four multiplex reactions (MS1 to MS4) following Godinho et al. 2011 and Godinho et al. 2015) (Supplementary Table 2). PCR products were run on an ABI3130xl Genetic Analyser (Applied Biosystems, Waltham, Massachusetts, USA) using GeneScan500 LIZ size standard. The results were checked manually, and alleles scored, using *Genemapper* 4.1 (Applied Biosystems).

Bayesian clustering analysis

We employed the Bayesian clustering algorithm of *Structure* 2.3.3 (Pritchard et al. 2000) to group all samples into clusters using their multilocus genotypes. The program was run for 2 million iterations after 500 thousand burn-in steps, using the admixture and correlated allele frequencies models, assuming values of the K parameter (number of populations) between 1 and 12. For each K, runs were repeated 20 times to ensure the consistency of the results. To select the most likely number of groups we tabulated the posterior probabilities of the data for each K (ln Pr(X|K)) and employed the procedure proposed by Evanno et al. 2005, which involves calculating the quantity ΔK for each pair of successive Ks, using *StructureHarvester* (Earl and vonHoldt 2012).

Having chosen K, we classified each individual in each cluster *i* based on the distribution of the individual membership proportions to that cluster (q_i). A cut-off value was set for each cluster at the point where a gap in the distribution of membership proportions was visible and above which membership proportions no longer varied considerably between individuals. Individuals above this threshold were considered to fully belong to the respective cluster (i.e., 'pure' individuals), while all individuals below the cut-off point were considered 'admixed', i.e., having a genetic background in more than one genetic cluster (Supplementary Table 1).

Spatial population analyses

Samples were spatially projected using QGIS 2.8.2 (QGIS Development Team 2015). We grouped samples according to similar posterior assignment probabilities to the different groups identified in our dataset in order to spatially identify geographical clusters. For a given genetic cluster, all adjacent individuals with a membership proportion to the same cluster higher than 0.50 were grouped. These individuals were used to define a minimum convex polygon for each genetic cluster (MCP_{total}). Individuals for which position and/or genetic makeup did not allow for a clear attribution were not included. In a second step, we defined another minimum convex polygon (MCP_{pure}) encompassing the area comprising only 'genetically pure' individuals. All 'pure' individuals (as defined by their genetic membership proportions, as explained above) that were not located within the MCP_{total} area of their respective population were considered as dispersants.

For each geographical population identified in space, we calculated the area of the MCP_{total} and MCP_{pure}, we computed the mean membership proportions, the percentage of 'admixed' individuals at the MCP_{total} level, and within the MCP_{pure}, the number of immigrants (i.e. dispersants from other populations falling with MCP_{total}). Moreover, in order to gain insights into the connection between genetic clusters, we also calculated the number of contributing genetic clusters to the "genetic pool" of every group, by considering only groups with >5% membership proportions, and the smallest distance to the closest group (km, measured as the smallest Euclidean linear distance between MCP_{total} borders).

Population genetics analyses

For each geographical group, as well as for each genetic cluster identified by *Structure* (including only 'pure' individuals), we calculated standard population genetics parameters including: i) the number of alleles, ii) observed and expected

heterozygosities, and iii) the fixation coefficient (F), using GenAlEx 6.501 (Peakall and Smouse 2012); iv) allelic richness and private allelic richness, rarefied to the smallest sample size, using ADZE 1.0 (Szpiech et al. 2008). We used Genepop 4.3 (Rousset 2008) to test for significant deviation from Hardy-Weinberg equilibrium and association between genotypes at pairs of loci (linkage disequilibrium). Statistical significance levels were adjusted using Bonferroni corrections(Rice 1989). Additionally, we also estimated the average relatedness between pairs of individuals for each group/cluster using the estimators of Queller and Goodnight (Queller and Goodnight 1989) and Lynch and Ritland (Lynch and Ritland 1999) with GenAlEx.

Differentiation between groups/clusters was measured by F_{ST} calculated from an AMOVA (Excoffier et al. 1992), using 1,000 permutations to test their significance, and Jost's D distance (Jost 2008) using GenoDive 2.07b27 (Meirmans and Van Tienderen 2004).

Spatial behavior of wolves

To investigate whether the identified genetic structure, based on genotypic and geographical data, is reflected in the behaviour of individual animals, we analyzed the spatial tracking data from collared wolves. Spatial tracking information from a total of 85 wolves, collected from 1982 to 2015 in the context of several research projects on the ecology of this species in Portugal and Spain, was used in these comparisons. Wolves were captured using Belisle® leg-hold snares (Edouard Belisle, Saint Veronique, PQ, Canada). See also references Pereira et al. 1985; Moreira 1992; Pimenta 1998; Grilo et al. 2002; Roque et al. 2011. Traps were monitored twice every day, in the early morning and late afternoon. Animals were chemically immobilized by intramuscular injection (handheld syringe, pole syringe, or blow-dart) of a mixture of ketamine (Imalgene®, Merial, Lyon, France) and medetomidine (Domitor®, Merial, Lyon, France) or sedated exclusively with medetomidine. Immobilization was usually reversed by the intramuscular injection of atipamezole (Revertor®, Merial, Lyon, France). Wolves were equipped with VHF (Followit, Sweden and Telonics, USA) or GPS-GSM/Iridium collars (Followit, Sweden and Vectronic, Germany). Out of the 85 wolves used, 30 animals were equipped with VHF collars; whereas 55 wolves had GPS collars. Our dataset was composed by 39 females and 46 males and contained information from 41 wolves with an estimated age <2 yrs and 44 wolves with an estimated age >2 yrs (Supplementary Table 3). Monitoring period for the wolves considered in our dataset averaged 378 days (overall range: 16-2,129 days; VHF collars: mean = 678 days, range = 44-2,129; GPS collars: mean = 214 days, range = 16-632) (Supplementary Table 3); and the mean

number of locations per wolf was 2685 (overall range 15-13,709; VHF collars: mean = 163, range = 15-417; GPS collars: mean = 3,573, range = 153-13,709) (Supplementary Table 3).

For each wolf, we calculated a minimum convex polygon (MCP) using all of the VHF/GPS locations during the entire monitoring time. Since in this study we were interested in detecting overlaps between individual wolf MCPs and the spatial distribution of the detected genetic groups (MCP_{total}), we maximized potential overlaps by considering the entire monitoring period of wolves regardless of changes in social status over time. Overall, the mean full monitoring MCP for wolves was 408 km^2 (range 14-2810) (Supplementary Table 3). Subsequently, we counted the number of genetic groups (i.e. MCP_{total}) overlapping with every individual wolf MCP.

Results

Clustering analysis and identification of admixed individuals

We employed the Bayesian clustering algorithm of *Structure* 2.3.3 (Pritchard et al. 2000) to group 218 wolf samples, spanning most of their distribution in the Iberian Peninsula (Fig. 2-1), into clusters based on their genotypes at 46 microsatellite loci. Posterior probability values Pr(X|K) increased with increasing values of K (number of assumed genetic clusters), starting to plateau after K=4, while ΔK values (Evanno et al. 2005) peaked at K=2 and K=4 (Supplementary Fig. 1). Results between *Structure* runs for the same K were generally very consistent, with variance increasing for higher values of K (Supplementary Fig. 1).

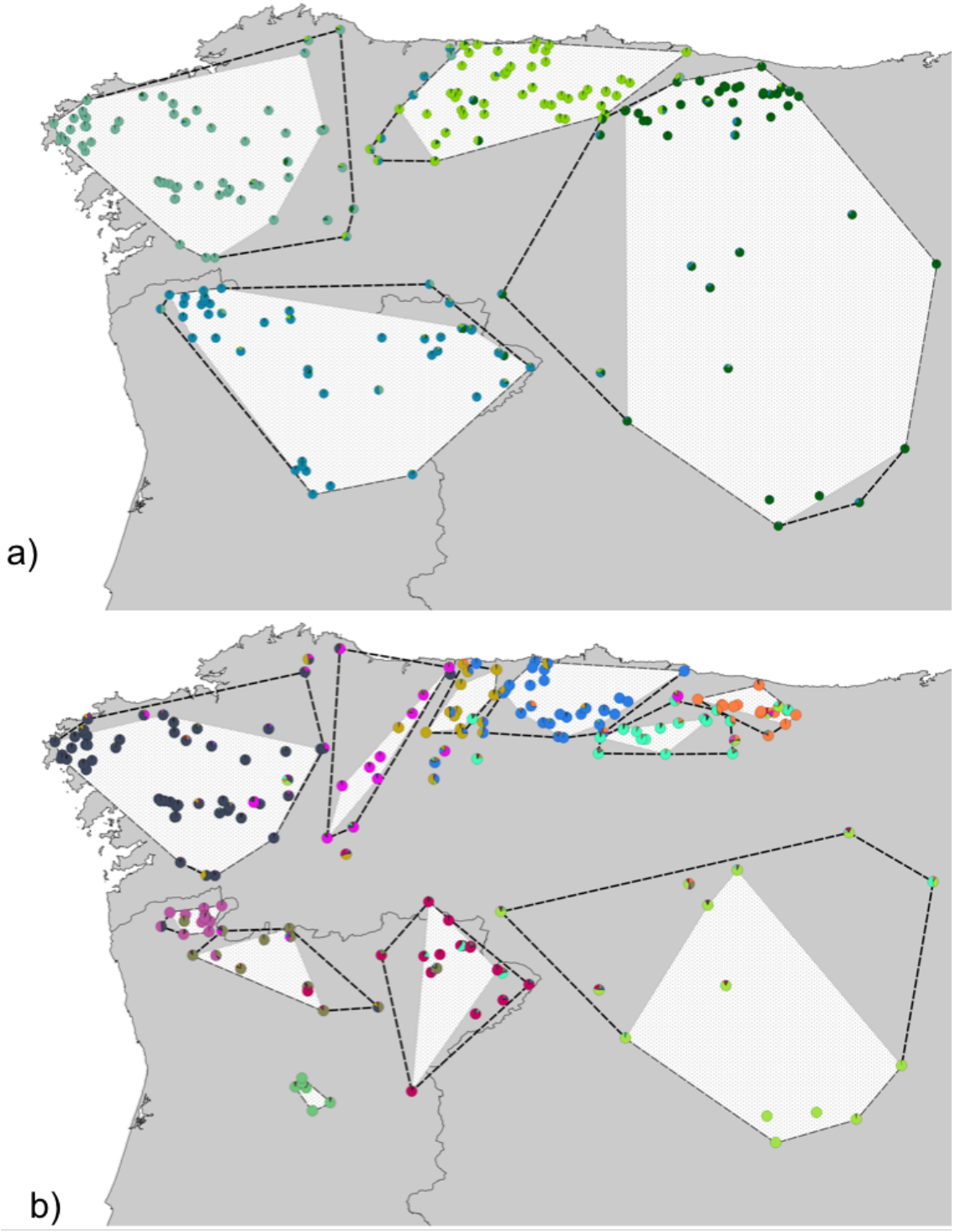


Fig. 2-2: Membership proportions of individuals sampled in this study, according to the Structure analysis, for K=4 (a) and K=11 (b). MCP_{total} and MCP_{pure} areas are represented as dashed black lines and white areas respectively.

Determining the number of groups from *Structure* runs is not straightforward: the program authors warn that Pr(X|K) serves only as indication of the true number of groups (Pritchard et al. 2000); manual of the program), and biological plausibility has to be taken into account. The ΔK statistic was developed as a further aid to this task, by measuring the rate of change in the probability of data between successive values of K (Evanno et al. 2005). In this study we chose to emphasize the geographical sense of the identified clusters, by taking in consideration both the likelihood of the results and the geographical sense of the identified clusters. While K=2 presented the highest ΔK value, this partitioning was not as geographically well supported (Supplementary Fig. 2), and probability values still increased substantially for higher K values (Supplementary Fig. 1). For subsequent analyses, we therefore considered two levels of genetic structure: K=4, for which a substantial change in ΔK was observed, and K=11, which showed the highest probability for which a geographically meaningful patterns could still be discerned (Supplementary Fig. 1 & 2).

Based on the distribution of membership proportions within each genetic cluster, we classified individuals as either 'genetically pure' or 'admixed'. Average cut-off values were high: for K=4: 0.83 (0.77-0.87), and for K=11: 0.88 (0.81-0.95) (Supplementary Table 1). ‘Pure’ individuals exhibited very high membership proportions (*qi*): K=4: 0.94-0.96, K=11: 0.90-0.97. In total, 166 (76%) and 137 (63%) our of 218 individuals were classified as ‘pure’ wolves, for K=4 and K=11, respectively.

Spatial population analyses

Spatial projection of samples and their respective membership proportions for K=4 revealed that the identified genetic clusters correspond to the regions of Galicia (NW Spain), Northern Portugal, the Western Cantabrian Mountains, and a large group that extends from Eastern Cantabrian Mountains to the plateau of Castilla y León (Fig. 2-2). Considering K=11, we identified the following groups (Fig. 2-2): in Portugal, three groups (Alto Minho, W and E Trás-os-Montes) to the North and one to the South (S Douro) of the Douro river; in Spain, two groups in Galicia (Western and Eastern Galicia), four groups in the Cantabrian Mountains (designated Western, Central, Eastern and Southeastern Asturias); and one group occupying the plateau in Castilla-León (Fig. 2-2). The E and SE Asturias populations showed greater affinity with the group of Castilla-León, clustering with this group at K=4. The group designated E Galicia at K=11 included individuals from both the Galician and Cantabrian populations at K=4.

We defined geographical populations based on the sample location, grouping together adjacent individuals with membership proportions to the same genetic cluster

>0.50. The locations of these individuals were used to define a minimum convex polygon for each population (MCP_{total}); similar MCPs determined by the locations of only 'pure' individuals were also defined - MCP_{pure} (Fig. 2-2). At K=4, the identified geographical groups contained a similar sample size (range 47-59 individuals) (Table 2-1). A higher variability was observed in sample sizes at K=11: the S Douro group showed the smallest size (6 individuals), while W Galicia was the group more intensively sampled (53 individuals) (Table 2-2). Individual membership proportions to genetic clusters within the respective geographic groups were, on average, higher for K=4 (mean Q_i over populations: 0.87) than for K=11 (0.79), while in both cases the genetic contribution from external groups was extremely low: each group received the genetic contribution of only one or two other groups (5% threshold), indicating low exchange of genes among groups, and not always from the closest ones. Groups from E and SE Asturias stand out as presenting substantial mutual genetic contribution (>0.10), while on the other hand, the isolation of the group of S Douro was remarkable (Fig. 2-3). Admixed individuals represented on average 23% (12-35%) of each group for K=4, and 34%% (15-59%) for K=11.

Most individuals (89-94%) occupied the area defined by the most external 'genetically pure' individuals (MCP_{pure}). 64% and 49% of admixed individuals were found inside the MCP_{pure} of their geographical population, for K=4 and K=11, respectively (Tables 2-1 & 2-2). MCP_{pure} areas represented on average 84% (74%-92%) and 63% (22%-100%) of MCP_{total} area at K=4 and K=11, respectively (Tables 2-1 & 2-2).

The proportion of individuals classified as dispersants in our dataset, i.e. presenting a membership proportion to one cluster above the cut-off value but found within the MCP_{total} of another, was extremely low. No dispersants were integrated in any of the K=4 groups, although two female individuals were identified as ‘dispersants’ originating from the Portuguese and Galician populations, but they could not be geographically attributed to a given destination group. The maximum number of dispersants identified per group for K=11 was one, and only five out of the eleven groups had such an immigrant. In total, we only identified seven dispersants (two females, five males) at K=11 (3.2% out of 218 individuals in our dataset). Two of them were not integrated in any clear destination group generated with our dataset. Dispersants originated from the populations of E and W Trás-os-Montes, W, C and SE Asturias, and W Galicia (Tables 2-1 & 2-2).

Combining our genetic results from spatial behavioral data from 85 radio- and GPS-collared wolves, we find a remarkable geographical overlap (Fig. 2-4). The majority (97.6%) of individual wolves' MCPs, defined by all recorded locations during the study

period, did not overlap more than one MCP_{total} as defined from the genetic dataset. The two wolves overlapping with more than one genetic group did it with two groups.

Population diversity and differentiation

A total of three and eight individuals were excluded at K=4 and K=11, respectively, for genetic population analyses because they could not be attributed to a particular geographical group. Due to the low number of dispersants in our dataset, genetic diversity values did not change substantially when considering either geographical groups or genetic clusters, both at K=4 and K=11 (Tables 2-3 & 2-4). For K=4, genetic diversity was similar between all groups, as measured by the mean expected heterozygosity (H_e), varying between 0.57 and 0.62 (average 0.59) (Table 2-3). H_e values were more variable between populations at K=11, varying between 0.48 and 0.59 for geographical groups (average 0.55) (table 2-4). Lowest H_e values were observed in E Asturias and S Douro populations (0.48); the latter also presenting the lowest allelic richness. However, the S Douro population had a higher private allelic richness (0.15) than the remaining populations (≤0.11). On the other hand, populations from E Trás-os-Montes and E Galicia showed the highest H_e and allelic richness values (Table 2-4).

We did not detect signs of inbreeding in the different groups, since observed and expected heterozygosities were similar, leading to F values close to zero (Tables 2-3 & 2-4). At K=4, populations did not appear to represent close familiar groups, as relatedness values were always lower than 0.2. However, mean relatedness values were slightly higher for K=11, with some populations presenting values that would be expected for half-siblings (0.25): S Douro (0.38), E Asturias (0.25-0.32), Alto Minho (0.25-0.28) (Table 2-4).

At K=4, all groups appeared similarly differentiated, as measured by pairwise F_{ST} (0.06-0.11) (Table 2-5). For K=11 values were more variable (0.03-0.25), the highest differentiation being observed between S Douro and E Asturias, and the lowest between E and W Trás-os-Montes (Table 2-6). On average, the S Douro group showed the highest differentiation with all other groups (average F_{ST}: 0.19), while E and W Trás-os-Montes showed the lowest (average F_{ST}: 0.09). Results with Jost's D were similar (Supplementary Tables 4 & 5).

Table 2-1: Statistics for geographical populations at K=4.

	Geographical populations			
	Asturias	**Portugal**	**Castilla y León**	**Galicia**
number of individuals				
N	52	47	55	59
inside MCP_p (% of total)	49 (94%)	44 (94%)	49 (89%)	53 (90%)
mean membership proportions to genetic clusters				
Asturias	0.89	0.04	0.07	0.03
Portugal	0.04	0.85	0.08	0.03
Castilla y León	0.06	0.04	0.83	0.03
Galicia	0.02	0.07	0.02	0.91
number of populations contributing >5%	1	1	2	0
number of admixed individuals				
$N_{admixed}$ (% of total)	6 (12%)	12 (28%)	19 (35%)	10 (17%)
inside MCPp (% of admixed)	3 (50%)	10 (83%)	13 (68%)	4 (40%)
number of dispersants from other populations (%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
closest population	Castilla y León	Castilla y León	Asturias	Asturias
approx. distance to closest population (km)	2	21	2	14
area (km^2)				
MCPt	7457	17417	45488	15712
MCPp	6828	15164	37467	11642
MCPp / MCPt (%)	92%	87%	82%	74%

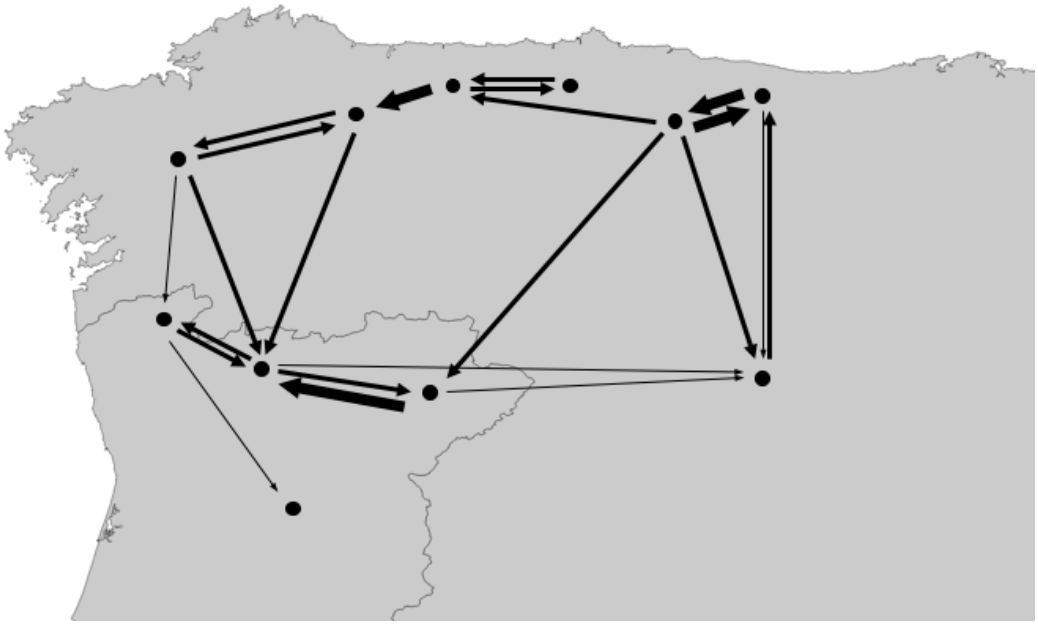


Fig. 2-3: Mean genetic contributions between pairs of subpopulations at K=11. The mean genetic proportion of subpopulation x in subpopulation y (Table 2-2) is represented as an arrow from x to y. Sizes of arrows represent different levels of mean membership proportions: small: 2.5-5%, medium: 5-10%, large: >10%.

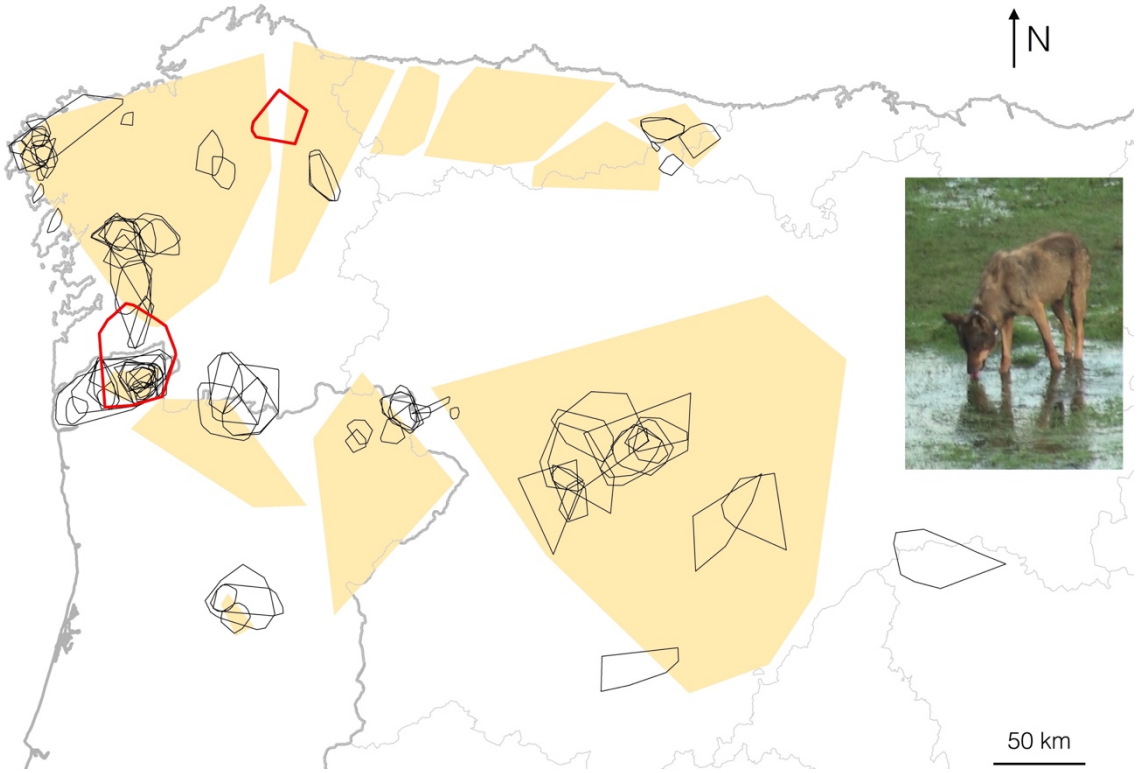


Fig. 2-4: MCPs from tracking data of radio-collared wolves (n=34) (black) and MCPt identified from our molecular and geographical analyses (gray).

Table 2-2: Statistics for geographical populations at K=11

	Geographical populations										
	Alto Minho	**E Trás-os-Montes**	**SE Asturias**	**W Asturias**	**Castilla y León**	**S Douro**	**W Galicia**	**C Asturias**	**E Galicia**	**W Trás-os-Montes**	**E Asturias**
number of individuals											
N	13	17	21	12	14	6	53	30	11	11	19
inside MCP_p (% of total)	12 (92%)	15 (88%)	16 (76%)	8 (67%)	9 (64%)	6 (100%)	46 (87%)	25 (83%)	9 (82%)	6 (55%)	17 (89%)
mean membership proportion to genetic clusters											
Alto Minho	0.82	0.01	0.00	0.00	0.00	0.03	0.01	0.00	0.02	0.08	0.00
E Trás-os-Montes	0.02	0.73	0.01	0.01	0.04	0.01	0.01	0.01	0.02	0.11	0.02
SE Asturias	0.01	0.08	0.77	0.08	0.06	0.00	0.01	0.01	0.01	0.01	0.10
W Asturias	0.00	0.01	0.01	0.77	0.01	0.00	0.02	0.07	0.10	0.02	0.01
Castilla y León	0.01	0.01	0.02	0.00	0.78	0.01	0.01	0.01	0.01	0.01	0.07
S Douro	0.00	0.02	0.00	0.00	0.01	0.91	0.00	0.01	0.00	0.01	0.00
W Galicia	0.04	0.01	0.00	0.00	0.01	0.01	0.84	0.00	0.05	0.04	0.00
C Asturias	0.00	0.01	0.03	0.09	0.01	0.00	0.01	0.85	0.01	0.02	0.02
E Galicia	0.01	0.02	0.01	0.01	0.01	0.01	0.06	0.01	0.76	0.05	0.02
W Trás-os-Montes	0.08	0.09	0.02	0.01	0.04	0.01	0.01	0.00	0.01	0.66	0.01
E Asturias	0.00	0.01	0.12	0.02	0.03	0.00	0.00	0.02	0.01	0.00	0.76
number of populations contributing >5%	1	2	1	2	1	0	1	1	2	2	2

Table 2-2 (continued)

	Geographical populations										
	Alto Minho	**E Trás-os-Montes**	**SE Asturias**	**W Asturias**	**Castilla y León**	**S Douro**	**W Galicia**	**C Asturias**	**E Galicia**	**W Trás-os-Montes**	**E Asturias**
number of admixed individuals											
$N_{admixed}$ (% of total)	2 (15%)	10 (59%)	8 (38%)	4 (33%)	5 (36%)	1 (17%)	17 (32%)	9 (30%)	4 (36%)	5 (45%)	7 (37%)
inside MCPp (% of admixed)	1 (50%)	5 (50%)	3 (37.5%)	1 (25%)	0 (0%)	1 (100%)	10 (59%)	4 (44.4%)	2 (50%)	3 (60%)	5 (71.4%)
number of dispersants from other populations (%)	1 (8%)	0 (0%)	0 (0%)	1 (8%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (9%)	1 (9%)	1 (5%)
closest pop	W Trás-os-Montes	W Trás-os-Montes	E Asturias	C Asturias	E Trás-os-Montes	E Trás-os-Montes	E Galicia	SE Asturias	W Galicia	Alto Minho	SE Asturias
approx. distance to closest pop (km)	5.5	10	1.5	4	21	57	6.5	2	6.5	5.5	1.5
area (km^2)											
MCPt	398	5135	1819	968	29071	176	11642	3501	4232	2710	859
MCPp	330	2584	627	676	14166	176	8770	2730	915	1595	600
MCPp / MCPt (%)	83	50	34	70	49	100	75	78	22	59	70

Table 2-3: Genetic diversity and relatedness of geographical populations and genetic clusters at K=4. N: sample size; Na: number of alleles; AR: allelic richness (rarefied to 46 samples); pAR: private allelic richness (rarefied to 46 samples); Ho: observed heterozygosity; He: expected heterozygosity; F: fixation index; r: relatedness (Lynch and Ritland, 1999)

Pop		N	Na	AR	pAR	Ho	He	F	r
Asturias	**pop**	50.11 ± 0.50	4.59 ± 0.23	4.44 ± 0.22	0.24 ± 0.07	0.55 ± 0.03	0.59 ± 0.03	0.06 ± 0.02	0.082 (0.072-0.088)
	cluster	44.52 ± 0.42	4.39 ± 0.24	4.09 ± 0.20	0.25 ± 0.06	0.54 ± 0.03	0.59 ± 0.03	0.07 ± 0.02	0.092 (0.084-0.100)
Portugal	**pop**	45.85 ± 0.27	5.13 ± 0.25	5.00 ± 0.24	0.50 ± 0.10	0.57 ± 0.02	0.62 ± 0.02	0.08 ± 0.02	0.056 (0.050-0.064)
	cluster	34.02 ± 0.22	4.85 ± 0.22	4.64 ± 0.20	0.55 ± 0.10	0.57 ± 0.02	0.61 ± 0.02	0.07 ± 0.02	0.079 (0.069-0.090)
Castilla y León	**pop**	52.13 ± 0.61	4.74 ± 0.27	4.56 ± 0.25	0.25 ± 0.07	0.53 ± 0.03	0.57 ± 0.03	0.07 ± 0.02	0.096 (0.090-0.104)
	cluster	33.22 ± 0.40	4.11 ± 0.24	3.92 ± 0.22	0.19 ± 0.05	0.52 ± 0.03	0.55 ± 0.03	0.06 ± 0.03	0.164 (0.151-0.176)
Galicia	**pop**	56.39 ± 0.50	4.85 ± 0.26	4.55 ± 0.23	0.32 ± 0.07	0.55 ± 0.03	0.57 ± 0.03	0.03 ± 0.01	0.092 (0.086-0.098)
	cluster	47.76 ± 0.47	4.57 ± 0.26	4.08 ± 0.21	0.31 ± 0.07	0.54 ± 0.03	0.55 ± 0.03	0.02 ± 0.02	0.110 (0.103-0.116)

Table 2-4 Genetic diversity and relatedness of geographical populations and genetic clusters at K=11. N: sample size; Na: number of alleles; Ne: number of effective alleles; AR: allelic richness (rarefied to 10 samples for the populations, and to 6 samples for the clusters); pAR: private allelic richness (rarefied to 10 samples for the populations, and to 6 samples for the clusters); Ho: observed heterozygosity; He: expected heterozygosity; F: fixation index; r: relatedness (Lynch and Ritland, 1999)

Pop		N	Na	AR	pAR	Ho	He	F	r Lynch & Ritland (1999)
Alto Minho	**pop**	12.78 ± 0.14	3.67 ± 0.16	2.93 ± 0.11	0.10 ± 0.03	0.53 ± 0.03	0.52 ± 0.03	-0.01 ± 0.04	0.245 (0.214-0.279)
	cluster	9.80 ± 0.12	2.94 ± 0.13	2.31 ± 0.09	0.09 ± 0.03	0.50 ± 0.04	0.47 ± 0.03	-0.02 ± 0.04	0.31 (0.28-0.339)
W Trás-os-Montes	**pop**	10.78 ± 0.07	3.78 ± 0.16	3.22 ± 0.12	0.05 ± 0.02	0.57 ± 0.03	0.56 ± 0.03	-0.02 ± 0.03	0.074 (0.048-0.102)
	cluster	5.96 ± 0.03	3.02 ± 0.16	2.54 ± 0.10	0.06 ± 0.03	0.61 ± 0.04	0.51 ± 0.03	-0.21 ± 0.04	0.204 (0.119-0.291)
E Asturias	**pop**	17.98 ± 0.25	3.46 ± 0.18	2.73 ± 0.12	0.03 ± 0.02	0.51 ± 0.04	0.49 ± 0.03	-0.04 ± 0.03	0.254 (0.225-0.285)
	cluster	10.35 ± 0.18	2.44 ± 0.13	2.07 ± 0.10	0.03 ± 0.02	0.46 ± 0.04	0.40 ± 0.03	-0.14 ± 0.04	0.445 (0.397-0.494)
E Trás-os-Montes	**pop**	16.39 ± 0.15	4.26 ± 0.21	3.33 ± 0.13	0.11 ± 0.04	0.62 ± 0.03	0.59 ± 0.03	-0.05 ± 0.02	0.089 (0.07-0.11)
	cluster	7.65 ± 0.09	3.37 ± 0.16	2.59 ± 0.11	0.08 ± 0.03	0.61 ± 0.04	0.52 ± 0.03	-0.17 ± 0.03	0.179 (0.134-0.23)
SE Asturias	**pop**	20.46 ± 0.15	4.07 ± 0.22	3.10 ± 0.14	0.06 ± 0.02	0.56 ± 0.03	0.55 ± 0.03	-0.01 ± 0.02	0.12 (0.104-0.137)
	cluster	14.54 ± 0.11	3.72 ± 0.20	2.59 ± 0.11	0.03 ± 0.01	0.56 ± 0.04	0.53 ± 0.03	-0.06 ± 0.03	0.157 (0.131-0.183)
W Asturias	**pop**	11.67 ± 0.11	3.65 ± 0.17	3.00 ± 0.12	0.10 ± 0.04	0.58 ± 0.03	0.56 ± 0.02	-0.047 ± 0.03	0.151 (0.116-0.19)
	cluster	7.67 ± 0.11	3.00 ± 0.16	2.43 ± 0.09	0.10 ± 0.04	0.55 ± 0.04	0.51 ± 0.02	-0.09 ± 0.05	0.232 (0.191-0.275)

Table 2-4 (continued)

Pop		N	Na	AR	pAR	Ho	He	F	r Lynch & Ritland (1999)
Castilla y León	**pop**	12.78 ± 0.27	3.70 ± 0.22	3.07 ± 0.16	0.09 ± 0.03	0.51 ± 0.03	0.54 ± 0.03	0.05 ± 0.03	0.148 (0.124-0.176)
	cluster	8.22 ± 0.17	3.20 ± 0.21	2.51 ± 0.12	0.09 ± 0.02	0.51 ± 0.03	0.51 ± 0.03	0.00 ± 0.04	0.214 (0.177-0.251)
S Douro	**pop**	5.89 ± 0.05	2.87 ± 0.14	2.76 ± 0.13	0.15 ± 0.05	0.55 ± 0.04	0.48 ± 0.03	-0.14 ± 0.04	0.377 (0.303-0.455)
	cluster	4.89 ± 0.05	2.57 ± 0.10	2.31 ± 0.08	0.17 ± 0.05	0.55 ± 0.04	0.46 ± 0.02	-0.21 ± 0.04	0.399 (0.316-0.501)
W Galicia	**pop**	50.67 ± 0.50	4.67 ± 0.26	3.04 ± 0.12	0.09 ± 0.02	0.54 ± 0.03	0.56 ± 0.03	0.02 ± 0.02	0.104 (0.098-0.11)
	cluster	35.26 ± 0.36	4.09 ± 0.23	2.51 ± 0.09	0.09 ± 0.03	0.53 ± 0.03	0.54 ± 0.03	0.02 ± 0.02	0.131 (0.121-0.14)
C Asturias	**pop**	29.01 ± 0.30	3.89 ± 0.20	2.95 ± 0.12	0.04 ± 0.01	0.53 ± 0.03	0.56 ± 0.03	0.03 ± 0.03	0.16 (0.147-0.174)
	cluster	21.24 ± 0.24	3.46 ± 0.18	2.46 ± 0.09	0.04 ± 0.01	0.50 ± 0.03	0.53 ± 0.03	0.03 ± 0.03	0.198 (0.177-0.217)
E Galicia	**pop**	10.13 ± 0.20	3.94 ± 0.21	3.31 ± 0.14	0.07 ± 0.02	0.62 ± 0.04	0.59 ± 0.03	-0.06 ± 0.03	0.106 (0.073-0.147)
	cluster	6.44 ± 0.15	3.17 ± 0.16	2.61 ± 0.11	0.06 ± 0.02	0.62 ± 0.4	0.53 ± 0.04	-0.17 ± 0.04	0.198 (0.128-0.275)

Table 2-5: Pairwise F_{ST} matrix between geographical populations (below diagonal) and genetic clusters (above diagonal) at K=4. All values were statistically significant based on 1000 permutations ($p<0.01$)

	Asturias	Portugal	Castilla y León	Galicia
Asturias	--	0.08	0.10	0.10
Portugal	0.07	--	0.12	0.08
Castilla y León	0.07	0.08	--	0.14
Galicia	0.09	0.06	0.11	--

Table 2-6: Pairwise F_{ST} matrix between geographical populations (below diagonal) and genetic clusters (above diagonal) at K=11. All values were statistically significant based on 1000 permutations ($p<0.05$)

	Alto Minho	E Trás-os-Montes	SE Asturias	W Asturias	Castilla Y León	S Douro	W Galicia	C Asturias	E Galicia	W Trás-os-Montes	E Asturias
Alto Minho	--	0.22	0.21	0.22	0.26	0.27	0.18	0.22	0.18	0.15	0.32
E Trás-os-Montes	0.13	--	0.14	0.16	0.16	0.20	0.13	0.17	0.11	0.13	0.26
SE Asturias	0.17	0.08	--	0.16	0.12	0.21	0.13	0.12	0.15	0.14	0.15
W Asturias	0.16	0.07	0.10	--	0.20	0.26	0.17	0.17	0.11	0.18	0.27
Castilla y León	0.20	0.10	0.08	0.14	--	0.23	0.17	0.18	0.17	0.19	0.21
S Douro	0.22	0.14	0.18	0.19	0.19	--	0.22	0.21	0.21	0.22	0.33
W Galicia	0.14	0.08	0.11	0.12	0.13	0.19	--	0.15	0.09	0.12	0.23
C Asturias	0.17	0.10	0.09	0.08	0.14	0.19	0.12	--	0.13	0.14	0.20
E Galicia	0.11	0.07	0.11	0.06	0.13	0.16	0.07	0.11	--	0.14	0.25
W Trás-os-Montes	0.07	0.03	0.09	0.08	0.12	0.17	0.07	0.09	0.06	--	0.28
E Asturias	0.23	0.14	0.07	0.15	0.12	0.25	0.16	0.13	0.16	0.16	--

Discussion

In this study we investigated the patterns of gene flow and the genetic structure of the Iberian wolf population. The extensive sampling and the integration of genetic and individual tracking data allowed us to capture the cryptic genetic structure in this population at different levels. Four main genetic groups can be identified in the Iberian wolf population, corresponding to the regions of Northern Portugal, Galicia, the Cantabrian mountain range and Castilla y Léon. At this level, there are no apparent large-scale topographical barriers nor gaps in the distribution that would provide a straightforward explanation for this partition. If this genetic structure is explored further, up to 11 geographically coherent groups can be recognized. At this scale, the size and genetic diversity of the groups is somewhat more variable. For instance, while four subpopulations can be distinguished along the Cantabrian mountain range, most of the regions of Galicia and Castilla y León seem to be occupied by animals from the one genetic group.

These groups appear to be characterized by very low levels of admixture and moderate to high genetic differentiation. Furthermore, the number of individuals identified as dispersants was also very low (3% of our sampled individuals). Combining these results with movement information recorded from 85 tracked individuals, we observe a remarkable overlap between the identified geographic/genetic clusters and the dispersal areas of individual wolves. Individual movement data is available for 8 of the 11 clusters, and in all of them except two, individuals do not cross more than one of the MCP_{total} areas defined by the genetic data. These results suggest that wolves present a very low mobility and very rarely disperse outside the general area where they were born.

Analyzing in detail the genetic contribution of each subpopulation to every other in a geographical context, there appear to be two main 'routes' through which gene flow is higher: in the North, along the Cantabrian mountain range, and to the South, along the regions North of the Douro river in Portugal. The isolation of the population from S Douro is also already evident from this analysis. In general, the identified genetic groups at K=11 are characterized by low levels of gene flow, which is reflected both in their genetic differentiation (F_{ST}) and the low proportion of miscegenation between groups.

Our results are consistent with other studies that show that wolf populations can be characterized by relatively low levels of gene flow and short dispersal distances, despite the potentially high dispersal capabilities of the species (Scandura et al. 2011; Jansson et al. 2012; Stronen et al. 2012). On a regional scale, this can result in a

metapopulation-like dynamic, with substantial genetic differentiation between regions (Jansson et al. 2012). In the case of the Iberian wolf population, our results show a remarkably reticulated population structure in a relatively small area and at such a fine scale that is not comparable with previous reports for other populations. Genetically differentiated wolf subpopulations have been described at continental scales, where habitats are more heterogeneous and gaps in the distribution exist, allowing for substantial genetic drift and adaptation to specific ecological conditions (e.g. Carmichael et al. 2007; VonHoldt et al. 2011; Stronen et al. 2013; Pilot et al. 2014). However, at local scales the number of described subpopulations is lower (e.g. Hindrikson et al. 2013; Stronen et al. 2013; Fabbri et al. 2014).

The Iberian wolf population has been affected by centuries of direct persecution and anthropogenic habitat changes, which has led to its severe decline and fragmentation. While at K=4 the identified groups present similar levels of genetic diversity, a larger variability exists at K=11. Larger and more diverse populations at this finer level might represent either areas where wolf abundance was consistently high even during the population minimum, such as the Galician populations (Valverde 1971; Garzón 1979), or a population that has been expanding rapidly in recent years, such as Castilla y León. The two different scales of our analysis also reveals that the wolves occupying the Castilian plateau, reaching the Madrid province, result from a population expansion from the Southern Cantabrian mountains (i.e. it merges with the E and SE Asturias populations at K=4), which is known to have served as a refuge during the minimum population levels in the 1970s (Valverde 1971). The genetic identity of some of the smaller subpopulations might be explained by the isolation imposed from this fragmentation, which then might have been reinforced by local geographical obstacles to dispersal. As an extreme example of this, the population of S Douro is mostly isolated from the Northern populations by the Douro river, while populations in every other direction have been extirpated. As a consequence, this population appears as genetically unique, with exceptionally low genetic diversity, the highest private allelic richness, and the highest genetic differentiation.

The genetic structure of a population results from a complex interaction between ecological traits, geographical features affecting gene flow and historical events. Given the current lack of an extensive sampling from before the decline and fragmentation of the population, it is difficult to assess from extant genetic data how much of the observed genetic population structure results from the anthropogenic impacts of the last centuries and the recent expansion, or if it reflects a more long-term trait of the Iberian wolf population, determined by ecological or geographical features. It has been described

that the tendency for wolves to disperse larger distances seems to be more common in situations of intense resource competition (Fuller et al. 2003). However, the persistence and recovery of Iberian wolves is in large part due to their reliance on livestock, which is relatively abundant and available, possibly reducing trophic competition (Llaneza et al. 2012; López-Bao et al. 2013; Llaneza and López-Bao 2015). Also, even at its minimum, the Iberian wolf population has maintained relatively large numbers, at least in its core areas: the population minimum is estimated at ca. 500 individuals in the 1970s (Valverde 1971; Garzón 1979). These factors could explain the apparent low dispersal of Iberian wolves. Clarifying the environmental constraints that may also contribute to the observed structure could benefit from further integration of direct tracking of individual animals and correlations with environmental conditions (climate, habitat, diet, human occupation, etc.). Knowledge of the factors determining genetic population structure and patterns of gene flow are important for the management and conservation of the Iberian wolf population.

Acknowledgements

Trapping took place under permits 338/2007/CAPT, 258/2008/CAPT, 286/2008/CAPT, 260/2009/CAPT, 332/ 2010/MANU, 333/2010/CAPT, 336/2010/MANU, 26/2012/ MANU, and 72/2014/CAPT (Portugal) and XXXXXXXXX-PNPE and 19/2006, 71/2009 and 86/2011 —Xunta de Galicia (Spain).

References

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: Conference ‘‘Frontiers of Wolf Recovery: Southwestern US and the World. p. 76–77.

Andreasen AM, Stewart KM, Longland WS, Beckmann JP, Forister ML. 2012. Identification of source-sink dynamics in mountain lions of the Great Basin. *Molecular Ecology* 21:5689–5701.

Aspi J, Roininen E, Kojola I, Ruokonen M, Vilà C. 2006. Genetic diversity, population structure, effective population size and demographic history of the Finnish wolf population. *Molecular Ecology* 15:1561–1576.

Blanco JC, Cortés Y. 2001. Ecología, Censos, Percepción y Evolución del Lobo en España: Análisis de un Conflicto. Málaga: SECEM.

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

Blanco JC, Cortés Y. 2012. Surveying wolves without snow: A critical review of the methods used in Spain. *Hystrix* 23:35–48.

Boulet M, Couturier S, Côté SD, Otto RD, Bernatchez L. 2007. Integrative use of spatial, genetic, and demographic analyses for investigating genetic connectivity between migratory, montane, and sedentary caribou herds. *Molecular Ecology* 16:4223–4240.

Broquet T, Ray N, Petit E, Fryxell JM, Burel F. 2006. Genetic isolation by distance and landscape connectivity in the American marten (Martes americana). *Landscape Ecology* 21:877–889.

Carmichael LE, Krizan J, Nagy J a., Fuglei E, Dumond M, Johnson D, Veitch a., Berteaux D, Strobeck C. 2007. Historical and ecological determinants of genetic structure in arctic canids. *Molecular Ecology* 16:3466–3483.

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

Coulon A, Guillot G, Cosson JF, Angibault JMA, Aulagnier S, Cargnelutti B, Galan M, Hewison AJM. 2006. Genetic structure is influenced by landscape features: Empirical evidence from a roe deer population. *Molecular Ecology* 15:1669–1679.

Dalén L, Fuglei E, Hersteinsson P, Kapel CMO, Roth JD, Samelius G, Tannerfeldt M, Angerbjörn A. 2005. Population history and genetic structure of the Arctic fox: a circumpolar species. *Biological Journal of the Linnean Society* 84:79–89.

Davis JM, Stamps JA. 2004. The effect of natal experience on habitat preferences. *Trends in Ecology and Evolution* 19:411–416.

Earl DA, vonHoldt BM. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* 4:359–361.

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. *Mammalian Biology* 76:217–221.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology* 14:2611–2620.

Excoffier L, Smouse PE, Quattro JM. 1992. Analysis of Molecular Variance Inferred From Metric Distances Among DNA Haplotypes: Application to Human Mitochondrial DNA Restriction Data. *Genetics* 131:479–491.

Fabbri E, Caniglia R, Kusak J, Galov A, Gomerčić T, Arbanasić H, Huber D, Randi E. 2014. Genetic structure of expanding wolf (Canis lupus) populations in Italy and Croatia, and the early steps of the recolonization of the Eastern Alps. *Mammalian Biology* 79:138–148.

Fuller TK, Mech LD, Cochrane JF. 2003. Wolf population dynamics. In: Mech LD, Boitani L, editors. Wolves: Behavior, Ecology, and Conservation. The University of Chicago Press. p. 161–191.

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

Gipson P, Ballard W, Nowak R, Mech L. 2000. Accuracy and Precision of Estimating Age of Gray Wolves by Tooth Wear. *Journal of Wildlife Management* 64:752–758.

Godinho R, Llaneza L, Blanco JC, Lopes S, Alvares F, García EJ, Palacios V, Cortés Y, Talegón J, Ferrand N, et al. 2011. Genetic evidence for multiple events of hybridization between wolves and domestic dogs in the Iberian Peninsula. *Molecular ecology* 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 non-invasive samples with ancestry informative markers. *Molecular Ecology Resources*:317–328.

Grilo C, Moço G, Cândido A, Alexandre S, Petrucci-Fonseca F. 2002. Bases para a definição de corredores ecológicos na conservação de uma população marginal e fragmentada: o caso da população lupina a sul do rio Douro – 1ª Fase. Relatório Técnico PRAXIS XXI.

Hindrikson M, Remm J, Männil P, Ozolins J, Tammeleht E, Saarma U. 2013. Spatial Genetic Analyses Reveal Cryptic Population Structure and Migration Patterns in a Continuously Harvested Grey Wolf (Canis lupus) Population in North-Eastern Europe. *PLoS ONE* 8:e75765.

Jansson E, Ruokonen M, Kojola I, Aspi J. 2012. Rise and fall of a wolf population: genetic diversity and structure during recovery, rapid expansion and drastic decline. *Molecular Ecology* 21:5178–5193.

Jost L. 2008. GST and its relatives do not measure differentiation. *Molecular Ecology* 17:4015–4026.

Llaneza L, López-Bao J V. 2015. Indirect effects of changes in environmental and agricultural policies on the diet of wolves. *European Journal of Wildlife Research* 61:895–902.

Llaneza L, López-Bao J V., Sazatornil V. 2012. Insights into wolf presence in human-dominated landscapes: The relative role of food availability, humans and landscape attributes. *Diversity and Distributions* 18:459–469.

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. *Biodiversity and Conservation* 24:2105–2108.

López-Bao JV, Sazatornil V, Llaneza L, Rodríguez A. 2013. Indirect Effects on Heathland Conservation and Wolf Persistence of Contradictory Policies that Threaten Traditional Free-Ranging Horse Husbandry. *Conservation Letters* 6:448–455.

Lynch M, Ritland K. 1999. Estimation of pairwise relatedness with molecular markers. *Genetics* 152:1753–1766.

Manel S, Schwartz MK, Luikart G, Taberlet P. 2003. Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution* 18:189–197.

McRae BH, Beier P, Dewald LE, Huynh LY, Keim P. 2005. Habitat barriers limit gene flow and illuminate historical events in a wide-ranging carnivore, the American puma. *Molecular Ecology* 14:1965–1977.

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.

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

Meirmans PG, Van Tienderen PH. 2004. GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes* 4:792–794.

Moreira L. 1992. Contribuição para o estudo da ecologia do lobo (Canis lupus signatus Cabrera 1907) no Parque Natural de Montesinho. Relatório de estágio.

Newsome TM, Boitani L, Chapron G, Ciucci P, Dickman CR, Dellinger JA, L??pez-Bao J V., Peterson RO, Shores CR, Wirsing AJ, et al. 2016. Food habits of the world's grey wolves. *Mammal Review*:1–15.

Peakall R, Smouse PE. 2012. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research--an update. *Bioinformatics* 28:2537–2539.

Pereira M, Fonseca F, Magalhães C. 1985. Wolf ecology in Portugal. In: Symposium Predateurs, Lisbonne 29/31.3.1985. Lisboa. p. 122–167.

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

Pilot M, Jędrzejewski W, Branicki W, Sidorovich VE, Jędrzejewska B, Stachura K, Funk SM. 2006. Ecological factors influence population genetic structure of European grey wolves. *Molecular Ecology* 15:4533–53.

Pimenta V. 1998. Estudo comparativo de duas alcateias no nordeste do distrito de Bragança: Utilização do espaço e do tempo e hábitos alimentares. Relatório de estágio.

Pimenta V, Barroso I, Álvares F, Correia J, Ferrão 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. Lisboa: Instituto da Conservação da Natureza / Grupo Lobo.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945–959.

QGIS Development Team. 2015. QGIS Geographic Information System. *Open Source Geospatial Foundation Project*.

Queller DC, Goodnight KF. 1989. Estimating Relatedness Using Genetic Markers. *Evolution* 43:258–275.

Ražen N, Brugnoli A, Castagna C, Groff C, Kaczensky P, Kljun F, Knauer F, Kos I, Krofel M, Luštrik R, et al. 2016. Long-distance dispersal connects Dinaric-Balkan and Alpine grey wolf (Canis lupus) populations. *European Journal of Wildlife Research*:137–142.

Rice WR. 1989. Analyzing Tables of Statistical Tests. *Evolution* 43:223–225.

Roque S, Godinho R, Cadete D, Pinto S, Pedro A, Bernardo J, Petrucci-Fonseca F, Álvares F. 2011. Plano de Monitorização do Lobo Ibérico nas áreas dos Projetos Eólicos das Serras de Montemuro, Freita, Arada e Leomil – Ano IV e Análise Integrativa dos Resultados (2006–2011). Relatório Final.

Rousset F. 2008. genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources* 8:103–106.

Row JR, Gomez C, Koen EL, Bowman J, Murray DL, Wilson PJ. 2012. Dispersal promotes high gene flow among Canada lynx populations across mainland North America. *Conservation Genetics* 13:1259–1268.

Rueness EK, Jorde PE, Hellborg L, Stenseth NC, Ellegren H, Jakobsen KS. 2003. Cryptic population structure in a large, mobile mammalian predator: the Scandinavian lynx. *Molecular Ecology* 12:2623–2633.

Sacks BN, Brown SK, Ernest HB. 2004. Population structure of California coyotes corresponds to habitat-specific breaks and illuminates species history. *Molecular ecology* 13:1265–75.

Sacks BN, Mitchell BR, Williams CL, Ernest HB. 2005. Coyote movements and social structure along a cryptic population genetic subdivision. *Molecular Ecology* 14:1241–1249.

Scandura M, Iacolina L, Capitani C, Gazzola A, Mattioli L, Apollonio M. 2011. Fine-scale genetic structure suggests low levels of short-range gene flow in a wolf population of the Italian Apennines. *European Journal of Wildlife Research* 57:949–958.

Stronen A V., Forbes GJ, Paquet PC, Goulet G, Sallows T, Musiani M. 2012. Dispersal in a plain landscape: Short-distance genetic differentiation in southwestern Manitoba wolves, Canada. *Conservation Genetics* 13:359–371.

Stronen A V, Jędrzejewska B, Pertoldi C, Demontis D, Randi E, Niedziałkowska M, Pilot M, Sidorovich VE, Dykyy I, Kusak J, et al. 2013. North-South differentiation and a region of high diversity in European wolves (Canis lupus). *PloS one* 8:e76454.

Surridge AK, Ibrahim KM, Bell DJ, Webb NJ, Rico C, Hewitt GM. 1999. Fine-scale genetic structuring in a natural population of European wild rabbits (Oryctolagus cuniculus). *Molecular Ecology* 8:299–307.

Szpiech ZA, Jakobsson M, Rosenberg NA. 2008. ADZE: A rarefaction approach for counting alleles private to combinations of populations. *Bioinformatics* 24:2498–2504.

Taberlet P, Fumagalli L, Wust-Saucy A-G, Cosson JF. 1998. Comparative phylogeography and postglacial colonization routes in Europe. *Molecular ecology* 7:453–64.

Tammeleht E, Remm J, Korsten M, Davison J, Tumanov I, Saveljev a., Männil P, Kojola I, Saarma U. 2010. Genetic structure in large, continuous mammal populations: The example of brown bears in northwestern Eurasia. *Molecular Ecology* 19:5359–5370.

Teacher AG, Thomas JA, Barnes I. 2011. Modern and ancient red fox (Vulpes vulpes) in Europe show an unusual lack of geographical and temporal structuring, and differing responses within the carnivores to historical climatic change. *BMC Evolutionary Biology* 11:214.

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

VonHoldt BM, Pollinger JP, Earl D a., Knowles JC, Boyko AR, Parker H, Geffen E, Pilot M, Jędrzejewski W, Jędrzejewska B, et al. 2011. A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. *Genome Research* 21:1294–1305.

vonHoldt BM, Stahler DR, Bangs EE, Smith DW, Jimenez MD, Mack CM, Niemeyer CC, Pollinger JP, Wayne RK. 2010. A novel assessment of population structure and gene flow in grey wolf populations of the Northern Rocky Mountains of the United States. *Molecular Ecology* 19:4412–4427.

Wabakken P, Sand H, Liberg O, Bjärvall A. 2001. The recovery, distribution, and population dynamics of wolves on the Scandinavian peninsula, 1978-1998. *Canadian Journal of Zoology* 79:710–725.

Waters JM, Fraser CI, Hewitt GM. 2013. Founder takes all: Density-dependent processes structure biodiversity. *Trends in Ecology and Evolution* 28:78–85.

Paper II - Historic Demography and Divergence of European Wolf Populations

Silva P et al.

(manuscript in preparation for submission)

Abstract

The gray wolf (*Canis lupus*) is one of the most widely distributed mammals, however little is known about its evolutionary history. Based on the analysis of eight canid genomes we examined the divergence between Eurasian wolf populations and their demographic history. Our results suggest that, while wolf abundance has been affected by the climatic oscillations of the Pleistocene, their differentiation into some of the genetic and morphologic partitions recognized today is relatively more recent. We estimate that two main clades of Old World wolves, representing classically recognized subspecies, diverged ca. 25 kya and subsequently experienced different demographic histories. We find that all Eurasian wolves suffered a drastic population bottleneck in the last 30 kya, which particularly affected wolves in Europe. We estimate that the events that led to the differentiation of currently isolated European wolf populations from the Iberian Peninsula, and Italy occurred at approximately at the same time, ca. 2.4-7.4 kya, confirming the long-term isolation of these populations that predates the extirpation of wolves in Central Europe in recent centuries, but are too recent to be directly associated with the end of the last glaciation.

Introduction

Gray wolves (*Canis lupus*) were once widely distributed across the Holarctic, occupying the roles of top predators in many ecosystems (Mech and Boitani 2003). Over the past two centuries, reductions in available habitat and natural prey, as well as direct human persecution, have resulted in their extinction in most of Central and Western Europe, and parts of North America (Ripple et al. 2014). In recent decades, due to the ability of wolves to disperse rapidly over long distances and the implementation of protection measures, wolves have recovered and even successfully reinvaded areas where they were previously extirpated (Chapron et al. 2014). In Central and Western Europe, several more or less isolated populations remain mainly in the southern peninsulas of the Balkans, Italy and Iberia, while larger and more interconnected populations exist in Eastern Europe that are in contact with populations from Russia and the rest of Asia (Chapron et al. 2014).

The worldwide distribution of wolves and their adaptation to diverse habitats has resulted in distinct morphologies, with several subspecies having been proposed based on morphometric data (Nowak 2003). According to some classifications, the widespread and medium-sized *Canis lupus lupus* is considered to occupy most of Eurasia, while *C. l. pallipes* is distributed in southern Eurasia, including the Middle East and southwestern Asia. In Europe, the smaller wolves of the Italian and Iberian peninsulas have also been proposed to constitute the subspecies *C. l. italicus* Altobello, 1921 and *C. l. signatus* Cabrera, 1907, respectively (Petrucci-Fonseca 1990; Vilà 1993; Nowak and Federoff 2002). It is still unclear however how much of this differentiation is due to historical restrictions of gene flow between global wolf populations or to adaptations to local environmental conditions, and not much is known about the timeframe in which these changes occurred.

Wolves probably originated in the Arctic regions of Asia in the early to middle Pleistocene, and subsequently expanded over Eurasia, reaching Europe ca. 0.8 Mya, but were not present in North America until the late Pleistocene, ca. 0.1 Mya (Nowak 2003; Wang and Tedford 2008). Wolves have therefore lived through the profound environmental changes of the cyclical Pleistocene glaciations and the transition to the Holocene that greatly affected the patterns of genetic diversity and differentiation of many species (Hewitt 2000; de Bruyn et al. 2011). While it is possible that these environmental transformations greatly affected wolves, they do not seem to have left strong genetic phylogeographic signatures as in many other species. A study of extant mtDNA variability in wolves has found relatively recent coalescent times and an absence

of any large-scale geographical structure, which has been hypothesized to be the result of repeated population contractions and rapid re-colonizations during the glacial cycles (Vilà et al. 1999). Genomic studies have also suggested that extant wolves in diverse areas of the world were recently derived from a bottlenecked population and then expanded to their worldwide distribution (Thalmann et al. 2013; Freedman et al. 2014; Fan et al. 2016). This does not mean however that there is a complete lack of genetic structure, since distinct genetic partitions, that might reflect adaptations to specific ecological conditions, can still be recognized based on genome-wide microsatellite and SNP data (Lucchini et al. 2004; VonHoldt et al. 2011; Pilot et al. 2014).

Wolves of the southern European peninsulas have mostly been isolated at least since the extinction of populations in Central Europe around the turn from the nineteenth to the twentieth century (Valière et al. 2003). Bayesian coalescent analyses of microsatellite data have suggested that Italian wolves suffered a severe decline over the last 2000 to 10,000 years, implying that this population might have been isolated for a much longer time, possibly since the end of the last glaciation (Lucchini et al. 2004). The distribution of extant mtDNA variability seems to support this hypothesis, given that Italian wolves possess a unique haplotype not found elsewhere, while a shared haplotype between Iberian and Eastern European wolves exists (Pilot et al. 2010). Also, divergence estimates based on genome-wide SNP data suggested that Iberian and Italian wolves separated at approximately the same time, 5600-3200 years ago (Pilot et al. 2014). On the other hand, wolves from Eastern Europe appear to have maintained much larger effective population sizes and connectivity (Pilot et al. 2014), non-withstanding some level of population structure and/or local bottlenecks (Pilot et al. 2006; Pilot et al. 2014).

In this study, we leverage the availability of full genome data from worldwide canids (Freedman et al. 2014; Fan et al. 2016) to further investigate the demographic history of European wolves, including the timing of their divergence, levels of gene flow and long-term effective population sizes. We interpret our findings in the light of recorded declines in recent centuries and older environmental conditions of the Pleistocene. We use recently developed and accurate demography inference methods capable of integrating information from many unlinked genomic segments, each of them representing a sample from the evolutionary process, and therefore informative about past population parameters (Dutheil and Hobolth 2012).

Materials and Methods

Canid genome data

We compiled a dataset of full genome sequences from eight canids at 12-26x coverage (table 2-7). Five canids were sequenced by Freedman et al. (2014): two wolves, from Croatia and Israel, dogs of the dingo and basenji breeds, and one golden jackal (*Canis aureus*) from Israel as an outgroup. In addition, we included three European wolves from Fan et al. (2016): two individuals from the Iberian Peninsula and one from Italy. Details regarding read alignment, genotyping and quality filtering procedures can be found in the original studies.

Table 2-7: Canid genomes used in this study

sample	region of origin	average genome coverage (X)	reference
Portuguese wolf	Minho, Portugal	26.1	Fan et al., 2016
Spanish wolf	Castilla y Léon, Spain	25.29	Fan et al., 2016
Italian wolf	Calabria, Italy	13.01	Fan et al., 2016
Croatian wolf	Perković, Croatia	25.3x	Freedman et al., 2014
Israeli wolf	Neve Ativ, Golan Heights, Israel	21.6x	Freedman et al., 2014
basenji	Maryland, USA	12.6x	Freedman et al., 2014
dingo	Bargo Dingo Sanctuary, Australia	25.75x	Freedman et al., 2014
golden jackal	Tel Aviv, Israel	23.8x	Freedman et al., 2014

Estimating population divergence times, effective sizes and migration

The Generalized Phylogenetic Coalescent Sampler (*G-PhoCS*) developed by Gronau et al. (2011) performs demographic inferences on genomic sequence data given information on the population phylogeny and which population pairs are likely to have experienced post-divergence gene flow (implemented as 'migration bands' that allow different rates for the two directions of gene flow). For the given phylogeny, *G-PhoCS* infers population sizes, population divergence times, and migration rates.

The demographic parameters are estimated based on inferred genealogies for thousands of neutrally evolving loci; therefore, a set of high-quality genomic regions that ideally have not been subject to strong selection are needed. We used the same putatively neutral regions defined in Freedman et al. (2014, Supplementary Text 9.2.1). Briefly, we excluded regions of the genome with assembly gaps, repeats, low mappability, missing data in all samples (i.e. bases not passing quality filters), coding regions (and respective 10kb flanking regions) and regions that are highly conserved in mammals. In total, 13,696 1kb regions of the genome fulfilled these criteria. For these regions, we constructed multiple sequence alignments for our samples in addition to the boxer reference genome (canFam3.1). Individual positions failing quality filters were

masked as N's; additionally, all 'CG' dinucleotides, as well as all positions with a C* dinucleotide in one genome and *G in another were also masked.

The assumed population phylogeny for *G-PhoCS* is represented in figure 2-4, according to the Maximum Likelihood phylogeny inferred by Fan et al. (2016). All MCMC runs were executed using the same default settings as in Gronau *et al*. (2011) and Freedman *et al.* (2014) (Supplementary Text 9.2.2 therein). The program was allowed to run for 500,000 iterations, the first 200,000 of which were discarded as burn-in. Convergence was inspected manually for each run using *Tracer* (Rambaut et al. 2014).

Given the limitations of the method when a large number of migration bands are used, we followed the same procedure as Freedman et al. 2014 (Supplementary Text 9.3 therein) in taking a two-step approach to infer gene flow in our data. First, we identified which population pairs showed signs of significant gene flow in two separate runs using the same settings as in the final run but using only a subset of 5,000 of the loci: in one run we considered migration bands between geographically adjacent wolf populations (Portuguese wolf <-> Spanish wolf, ancestral Iberian wolves <-> ancestral Italian wolves, Italian wolf <-> Croatian wolf, Croatian wolf <-> Israeli wolf) and in the other run, between every wolf population and the boxer. A migration band was inferred to have significant gene flow if the 95% Bayesian credible interval of the total migration rate for that band did not include 0, or if the total migration rate was estimated to be greater than 0.03 with posterior probability greater than 50%. We then performed a final run considering only the migration bands with significant gene flow as well as migration bands in the opposite direction.

Parameters estimates θ and τ given by *G-PhoCS* are scaled by the mutation rate μ; effective population sizes (N_e, in number of individuals) can be calculated by $\theta=4N_e\mu$, and divergence times (T, in years) by $\tau=T\mu/g$, where g is the average generation time (in years). For these conversions, we assumed a 3 year generation time (Mech and Boitani 2003), and two distinct mutation rates, given the considerable uncertainty of this rate for canids. We interpret our results using the rate of 0.4×10^{-8} mutations/bp/generation recently estimated based on the comparison of an ancient Asian wolf genome with modern wolves (Skoglund et al. 2015), but also report time estimates of key events with the previously used rate of 1×10^{-8} mutations/bp/generation (Lindblad-Toh et al. 2005; Freedman et al. 2014). The *G-PhoCS* model also uses a scaled version of migration rate, $M=m/\mu$, where m is the probability of migration across a given band in a single generation. The level of gene flow across a given migration band is measured by the total migration rate, which is the migration rate scaled by the time span of the migration

band (τ_m): $m_{tot} = M\tau_m$. The time span of a migration band is defined using the start and end times of the two populations involved.

To confirm the reliability of the inferred divergence times between the Portuguese and the Spanish wolves, both of the Iberian wolf population, (see Results), we reran the *G-PhoCS* analyses with simulated data. For all the simulations, we assumed a recombination rate of 0.92×10^{-8} per generation, based on the mean recombination rate estimated by Wong et al. (2010) from a linkage map of the domestic dog genome, constructed from microsatellite data. *ms* (Hudson 2002) was used to simulate gene trees at 15,000 loci under the parameters inferred by *G-PhoCS* in the analyses described above (Supplementary Information (SI), command line 1); *seq-gen* (Rambaut and Grassly 1997) was then used to build 1kb sequences for each of those loci (see SI command line 2 for example). We assumed the simulated loci evolved under the JC69 model. The alignments produced in this manner were then used as input for *G-PhoCS* using the same settings as described above for the main analyses. In addition to this control simulation where all parameters corresponded to the ones inferred by *G-PhoCS*, we performed two simulations in which 1) the divergence time between the Portuguese and the Spanish wolves was assumed to be 90% smaller than inferred in the main analysis (SI command line 3), and 2) complete gene flow (full panmixia) exists until this day between the Portuguese and Spanish populations (SI command line 4).

Results

Divergence times, long-term effective population sizes and migration rates estimates inferred by *G-PhoCS* for the demographic history of European wolf populations are summarized in figure 2-4 and Supplementary Table S1. Assuming a mutation rate of 0.4×10^{-8}, the divergence of Middle Eastern wolf populations, represented in this analysis by the Israeli wolf, appears to have been established around the same time as the inferred wolf/dog divergence, ca. 25 kya. For the remaining European wolf populations, split times are estimated as much more recent, and very close together: 7.4 kya for Croatian wolves, and 7 kya for Italian/Iberian wolves. The two samples from the Iberian Peninsula are estimated to have a divergence time of ca. 6 thousand years.

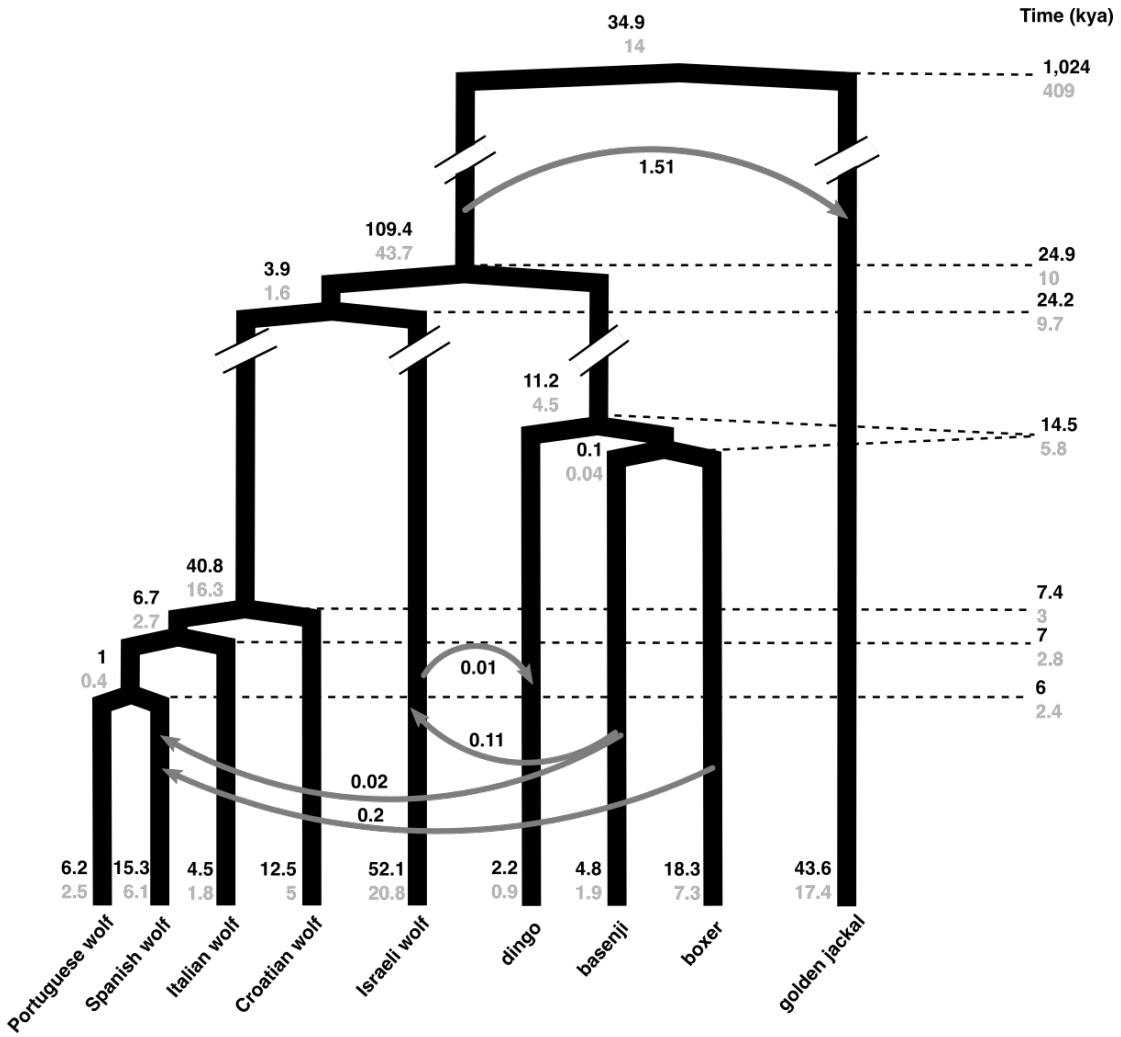


Fig. 2-5: Demographic model and parameter estimates tested in *G-PhoCS*. Values to the left of branches are inferred effective population sizes (in thousands of individuals); arrows represent gene flow; values to the right of the diagram are divergence times (in thousands of years). Times and population sizes were converted using an average mutation rate of 0.4×10^{-8}; lower values in gray are the same estimates converted using a rate of 1×10^{-8}.

The unexpectedly high divergence times inferred between the Portuguese and Spanish wolves prompted us to perform simulations to rule out that this estimate results from a limitation of the method. However, estimates based on simulated data with 1) lower divergence, or 2) panmixia between these two wolves show that *G-PhoCS* would accurately infer the split times in those scenarios (Supplementary Table S2, Supplementary Figure S1). That is to say, *G-PhoCS* appears to not be limited in making inferences from recently diverged genomes, and the obtained old divergence estimate results from the actual information in the data.

The ancestral wolf population of all sequences sampled in this study is estimated to have an effective size of ~4,000 individuals, while the population leading to European wolves was much larger (~41,000 individuals). Italian and Iberian wolves seem to have originated from much smaller populations (~6,700 and ~1,000 individuals, respectively). Since their divergence, Middle Eastern wolves appear to have maintained larger

population sizes (in the order of ~52,000 individuals) than European populations (Italian, 4,500; Portuguese, 6,200 thousand; Spanish, 15,000). We also detected significant gene flow from basenji to Israeli wolf (~11%), which had been described for the same samples in (Freedman et al. 2014), and from boxer (~20%) and basenji (~2%) to the Spanish wolf. No significant gene flow was detected between wolf populations, not even between the Portuguese and Spanish wolf, which are both from the Iberian wolf population.

Assuming a mutation rate of $1x10^{-8}$ implies more recent divergence times and smaller effective population size estimates, that are ca. 40% of the values stated above (figure 2-4, Supplementary Table S1).

Discussion

Changes in wolf abundance during the late Pleistocene and Holocene

Our demographic analyses imply a dramatic reduction of all studied Eurasian wolf populations ca. 30,000 or 10,000 years ago, depending on the assumed mutation rate. The ancestral wolf population of all modern populations analyzed in this study is estimated by *G-PhoCS* to have had an effective size of only ~4,000, compared to the ~100,000 for the ancestral wolf population from which both wolves and dogs descend. This corresponds to the bottleneck inferred from other genome combinations and methods (Thalmann et al. 2013; Freedman et al. 2014; Fan et al. 2016), and supports the idea that profound environmental changes affected wolf abundance in the late Pleistocene. Depending on the mutation rate used, this decline may reflect the entrance and expansion of modern humans in Eurasia, ca. 30kya, or the dramatic declines associated with megafaunal extinctions, ca. 10kya (Koch and Barnosky 2006; Thalmann et al. 2013; Freedman et al. 2014).

During the Holocene, wolf populations expanded again, with inferred population sizes of approximately half of those before the decline: effective sizes of ~41,000 and ~52,000 are inferred for the ancestral European wolf population and Middle Eastern wolves, respectively.

Divergence of European wolf populations

In our *G-PhoCS* analysis, only the Israeli wolf is included as a representative of *C. l. pallipes*, while all other sampled wolves belong to *C. l. lupus*. We estimate the divergence time for these two groups at ca. 25-10 kya, with limited post-divergence gene flow. Subsequently, these two groups also experienced distinct demographic trajectories

and therefore seem to represent morphologically and genetically distinct evolutionary groups.

Our analyses also suggest that the divergence of the three sampled European populations occurred very closely in time, followed by negligible gene flow between them. Even taking into account the uncertainties regarding the mutation rate, our estimated divergence dates (2.4-7.4 kya) are more recent than the Last Glacial Maximum (LGM, ca. 20-14 kya), suggesting that the environmental changes of Pleistocene glaciations were not the main factor driving the divergence of these populations, although these climatic changes might still have affected their abundance (see below). The Southern European peninsulas of the Balkans, Italy and Iberia served as refugia for many species during the Pleistocene glacial periods (Hewitt 2000; de Bruyn et al. 2011), leading in many cases to their diversification into distinguishable groups that then recolonized the northern regions during warmer periods. For wolves however, no significant loss of distribution during cold periods is observable in the fossil record (Sommer and Benecke 2005), arguing against a retreat to refugia and subsequent recolonization found in many species. It has been proposed nonetheless, based on nuclear microsatellite data, that the Italian wolf population has been isolated for a long time, possibly since the LGM (Lucchini et al. 2004). Our results support the isolation of both the Italian and the Iberian populations but suggest that the onset of this divergence is more recent. Our estimated divergence dates are, however, concordant with those of Pilot et al. (2014) based on 61 thousand SNPs, who estimated a nearly simultaneous separation of the same European wolf populations ca. 5,600-3,200 ya.

An unexpected result is the relatively old divergence time estimated for the Portuguese and Spanish wolves, that is similar in magnitude to the divergence between the Iberian and Italian populations. Iberian wolves currently form a nearly continuous and expanding population in the Northwestern region of the Iberian Peninsula, spanning the Northern parts of both Portugal and Spain (Blanco and Cortés 2001; Pimenta et al. 2005). Our simulations indicate that *G-PhoCS* would be capable of inferring very recent divergence times between genomes sampled from a panmictic population or two very recently diverged populations. Since the comparatively high inferred divergence between Portuguese and Spanish wolves is unlikely to result from actual long-term geographical isolation between them, a possible alternative explanation is the existence of cryptic population structure in the Iberian wolf population. Such population structure has been described in several other apparently connected and uniform wolf populations, and might reflect local adaptations to environmental and prey conditions (Carmichael et al. 2001; Geffen et al. 2004; Pilot et al. 2006; Musiani et al. 2007; Muñoz-Fuentes et al. 2009;

Scandura et al. 2011; Stronen et al. 2013; Fabbri et al. 2014). Preliminary data also indicate the existence of population genetic structure in the Iberian Peninsula (Silva *et al*., in prep. Paper I). The Portuguese and Spanish wolves used in the present study belong to distinct genetic clusters (Alto Minho and Castilla y León, respectively), which might partially explain their differentiation.

Italian and Iberian wolves have been found to appear as highly distinct in previous studies using microsatellite loci, mtDNA sequence divergence and SNP genotypes (Vilà et al. 1999; Lucchini et al. 2004; Pilot et al. 2014). Several morphological differences have been proposed to categorize Italian and Iberian wolves as subspecies distinct from *C. l. lupus* (*C. l. italicus* Altobello, 1921 and *C. l. signatus* Cabrera, 1907, respectively), which include a smaller body size, differences in coat color and pattern and skull morphology (Petrucci-Fonseca 1990; Vilà 1993; Nowak and Federoff 2002). These may result from changes in a limited number of loci as adaptations to unique local conditions, an hypothesis that is supported by a recent study that showed an increase in the distinctiveness of these populations in a PCA when only loci putatively under selection are considered (Pilot et al. 2014). The distinctiveness of these populations could have been further exacerbated by their relatively old divergence, as inferred in our analysis, as well as long-term inbreeding inferred from genomic runs of homozygosity by (Fan et al. 2016).

References

Blanco JC, Cortés Y. 2001. Ecología, Censos, Percepción y Evolución del Lobo en España: Análisis de un Conflicto. Málaga: SECEM.

de Bruyn M, Hoelzel a R, Carvalho GR, Hofreiter M. 2011. Faunal histories from Holocene ancient DNA. Trends in ecology & evolution 26:405–13.

Carmichael LE, Nagy J, Larter N, Strobeck C. 2001. Prey specialization may influence patterns of gene flow in wolves of the Canadian Northwest. Molecular Ecology 10:2787–2798.

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

Dutheil J, Hobolth A. 2012. Ancestral Population Genomics. In: Anisimova M, editor. Evolutionary Genomics SE - 12. Vol. 856. Humana Press. (Methods in Molecular Biology). p. 293–313.

Fabbri E, Caniglia R, Kusak J, Galov A, Gomerčić T, Arbanasić H, Huber D, Randi E. 2014. Genetic structure of expanding wolf (Canis lupus) populations in Italy and Croatia, and the early steps of the recolonization of the Eastern Alps. Mammalian Biology 79:138–148.

Fan Z, Silva P, Gronau I, Wang S, Armero AS, Schweizer M, Ramirez O, Pollinger J, Galaverni M, Ortega-Del Vecchyo D. 2016. Worldwide patterns of genomic variation and admixture in gray wolves. Genome Research 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 genetics 10:e1004016.

Geffen E, Anderson MJ, Wayne RK. 2004. Climate and habitat barriers to dispersal in the highly mobile grey wolf. Molecular Ecology 13:2481–2490.

Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A. 2011. Bayesian inference of ancient human demography from individual genome sequences. Nature Genetics 43:1031–1034.

Hewitt G. 2000. The genetic legacy of the Quaternary ice ages. Nature 405:907–13.

Hudson RR. 2002. Generating Samples under a Wright-Fisher Neutral Model of Genetic Variation. Bioinformatics 18:337–338.

Koch PL, Barnosky AD. 2006. Late Quaternary Extinctions: State of the Debate. Annual Review of Ecology, Evolution, and Systematics 37:215–250.

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

Lucchini V, Galov A, Randi E. 2004. Evidence of genetic distinction and long-term population decline in wolves (Canis lupus) in the Italian Apennines. Molecular Ecology 13:523 – 536.

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

Muñoz-Fuentes V, Darimont CT, Wayne RK, Paquet PC, Leonard J a. 2009. Ecological factors drive differentiation in wolves from British Columbia. Journal of Biogeography 36:1516–1531.

Musiani M, Leonard JA, Dean Cluff H, Cormack Gates C, Mariani S, Paquet PC, Vilà C, Wayne RK. 2007. Differentiation of tundra/taiga and boreal coniferous forest wolves: genetics, coat colour and association with migratory caribou. Molecular Ecology 16:4149–4170.

Nowak R, Federoff N. 2002. The systematic status of the Italian wolf Canis lupus. Acta theriologica 47:333–338.

Nowak RM. 2003. Wolf evolution and taxonomy. In: Wolves: Behavior, ecology, and conservation. University of Chicago Press: Chicago, IL, USA. p. 239–258.

Petrucci-Fonseca F. 1990. O lobo ibérico (Canis lupus signatus Cabrera, 1907) em Portugal. PhD Thesis, Faculdade de Ciencias da Universidade de Lisboa, Lisbon, Portugal.

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

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

Pilot M, Jędrzejewski W, Branicki W, Sidorovich VE, Jędrzejewska B, Stachura K, Funk SM. 2006. Ecological factors influence population genetic structure of European grey wolves. Molecular Ecology 15:4533–53.

Pimenta V, Barroso I, Álvares F, Correia J, Ferrão 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. Lisboa: Instituto da Conservação da Natureza / Grupo Lobo.

Rambaut A, Grassly NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. CABIOS: Computer Applications in the Biosciences 13:235–238.

Rambaut A, Suchard MA, Xie D, Drummond AJ. 2014. Tracer v1. 6. Computer program and documentation distributed by the author, website http://beast.bio.ed.ac.uk/Tracer.

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 343:1241484.

Scandura M, Iacolina L, Capitani C, Gazzola A, Mattioli L, Apollonio M. 2011. Fine-scale genetic structure suggests low levels of short-range gene flow in a wolf population of the Italian Apennines. European Journal of Wildlife Research 57:949–958.

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. Current Biology:1–5.

Sommer R, Benecke N. 2005. Late-Pleistocene and early Holocene history of the canid fauna of Europe (Canidae). Mammalian Biology - Zeitschrift für Säugetierkunde 70:227–241.

Stronen A V, Jędrzejewska B, Pertoldi C, Demontis D, Randi E, Niedziałkowska M, Pilot M, Sidorovich VE, Dykyy I, Kusak J, et al. 2013. North-South differentiation and a region of high diversity in European wolves (Canis lupus). PloS one 8:e76454.

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 (New York, N.Y.) 342:871–4.

Valière N, Fumagalli L, Gielly L, Miquel C, Lequette B, Poulle M-L, Weber J-M, Arlettaz R, Taberlet P. 2003. Long-distance wolf recolonization of France and Switzerland inferred from non-invasive genetic sampling over a period of 10 years. Animal Conservation 6:83–92.

Vilà C. 1993. Aspectos morfológicos y ecológicos del lobo ibérico (Canis lupus L.). PhD Thesis. Universidad de Sevilla, Sevilla, Spain.

Vilà C, Amorim IR, Leonard JA, Petrucci-Fonseca F, Posada D, Crandall KA, Castroviejo J, Ellegren H, Wayne RK. 1999. Mitochondrial DNA phylogeography and population history of the grey wolf canis lupus. Molecular Ecology 8:2089–103.

VonHoldt BM, Pollinger JP, Earl D a., Knowles JC, Boyko AR, Parker H, Geffen E, Pilot M, Jędrzejewski W, Jędrzejewska B, et al. 2011. A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. Genome Research 21:1294–1305.

Wang X, Tedford RH. 2008. Dogs: Their Fossil Relatives and Evolutionary History. Columbia University Press.

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.

Chapter 3 - HISTORICAL DEMOGRAPHY OF WOLVES AND THE DOMESTICATION OF THE DOG

Paper III: Freedman AH, Gronau I, Schweizer RM, Ortega Del-Vecchyo D, Han E, **Silva PM**, Galaverni M, Fan Z, Marx P, Lorente-Galdos B, Beale H, Ramirez O, Hormozdiari F, Alkan C, Vilà C, Squire K, Geffen E, Kusak J, Boyko AR, Parker HG, Lee C, Tadigotla V, Siepel A, Bustamante CD, Harkins TT, Nelson SF, Ostrander EA, Marques-Bonet T, Wayne RK & Novembre J (2014)
"Genome Sequencing Highlights the Dynamic Early History of Dogs".
PLoS Genetics 10(1): e1004016.
doi:10.1371/journal.pgen.1004016

Paper IV: Fan Z*, **Silva P***, Gronau I, Wang S, Armero AS, Schweizer RM, Ramirez O, Pollinger J, Galaverni M, Ortega-Del Vecchyo D, Lianming D, Zhang W, Zhang Z, Xing J, Vilà C, Marques-Bonet T, Godinho R, Yue B & Wayne RK (2016)
"Worldwide patterns of genomic variation and admixture in gray wolves"
Genome Research 26:163-173
doi:10.1101/gr.197517.115
* equal contribution

Paper V: **Silva P** et al. (in prep) The Effects of Population Structure and Sampling Scheme on Demographic Inferences from Microsatellite Data: an Empirical Test on the Iberian Wolf Population

Paper III - Genome Sequencing Highlights the Dynamic Early History of Dogs

Adam H. Freedman[1], Ilan Gronau[2], Rena M. Schweizer[1], Diego Ortega-Del Vecchyo[1], Eunjung Han[1], **Pedro M. Silva**[3], Marco Galaverni[4], Zhenxin Fan[5], Peter Marx[6], Belen Lorente-Galdos[7], Holly Beale[8], Oscar Ramirez[7], Farhad Hormozdiari[9], Can Alkan[10], Carles Vilà[11], Kevin Squire[12], Eli Geffen[13], Josip Kusak[14], Adam R. Boyko[15], Heidi G. Parker[8], Clarence Lee[16], Vasisht Tadigotla[16], Alan Wilton[17], Adam Siepel[2], Carlos D. Bustamante[18], Timothy T. Harkins[16], Stanley F. Nelson[12], Elaine A. Ostrander[8], Tomas Marques- Bonet[7,19], Robert K. Wayne[1]*, John Novembre[1¤*]

[1] Department of Ecology and Evolutionary Biology, University of California, Los Angeles, Los Angeles, California, United States of America, [2] Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York, United States of America, [3] CIBIO-UP, University of Porto, Vairão, Portugal, [4] ISPRA, Ozzano dell'Emilia, Italy, [5] Key Laboratory of Bioresources and Ecoenvironment, Sichuan University, Chengdu, China, [6] Department of Measurement and Information Systems, Budapest University of Technology and Economics, Budapest, Hungary, [7] Institut de Biologia Evolutiva (CSIC-Univ Pompeu Fabra), Barcelona, Spain, [8] National Institutes of Health/NHGRI, Bethesda, Maryland, United States of America, [9] Department of Computer Science, University of California, Los Angeles, Los Angeles, California, United States of America, [10] Bilkent University, Ankara, Turkey, [11] Estación Biológia de Doñana EBD-CSIC, Sevilla, Spain, [12] Department of Human Genetics, University of California, Los Angeles, Los Angeles, California, United States of America, [13] Department of Zoology, Tel Aviv University, Tel Aviv, Israel, [14] University of Zagreb, Zagreb, Croatia, [15] Department of Veterinary Medicine, Cornell University, Ithaca, New York, United States of America, [16] Life Technologies, Foster City, California, United States of America, [17] School of Biotechnology & Biomolecular Sciences, Faculty of Science, The University of New South Wales, Sydney, NSW 2052, Australia. [18] Stanford School of Medicine, Stanford, California, United States of America, [19] Institució Catalana de Recerca i Estudis Avançats (ICREA). 08010, Barcelona, Spain

* E-mail: rwayne@ucla.edu (RKW); jnovembre@uchicago.edu (JN)

¤ Current address: Department of Human Genetics, University of Chicago, Chicago, Illinois, United States of America.

PLoS Genetics 10(1): e1004016. doi:10.1371/journal.pgen.1004016
published January 16 2014

Abstract

To identify genetic changes underlying dog domestication and reconstruct their early evolutionary history, we generated high-quality genome sequences from three gray wolves, one from each of the three putative centers of dog domestication, two basal dog lineages (Basenji and Dingo) and a golden jackal as an outgroup. Analysis of these sequences supports a demographic model in which dogs and wolves diverged through a dynamic process involving population bottlenecks in both lineages and post-divergence gene flow. In dogs, the domestication bottleneck involved at least a 16-fold reduction in population size, a much more severe bottleneck than estimated previously. A sharp bottleneck in wolves occurred soon after their divergence from dogs, implying that the pool of diversity from which dogs arose was substantially larger than represented by modern wolf populations. We narrow the plausible range for the date of initial dog domestication to an interval spanning 11–16 thousand years ago, predating the rise of agriculture. In light of this finding, we expand upon previous work regarding the increase in copy number of the amylase gene (AMY2B) in dogs, which is believed to have aided digestion of starch in agricultural refuse. We find standing variation for amylase copy number variation in wolves and little or no copy number increase in the Dingo and Husky lineages. In conjunction with the estimated timing of dog origins, these results provide additional support to archaeological finds, suggesting the earliest dogs arose alongside hunter-gathers rather than agriculturists. Regarding the geographic origin of dogs, we find that, surprisingly, none of the extant wolf lineages from putative domestication centers is more closely related to dogs, and, instead, the sampled wolves form a sister monophyletic clade. This result, in combination with dog-wolf admixture during the process of domestication, suggests that a re-evaluation of past hypotheses regarding dog origins is necessary.

Introduction

Gray wolves have been dominant predators across Eurasia and North America, often exerting top-down impacts on the ecological communities they inhabit [1,2]. As humans expanded out of Africa into Eurasia, they came into contact with gray wolves and, through a complex and poorly understood process, dogs emerged as the first human companion species and the only large carnivore to ever be domesticated. Archaeological evidence provides partial clues about dog origins. For example, dog-like canids first appear in the fossil record as early as 33,000 years ago in Siberia [3]. However, it is not clear if these proto-dog fossils are ancestral to living dogs or instead represent failed domestication attempts or simply morphologically distinct wolves [3]. Similarly, the geographic origin of dogs is uncertain, with distinct lines of evidence supporting Southeast Asia, the Middle East, and Europe as potential domestication centers, and ruling out Africa, Australia, and North America [4–10]. Nonetheless, several recent studies have begun to illuminate the genetic basis of traits that changed during dog domestication and breed formation, advancing the general understanding of how genetic mechanisms shape phenotypic trait diversity [11–14]. For example, a recent study found an increase in copy number of the amylase gene (AMY2B) during dog domestication suggesting adaptation to starch-rich diets [15]. Given the unique behavioral adaptations of dogs, including docility and the ability to form social bonds with humans [16], comparative genomics analyses of dogs and wolves holds great promise for identifying genetic loci involved in complex behavioral traits [14]. However, the demographic context of selection must first be understood to determine how it may have affected patterns of genetic divergence between dogs and wolves.

To advance the understanding of dog origins and genetic changes early in dog domestication, we sequenced the genomes of six canid individuals, including three wolves (*Canis lupus*), an Australian Dingo, a Basenji and a golden jackal (*Canis aureus*).

The three wolves were chosen to represent the broad regions of Eurasia where domestication is hypothesized to have taken place (Europe, the Middle East, and East/Southeast Asia) [6], and specifically, were sampled from Croatia, Israel, and China (Figure 3-1). The Dingo and Basenji represent divergent lineages relative to the reference Boxer genome [10] and maximize the opportunity to capture distinct alleles present in the earliest dogs. These lineages are also geographically distinct, with modern Basenjis tracing their history to hunting dogs of western Africa, while Dingoes are free-living semi-feral dogs of Australia that arrived there at least 3,500 years ago (Figure 3-1) [17]. As a result of their geographic isolation, the natural range of wolves has never extended as far south as the geographic sources for these two dog lineages [6], thus

they are less likely to have overlapped and admixed with wolves in the recent past. Sequencing the golden jackal in principle allows us to infer the ancestral state of variants arising in dogs and wolves (Text S1, S2), though in practice this was complicated by the observation of wolf-jackal admixture (see below). For some analyses, we also leverage data from a companion study of 12 additional dog breeds (Text S1).

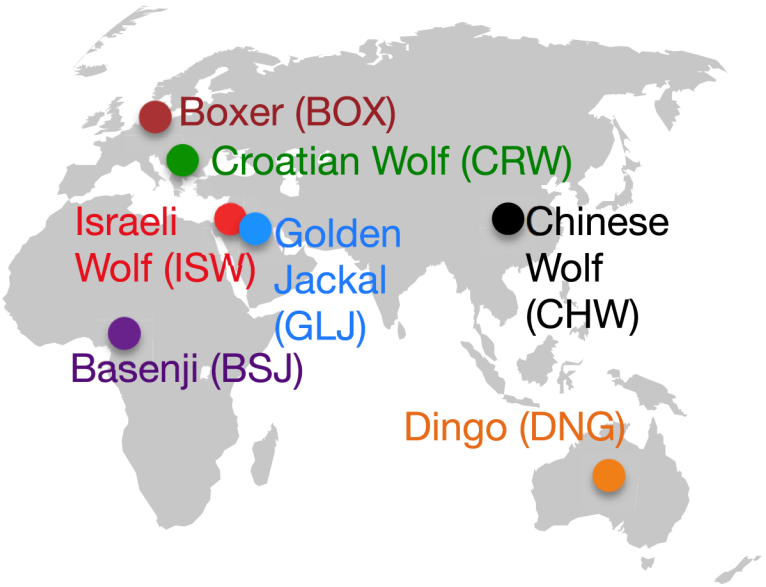


Fig. 3-1: Geographic distribution of sampled lineages.

We chose to sequence a small number of individual genomes to high coverage, rather than larger numbers of (pooled) individuals at low coverage, to take advantage of recently developed demography inference methods based on small numbers of high quality genomes [18–20]. These methods allowed us to disentangle the effects of incomplete lineage sorting (ILS) – the discordance from the population phylogeny at individual loci resulting from deep coalescence – and post-divergence gene flow, which pose a particular challenge in analysis of such recently diverged species as dogs and wolves [21]. Combining the results of multiple complementary methods provided us with an integrated, robust view of the shared history of dogs and wolves, including population divergence times, ancestral population sizes, and rates of gene flow. Using polymorphism data from 10 million single-nucleotide variant sites, we investigated: 1) the size of the ancestral wolf population at the time of wolf/dog divergence; 2) the geographic origins and timing of dog domestication; 3) post-divergence admixture between dogs and wolves; and 4) lineage-specific characteristics of the recently discovered dog-specific *AMY2B* expansion [15].

Materials and Methods

Samples and sequencing

We selected six samples for genome sequencing and generated single end and long mate pair SOLiD reads. We generated additional paired end (PE) sequence data on the Illumina HiSeq platform (Text S1). For most downstream analyses, we also utilized sequence information from the Boxer reference genome (CanFam 3.0).

Sequence alignment, genotyping, and filters

We aligned sequence reads to CanFam 3.0, with post-processing of aligned reads including the removal of duplicates, local realignment, and base quality recalibration (Text S3). We then genotyped each sample individually, using the Genome Analysis Toolkit (GATK) pipeline [34]. For SNV genotyping and analysis, we excluded repeats of recent origin, CpG sites, regions falling in copy number variants, and triallelic sites, and at the sample level we filtered out genotypes proximate to called indels, with excess depth of coverage, with low genotype quality scores, or where the SNV fell within five base pairs of another SNV (Text S4).

Genotype validation

We compared genotype calls based upon sequencing to those from the same samples using the Illumina CanineHD BeadChip, which consists of >170,000 markers evenly spaced throughout the dog genome (Text S5). We also analyzed variants overlapping those generated in a previous SNP array study of a large panel of dogs and wolves [10], and performed PCA on the combined data set to verify that NGS genotypes clustered with array genotypes for the same lineages (Text S5).

Structural variant detection

We delineated segmental duplications in our six genomes by identifying regions with a significant excess depth of coverage (Text S6). For this purpose, we aligned Illumina and SOLiD reads with MrFAST [35] and drFAST [36] respectively. Absolute copy numbers were calculated using mrCaNaVar version 0.31 (http://mrcanavar.sourceforge.net/). In the particular case of the previously reported *AMY2B* expansion in the dog lineage [15] we also examined patterns of copy number across 52 breed dogs, six Dingoes, and 40 wolves using qPCR (Text S6).

Functional element annotation

In order to conduct demographic analyses on putatively neutral genomic regions without any apparent functional annotation, we first identified genic region using annotations from the union of refGene, Ensembl and SeqGene annotation databases, with the condition that all annotated transcripts had proper start and stop codons, and contained no internal stop codons (Text S7). In addition, we defined conserved non-coding elements (CNEs) on the basis of phastCons scores [37] (Text S7).

N_e through time: Pairwise-Sequential-Markov-Coalescent (PSMC)-based inference

We used the PSMC methods developed by Li and Durbin [20] to infer the trajectory of population sizes across time for the six canid genome sequences (Text S8).

Testing for admixture: ABBA-BABA

To investigate the extent of gene flow between wolves and dogs subsequent to their divergence, we employed a method recently developed by Durand et al. [18]. This method tests for directional gene flow by testing for asymmetries in allele sharing between a source lineage (P3), and either of two receiving lineages (P1, P2) with reference to an outgroup (O). To focus on gene flow most germane to evolutionary processes influencing wolf-dog divergence, we restricted testing to those cases where one of the dog samples was P3, the other two (P1 and P2) were wolves, and viceversa (P3=wolf, P1 and P2=dogs). For more details, see Text S8.

Demographic model for dog domestication

Our main demographic analysis is based on the Generalized Phylogenetic Coalescent Sampler (G-PhoCS) developed by Gronau et al. [19] and which we applied to 16,434 1 kb loci chosen via a strict set of criteria to obtain putatively neutral loci (Text S9). The prior distributions over model parameters was defined by a product of Gamma distributions using the default setting chosen by Gronau et al. [19]. Markov Chain was run for 100,000 burn-in iterations, after which parameter values were sampled for 200,000 iterations every 10 iterations, resulting in a total of 20,001 samples from the approximate posterior. Convergence was inspected manually for each run. We conditioned inference on the population phylogeny based upon the neighbor-joining tree constructed from the genome-wide distance matrix described above (Fig. 3-2A). We also constructed models under a 'regional domestication' scenario, in which each dog lineage originated from a wolf lineage from the same geographic region, i.e. Basenji from Israeli

wolf, Boxer from Croatian wolf, and Dingo from Chinese wolf. We assessed models in which the branch ancestral to dogs was sister to a particular extant wolf population, or one of internal branches in the wolf clade. In addition, we investigated the sensitivity of parameter estimates to choice of locus length, number of loci, intra-locus recombination, distance from coding exons, and selection on linked sites. For more details, see Text S9.

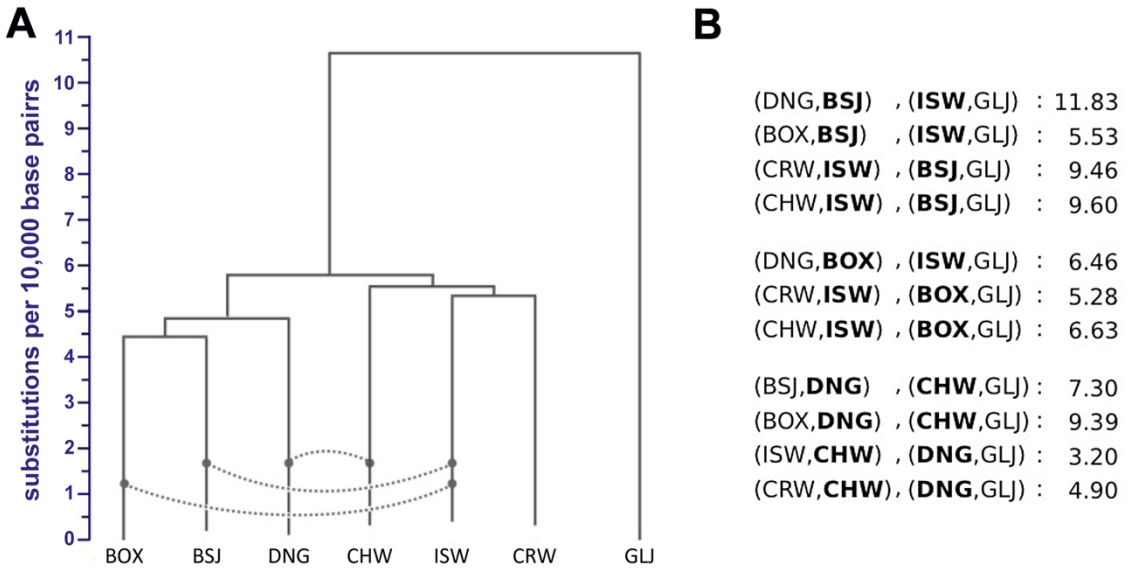


Fig. 3-2: Neighbor-joining tree and admixture signatures from ABBA/BABA tests. (A) NJ tree constructed from genome-wide pairwise divergence, calculated using equation E8.1 in Text S8. All nodes have 100% bootstrap support. Dashed lines indicate admixture edges that were statistically significant in ABBA/BABA tests. (B) ABBA/BABA tests with significant Z-scores (values >3 are significant). All comparisons made are shown in Table S11. For each row, boldfaced labels indicate admixing lineages.

Results

Individual-level genome sequences

For each of the six samples, we generated high-quality genome sequences. Cumulative coverage was 726 for the wolves (246 average per individual), 386 coverage for the two dogs (196 average per individual), and 246for the golden jackal, for a total of 335 Gb of uniquely aligned sequence from 11.2 billion reads (Table S1). Surveys of wolf genetic diversity to date have been limited to shotgun sequencing with incomplete genomic coverage [22], small numbers of sequence loci [23], limited pooled sequencing (66 average from a pool of 12 wolves, 306 average from a pool of 60 dogs) [15] or lower coverage sequencing (9–116 coverage of 4 wolves, 9–146of 7 dogs) [24].

Our analyses draw on 10,265,254 high quality variants detected by our genotyping pipeline (Text S3, S4, S5), of which 6,970,672 were at genomic positions with no missing data for any lineage (Tables S2, S3). We estimate genotype error rates to be very low based on comparison to genotype calls from genotyping arrays (e.g.

heterozygote discordance rates of 0.01–0.04%, Tables S4, S5, Text S5). Further, PCA on the intersection of sequencing and genotyping array variants show the novel samples cluster appropriately, suggesting batch effects due to technology have been minimized (Figure 3-3, Text S5).

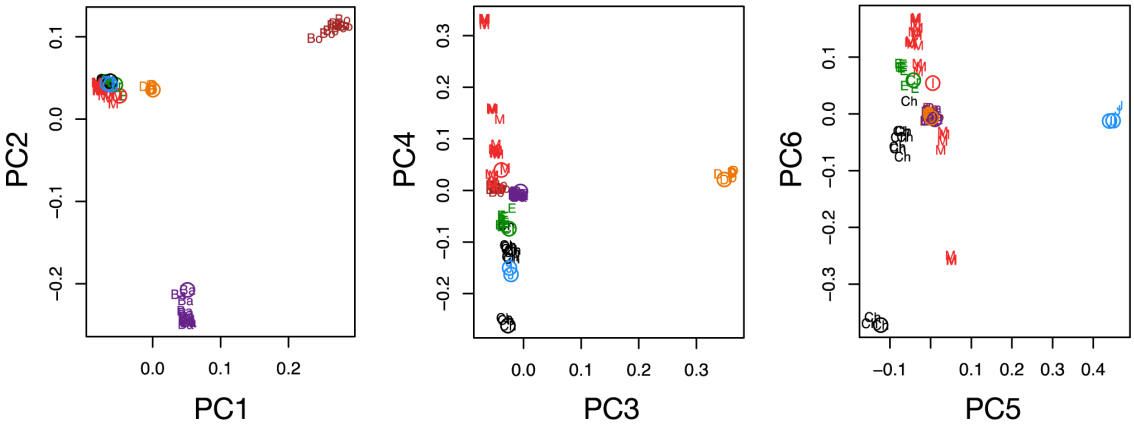


Fig. 3-3: Comparison of next generation sequencing with array typed samples, and historical changes in effective population size. PCA plot of next-generation sequencing (NGS) samples generated in this study (open circles) along with corresponding samples genotyped on the Affymetrix canid array [10] (colors and two letter codes: red M = Mid-East Wolf, green E = European Wolf, black Ch = Chinese Wolf, purple Ba = Basenji, brown Bo = Boxer, orange D = Dingo, cyan J = Golden Jackal).

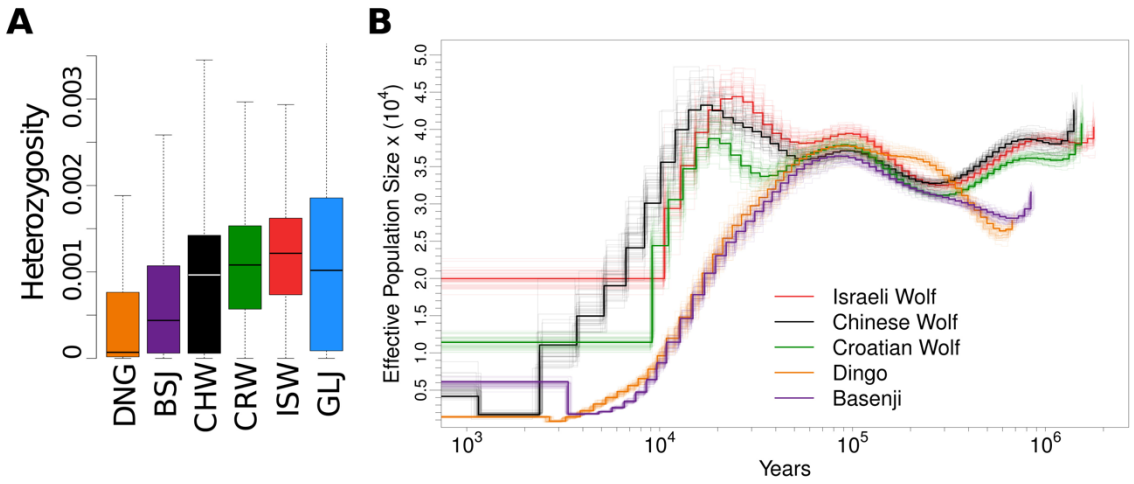


Fig. 3-4: Heterozygosity and historical changes in effective population size. (A) Box plots of heterozygosity measured in 5000 100 kb windows for each sample. (B) Reconstruction of historical patterns of effective population size (Ne) for individual genome sequences. Based upon the genomic distribution of heterozygous sites using the pairwise sequential Markovian coalescent (PSMC) method of Li and Durbin 2011 [20]. Time scale on the x-axis is calculated assuming a mutation rate of $1x10^{-8}$ per generation (see Text S8); estimates from the full data and 50 bootstraps are depicted by darker and lighter lines, respectively.

Ancestral population sizes of dogs and wolves

Genome-wide patterns of heterozygosity provide useful information on long-term effective population sizes. The mean heterozygosity rates (per nucleotide position) observed in the genome sequences of the Basenji and Dingo were $9x10^{-4}$ and $6x10^{-4}$,

respectively (Figure 3-4A, Table S6), consistent with a rate of $6x10^{-4}$ previously observed in modern dog breeds [22], and considerably smaller than the rates observed in the three wolf genomes ($1.2x10^{-3}$–$1.6x10^{-3}$). This twofold reduction in heterozygosity observed in dogs relative to wolves can be superficially interpreted to reflect a relatively weak two-fold reduction in effective population size of dogs relative to their ancestors, assuming that genetic variation in modern wolves is representative of the ancestral population.

To better understand the changes in ancestral population sizes that influenced dogs and wolves, we employed the pairwise sequential Markovian coalescent (PSMC) method [20]. This method infers ancestral effective population sizes (N_e) over time, based on a probabilistic model of coalescence with recombination and changes in heterozygosity rates along a single diploid genome. We applied PSMC to each of the five genomes (Figure 3-4B, Text S8) and converted the mutation-scaled estimates of time (to years) and population size (to numbers of individuals) by assuming an average mutation rate per generation of $\mu = 1x10^{-8}$ and an average generation time of three years [22,25] (see Discussion). The inferred tracks of ancestral N_e in dogs show a population decline of at least 16-fold over the past 50 thousand years, from greater than 32,000 individuals (ancestral N_e for Basenji lineage: 32,100–35,500; for Dingo lineage: 32,500–37,400 95% bootstrap CI) to less than 2,000 individuals (Basenji lineage: 1640–1980 at 4,000 years ago; Dingo lineage: 704–1042 at 3,000 years ago). Interestingly, wolves also show a considerable, yet milder, 3-fold reduction in effective population size to present estimates between 10,000 and 17,000 for the three wolf samples. For clarity, we note that with PSMC the population size trajectories are effective sizes for the lineages that eventually lead to the canid samples as they are known today (e.g. as Basenji or as Dingo) and that looking backwards in time eventually trace back to the common ancestral lineage of dogs and wolves. Our observations do not appear to be biased by very recent inbreeding in dogs and wolves, as we found that runs of homozygosity do not affect our inferences of ancestral N_e (Text S8). These results indicate the ancestral wolf population from which dogs were domesticated was considerably larger than estimated from current levels of diversity in wolves and suggest that simple comparisons of nucleotide diversity in present-day dogs and wolves lead to substantial underestimates of the severity of the bottleneck in dogs.

Phylogenetic relationships and admixture between dogs and wolves

Individual genome sequences include valuable information about phylogenetic relationships between our samples. However, interpretation of these phylogenetic signals is challenging due to the possibility of post-divergence gene flow between dogs

and wolves, as well as ILS, which is an expected consequence of the large ancestral population sizes inferred by PSMC. Indeed, we observed predominant ancestral polymorphism in our data: for variant sites with no missing data, and where variants were observed in dogs or wolves, 32.0% of variant sites were shared across dogs and wolves, 47.3% were private to wolves, 20.2% were private to dogs, and only 0.5% were fixed between dogs and wolves (Table S3). Pairwise sequence divergence captures mean coalescent times that are robust to both ILS and moderate levels of gene flow (see below). Thus, to provide accurate estimates of phylogeny given these demographic processes, we constructed a neighbor-joining (NJ) tree from a conservative estimator of genome-wide pairwise sequence divergence for all pairs in our seven genomes, including the Boxer reference and using the golden jackal as an outgroup (Figure 3-2A, Text S8, Table S7). Bootstrap support for all nodes was 100%, with dogs and wolves recovered as monophyletic sister clades. Surprisingly, the Boxer reference is only slightly more divergent from the three wolf genomes than it is from the two dog genomes. To evaluate the robustness of our phylogenetic inference, we also constructed a NJ tree using an estimator of sequence divergence for which all possible mismatches between alleles from a pair of individuals are counted (Table S8). The consensus tree based on this metric places the Chinese wolf at a position sister to a clade of our other wolf and dog samples (Figure S1), but the bootstrap support for this relationship is low (54%), suggesting poorer resolution with this estimator. Importantly, both approaches and additional phylogenetic analyses strongly support the hypothesis of dogs forming a distinct clade (Text S8, Tables S9, S10).

One important factor that could complicate inference of divergence between dogs and wolves is post-divergence gene flow. To examine admixture in our sampled genomes, we employed the nonparametric 'ABBA-BABA' test for gene flow between two divergent populations, such as humans and Neandertals [26], from individual genome sequences. This method tallies site patterns for four taxa, compares them to those expected under an assumed phylogeny and then uses this comparison to identify significant pattern asymmetries that cannot be explained by ILS or sequencing errors. We applied this test to all dog-wolf sample pairs, using the golden jackal as an outgroup and one of the other four samples as an additional ingroup (Text S8). We found significant evidence of admixture for three population pairs: Israeli wolf and Basenji, Chinese wolf and Dingo, and Israeli wolf and Boxer (Figure 3-2B, see also Table S11). Care should be taken in interpreting these results, as the detected admixture signals may reflect gene flow between lineages ancestral to our contemporary samples. The signal for Chinese wolf and Dingo likely represents ancient admixture in Eastern Eurasia,

and the signal observed for Israeli wolf, Basenji, and Boxer likely represents ancestral admixture that occurred in western Eurasia. The resulting phylogeny with admixture edges (Figure 3-2A) is used as the starting point for a more comprehensive examination of joint demographic model for dogs and wolves.

A complete demographic model for dogs and wolves

We next inferred a complete demographic model for dogs and wolves, including population divergence times, ancestral population sizes, and rates of post-divergence gene flow by jointly analyzing all seven genomes using the Generalized Phylogenetic Coalescent Sampler (G-PhoCS) [19], a recently developed Bayesian demographic inference method. The method is based on a full coalescent-based probabilistic model that considers both ILS (by modeling ancestral population size) and post-divergence gene flow (by allowing lineages to migrate between populations through designated migration bands). G-PhoCS conditions its inference on a given population phylogeny, and uses information on local genealogies at a large number of short, unlinked, neutrally evolving loci to generate samples of demographic parameters from an approximate posterior distribution. We applied G-PhoCS to a multiple sequence alignment of the six genomes and Boxer reference in 16,434 carefully filtered putative neutral autosomal loci using the NJ tree to indicate the topology of the population phylogeny (Text S9, see discussion on alternative topologies below).

Initially, we considered various migration bands with significant signatures of gene flow (Text S9). We found evidence of bi- directional gene flow between Israeli wolf and Basenji, as well as Chinese wolf and Dingo, consistent with our findings from the non-parametric ABBA-BABA test. Interestingly, the joint analysis of all genomes indicated that admixture inferred by the ABBA-BABA test for the Israeli wolf and the Boxer is likely a result of gene flow from a population ancestral to Basenji into a population ancestral to Israeli wolves. We base this conclusion on the observation that there is no significant signature of admixture between Boxer and Israeli wolf in the ABBA-BABA test or the G-PhoCS inference when Basenji is also included in the analysis. Using G-PhoCS we were also able to examine signatures of admixture in the jackal outgroup, which cannot be detected using the ABBA-BABA test, and found significant gene flow between the golden jackal and the Israeli wolf, as well as the population ancestral to all dog and wolf samples.

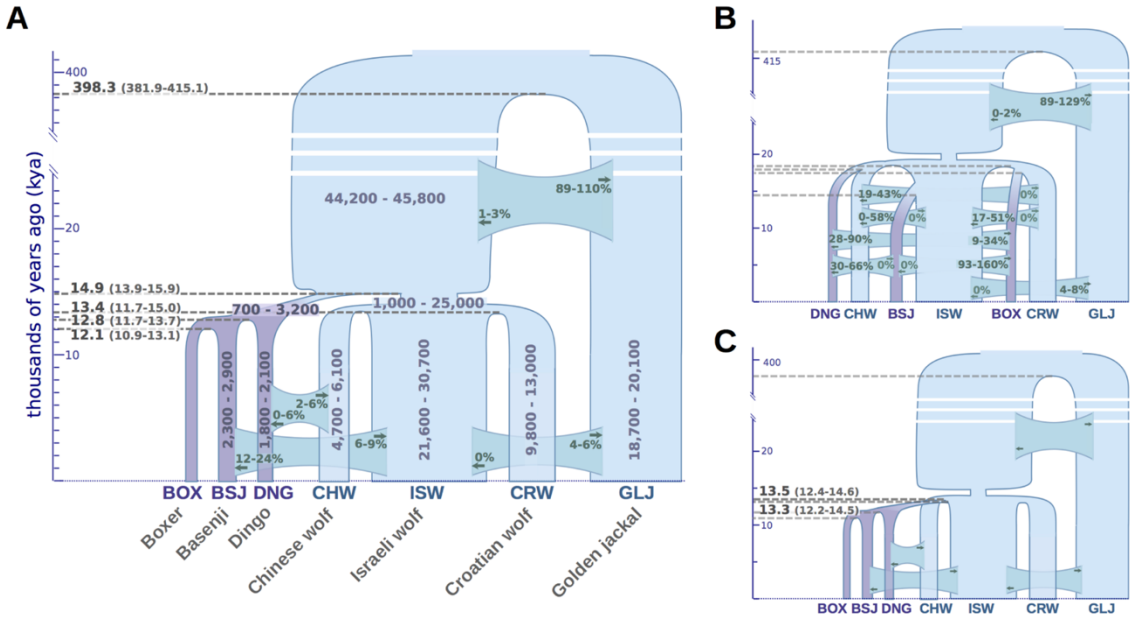


Fig. 3-5: Demographic model of domestication. Divergence times, effective population sizes (Ne), and post-divergence gene flow inferred by G-PhoCS in joint analysis of the Boxer reference genome, and the sequenced genomes of two basal dog breeds, three wolves, and a golden jackal. The width of each population branch is proportional to inferred population size, and stated ranges of parameter estimates indicate 95% Bayesian credible intervals. Horizontal gray dashed lines indicate timing of lineage divergences, with associated means in bold, and 95% credible intervals in parentheses. Migration bands are shown in green with associated values indicating estimates of total migration rates, which equal the probability that a lineage will migrate through the band during the time period when the two populations co-occur. Panels show parameter estimates for (A) the population tree best supported by genome-wide sequence divergence (Fig. 4A) (B) a regional domestication model, and (C) a single wolf lineage origin model in which dogs diverged most recently from the Israeli wolf lineage (similar star-like divergences are found assuming alternative choices for the single wolf ancestor. Estimated divergence times and effective population sizes are calibrated assuming an average mutation rate of $1x10^{-8}$ substitutions per generation and an average generation time of three years. See Text S9 and Table S12 for details.

Our divergence time estimates imply that dogs and wolves diverged 14.9 thousand years ago (kya) with 13.9–15.9 kya Bayesian 95% credible interval (CI), assuming an average mutation rate per generation of $\mu = 1x10^{-8}$ and three years per generation (Figure 3-5A). Divergence times between wolf populations were tightly clustered at 13.4 kya (11.7–15.1 kya), and divergence between dogs was estimated to have occurred slightly more recently, at 12.8 kya (11.8–13.7 kya; divergence of Dingo) and 12.1 kya (10.9–13.1 kya; divergence between Boxer and Basenji). Interestingly, we inferred a divergence time of 398 kya (382–415 kya) between the golden jackal and the population ancestral to dogs and wolves, which is considerably more recent than previously reported [27]. To validate this finding, we ensured that our estimates appropriately account for ancestral gene flow into the golden jackal population (Text S9) and validated the position of our sample within the golden jackal lineage by comparing polymorphism data from that genome to a larger panel of wolves and jackals (Text S5, S11).

G-PhoCS produced estimates of ancestral effective population sizes compatible with the ones inferred by PSMC, with a large effective population size of 45,000

individuals (44,200–44,800) for the population ancestral to dogs and wolves, followed by a 22-fold reduction to 2,000 individuals (700–3,200) in the population ancestral to all dogs, and a more moderate 3.6-fold reduction to 12,600 individuals (1,000–25,000) in the population ancestral to all wolves. As with our inferences based on PSMC, we estimate a far more severe domestication bottleneck than previously reported [22,23].

The main discrepancy between PSMC and G-PhoCS concerns the timing of these changes. G-PhoCS associates this reduction in N_e with the divergence between dogs and wolves at around 15 kya, whereas PSMC infers a gradual population decline starting as early as 50 kya (Figure 3-4B). As PSMC is based upon the density of heterozygous sites within the genome sequence of an individual, it does not directly infer divergence times. However, one can informally estimate them as the points when N_e trajectories that are overlapping diverge moving forward in time towards the present. The discrepancy between G-PhoCS and PSMC reflects the distinct models used by these methods: G-PhoCS assumes a constant population size for every branch of the phylogeny, which prevents it from characterizing gradual changes in population size, whereas PSMC tends to produce smoothed traces of ancestral N_e, which may limit its ability to capture rapid population bottlenecks. To test which of the inferred models has a better fit to the data, we simulated data under both models, and then used each method to analyze the data simulated under the model inferred by the other method (Text S8, S9). These two reciprocal tests indicated that the early and gradual population decline inferred by PSMC is compatible with a more recent dramatic reduction (Text S8, Figure S2), and that divergence time estimates of G-PhoCS were not compromised by its inability to model gradual changes in N_e (Figure S3). Both results support the demographic model inferred by G-PhoCS, which has a relatively recent divergence between dogs and wolves followed by a dramatic reduction in population size. We additionally validated the robustness of our demographic parameter estimates under the set of loci chosen for the analysis as well as assumptions made on intra-locus recombination (Text S9).

Alternative models for dog domestication

The demographic model we inferred using G-PhoCS reflects the population phylogeny estimated in the NJ analysis. To validate the robustness of our inference to this assumption, we considered a series of alternative topologies that correspond to plausible scenarios of the shared histories dogs and wolves. When we assume a model in which each dog population originated from the wolf population corresponding to its geographic origin (a model of regional domestication, e.g. Figure 3-5B), G-PhoCS infers

extremely large rates of post-divergence gene flow between dogs and between wolves. For instance, the total rate of gene flow from Basenji to Boxer is inferred to be $m^{tot} = 1.24$ (0.93–1.59, 95% Bayesian CI), implying a probability near 100% for any Boxer lineage to have migrated from a population ancestral to Basenji. Total rates above 30% were inferred for additional migration bands, such as Basenji-to-Dingo (0.47), Croatian-to-Israeli wolf (0.33), and Croatian-to-Chinese wolf (0.33) (Figure S4). In terms of the number of migrants per generation ($4N_em$), these estimates translate into 0.26 (CI: 0.15–0.38), 4.48 (CI: 2.52–6.36), and 0.89 (CI: 0.56–1.23), reflecting large amounts of gene flow, which is unlikely given historical separation of these geographically distinct populations. In contrast, the migration rates estimated in our original inference were considerably lower, with nearly all total rates falling below 10% (Figure 5, Text S9, Table S12), indicating a better fit of that topology to the data.

Another set of alternative topologies we examined is one in which the dog clade originates from one of the four branches in the wolf sub-phylogeny (e.g. Figure 3-5C). Assuming such topologies, G-PhoCS infers that dogs diverged from wolves less than 200 years after wolves diverged from each other (Figure S5), whereas in the original inference conditioned on the NJ tree, the divergence between dogs and wolves was estimated to have occurred 1,400 years before the divergence between wolf populations. All other parameter estimates were not significantly affected by the choice of origin population for the dog clade. Thus regardless of our assumptions on the identity of the wolf population from which dogs originated, we infer that dogs diverged from the sampled wolf populations at about the same time these wolf populations diverged from each other. Additionally, the greater difference between estimated divergence times in our original analysis provides some support for our initial assumption that dogs and wolves form sister clades.

Assessment of models in lights of site configuration statistics

Because G-PhoCS does not yet support statistical tests for model selection, we assessed relative support for the alternative models by performing simulations under each model, and comparing the simulated and real data with respect to a series of site configuration statistics informative about the topologies of local genealogies. For every quartet in our sample set that contains the jackal outgroup, we computed the relative frequencies of bi-allelic sites in which each of the two alleles (denoted A and B) is present in exactly two of the four individuals. Similar statistics are used in the ABBA- BABA test for admixture, but in our case we were also interested in the frequency of the BBAA configuration, which is the one compatible with the topology of the assumed phylogeny

(see Text S8 for more information). We compared frequencies of the three configurations in 20 quartets observed in our data with those observed in data simulated under the three demographic models shown in Figure 5, denoted as ‘‘dog/wolf reciprocal monophyly’’ (Figure 3-5A), ‘‘regional domestication’’ (Figure 3-5B), and ‘‘ISW-source’’ (Figure 3-5C). This comparison allowed us to draw conclusions regarding the fit of each of these models to the data with respect to the distribution of local genealogies it implies (Table S10).

The three models appeared to be fairly compatible with the data overall, with the reciprocal monophyly model showing the lowest discrepancy (absolute error=0.43), followed closely by the ISW-source model (absolute error = 0.47) and then trailed by the regional domestication model (absolute error = 0.82). The regional domestication model showed the largest discrepancy in quartets including Dingo and at least one other dog, indicating considerably weaker support for the dog clade and its internal structure than present in the data. This implies that the patterns of sequence similarity between dogs are more compatible with a distinct dog clade than they are with similarity solely generated by gene flow between the different dog lineages. The ISW-source model showed high discrepancy in quartets containing the Croatian and Israeli wolves, indicating that the model has problems capturing the phylogenetic relationships between those wolves and the dogs. The reciprocal monophyly model provided the best fit to the data, but it did show some discrepancy in quartets containing both the Dingo and the Chinese wolf. This is perhaps related to the large credible intervals for the rates of gene flow between these populations in the G-PhoCS inference (CHW®DNG, 0–6%; DNG®CHW, 2–6%). In conclusion, these tests show that topological signatures in the data provide strong support for a monophyletic dog clade and somewhat weaker support for a monophyletic wolf clade.

Amylase expansion and dog origins

Our inference of a pre-agriculture origin of dogs provides an important context for re-assessing the recent hypothesis that copy number expansion at the amylase locus (*AMY2B*) in dogs was an important part of the domestication process [15]. In that study, copy number segregated between species, with only two copies of the gene in each of the 35 wolves genotyped and an average 7.4- fold increase across 136 dogs. This finding was interpreted to suggest that AMY2B expansion enabled early dogs to exploit a starch-rich diet as they fed on refuse from agriculture. Surprisingly, and using the corrected depth of coverage to estimate discrete gene copy number, we find the Dingo has just two copies of *AMY2B* (Figure 3-6A, Text S6), suggesting that the *AMY2B* copy number

expansion was not fixed across all dogs early in the domestication process. In a survey of sequence data from 12 additional domestic dog breeds, we find that the Siberian Husky, a breed historically associated with nomadic hunter gatherers of the Arctic, has only three to four copies of *AMY2B*, whereas the Saluki, which was historically bred in the Fertile Crescent where agriculture originated, has 29 copies (Figure S6). In order to validate the results, we used real-time quantitative PCR (qPCR) to explore the variation in *AMY2B* copies across additional breed dogs (n = 52), additional dingoes (n = 6) and a worldwide distribution of wolves (n = 40) (Text S6). The qPCR results show modern dog breeds on average have a high copy number of AMY2B and that wolves and Dingoes do not (Figure 3-6B, Table S13). However, the qPCR results also shows that the *AMY2B* expansion is polymorphic in wolves (16 of 40 wolves with >2 copies Figure 3-6B) and thus is not restricted to dogs.

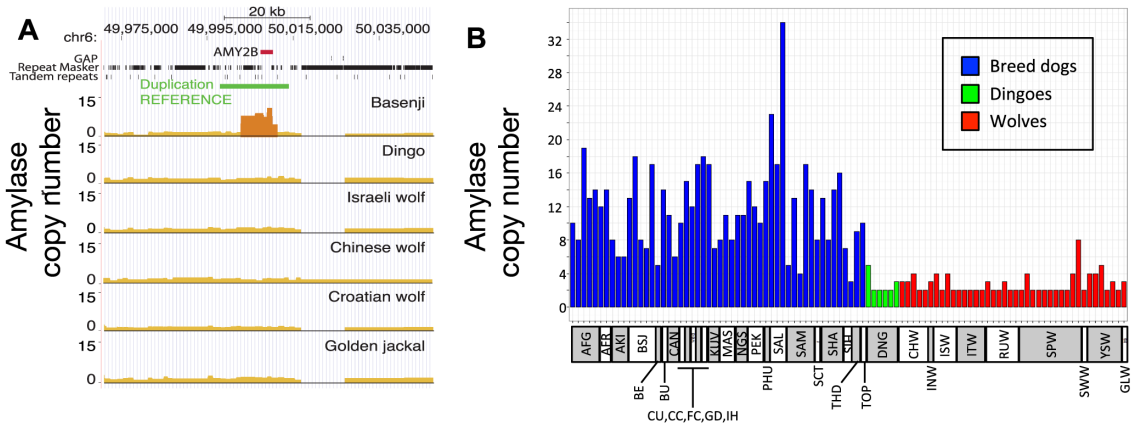


Fig. 3-6: Copy number variation at amylase (AMY2B) locus. (A) Copy number variation (CNV) at AMY2B estimated from whole genome sequence data, showing presence of elevated copy number in Basenji but not in other lineages. Results are based on SOLiD data, except for the Chinese wolf (see Text S6 for supporting results and Text S10 for CNV analyses in an additional 12 dog breeds). (B) qPCR results on CNV state in an expanded set of wolf and dog lineages. Abbreviations for lineages are: AFG, Afgan Hound; AFR, Africanis; AKI, Akita; BSJ, Basenji; BE, Beagle; BU, Bulldog, CAN, Canaan Dog; CU, Chihuahua; CC, Chinese Crested; FC, Flat-coated Retriever; GD, Great Dane; IH, Ibizan Hound; KUV, Kuvasz; MAS, Mastiff; NGS, New Guinea Singing Dog; PEK, Pekinese; PHU, Phu Quoc; SAL, Saluki; SAM, Samoyed; SCT, Scottish Terrier; SHA, Shar Pei; SIH, Siberian Husky; THD, Thai Dog; TOP, Toy Poodle; DNG, Dingo; CHW, Chinese wolf; INW, Indian wolf; ISW, Israeli wolf; ITW, Italian wolf; RUW, Russian wolf; SPW, Spanish wolf; YSW, Yellowstone wolf; GLW, Great Lakes wolf.

Discussion

In this study, we generated high-quality individual canid genomes, and used them to uncover the history of dogs and gray wolves. Interpretation of the phylogenetic signals in these genomes was particularly challenging due to high levels of incomplete lineage sorting and post-divergence gene flow. We were able to disentangle the effects of these factors by using an array of recently developed statistical methods that together provided a detailed and robust inference of past demography for these canids. We used methods

that rely on different aspects of this dataset: 1) whole- genome patterns of heterozygosity in single individuals (PSMC), 2) a subset of sites that are informative for post-divergence admixture (ABBA/BABA analyses) and 3) a set of neutral loci analyzed jointly across all individuals (G-PhoCS).

We found evidence of wolf-dog admixture in two divergent dog lineages (Basenji and Dingo). The fact that these lineages have been geographically isolated from wolves in the recent past suggests that this gene flow was ancestral and thus likely impacted multiple (if not most) dog lineages [28,29]. Admixture has likely complicated previous inferences of dog origins. For instance, the presence of long shared haplotypes in Middle East wolves with several dog breeds [10] may reflect historic admixture rather than recent divergence. Similarly, elevated genetic diversity in East Asian dogs and affinities between East Asian village dogs and wolves [7,9,24] may be confounded by past admixture with wolves. In areas where village dogs [30] roam freely and wolves have historically been in close proximity, admixture may also be present and exert a non-trivial impact on patterns of genetic variation [21].

Our inferences of ancestral population size from both PSMC and G-PhoCS revealed an unexpected, roughly threefold population bottleneck in wolves. With PSMC, we detect the start of this bottleneck as early as 20 kya, while with G-PhoCS the bottleneck occurs at the timing of dog-wolf divergence, approximately 15 kya. Because our simulations indicated that the timing of abrupt changes in N_e are overestimated by PSMC (Text S8, S9, Figure S2), we place higher confidence in the more recent date inferred with G-PhoCS. Regardless of the method chosen, the bottleneck in wolves appears to have occurred well in advance of direct extermination campaigns by humans and within the timeframe of environmental and biotic changes associated with the ending of the Pleistocene era. Although the specific cause of this bottleneck is unknown, it has important implications for dog domestication. Because of this bottleneck, we expect that at the onset of domestication, there was substantially more genetic diversity for selection to act on than what is observed in modern wolves. Direct comparisons of dog and wolf diversity (such as comparisons of heterozygosity) will not show as large a difference and thus previous studies that did not consider a wolf population decline [22,23] have underestimated the bottleneck associated with domestication. These previous studies estimated a two to fourfold reduction in dog N_e, a far milder population contraction than the at least 16-fold reduction we infer here.

We provide several lines of evidence supporting a single origin for dogs, and disfavoring alternative models in which dog lineages arise separately from geographically distinct wolf populations (Figures 3-2 and 3-5, Table S10). Considering a

full multi-population demographic model with gene flow, we infer that dogs diverged from wolves at around 15 kya (CI: 14–16 kya). Examination of previous estimates shows a wide range of suggested divergence times [24,25]. However, most of the discrepancy between different studies can be traced to differences in the assumed mutation rate. We assume an average mutation rate per generation of $1x10^{-8}$ and an average generation time of three years. However, we observed that CpG di-nucleotides, which we filtered out from the data, contribute roughly 30% of mutations in these canid genomes, similar to what was observed in human genomes [19]. Thus our assumptions regarding mutation rate imply a genome-wide rate (i.e. including filtered sites) of $1.4x10^{-8}$. Other studies of dog domestication assume slightly lower genome-wide rates. For instance, a recent study based on shotgun sequencing data [25] assumes a mutation rate of $1x10^{-8}$ and estimates the divergence time to be 14 kya (CI:11–18 kya) or 30 kya (CI:15–90 kya), depending on the assumed amount of gene flow. Another recent study [20] assumes an even lower mutation rate of $0.66x10^{-8}$ and estimates the divergence time at roughly 32 kya. Calibrating the different estimates using the same mutation rate shows a remarkable consistency with our conclusions. Unfortunately, very little is known about dog mutation rates, and estimates of mammalian mutation rates range from $0.22x10^{-8}$ per year (i.e., $0.66x10^{-8}$ per generation) [31] to $1.8x10^{-8}$ per generation [32]. Considering this wide range expands the credible interval for the divergence time of dogs and wolves from 14–16 kya to 11–34 kya. Importantly, our study was able to eliminate much of the uncertainty in the mutation-scaled divergence time (CI: $0.46x10^{-4}$–$0.53x10^{-4}$), leaving the mutation rate as the dominant source of uncertainty in dating the origin of dogs.

The divergence time between dogs and wolves provides an estimated upper bound for the time of domestication. We can also estimate a lower bound as the divergence time between the Dingo and the population ancestral to Basenji and Boxer, which we infer at 13 kya (CI: 11–12 kya, 9–25 kya assuming a range of mutation rates). Thus, our demographic analysis strongly suggests that domestication occurred between about 11 and 16 kya (9–34 kya with mutation rate uncertainty), which would place it prior to the adoption of extensive agriculture by humans. This finding is consistent with the fossil record, but it raises questions regarding the hypothesis that the advent of agriculture created a novel niche that was the driving force in dog domestication [15]. Our examination of AMY2B confirmed previously reported high copy numbers across almost all dog breeds [15]. However, we also found variation in copy numbers across wolf populations, and low copy numbers in dog lineages that are not associated with agrarian societies (Dingo and Husky). This suggests a more complicated history of the

high copy number variants of AMY2B, which likely existed already as standing variation in early domestic dogs, but expanded more recently with the development of large agriculturally based civilizations in the Middle East, Europe and Eastern Asia.

Overall, the genomes sequenced in this study reveal a dynamic and complex genetic history interrelating dogs and wolves. One question that remains unanswered is that of the geographic origin of dogs and the wolf lineage most closely related to them. Our analysis suggests that none of the sampled wolf populations is more closely related to dogs than any of the others, and that dogs diverged from wolves at about the same time that the sampled wolf populations diverged from each other (Figures 3-5A, 3-5C). One possible implication of this finding is that a more closely related wolf population exists today, but was not represented by our samples. We consider this unlikely, as we sampled the three major putative domestication regions, and previous SNP array studies demonstrate that wolf populations are only weakly differentiated, indicating that the wolves we sampled should serve as good proxies for wolves in each broad geographic region [10].

Another alternative is that the wolf population (or populations) from which dogs originated has gone extinct and the current wolf diversity from each region represents novel younger wolf lineages, as suggested by their recent divergence from each other (Figure 3-5A). Our inference that wolves have gone through bottlenecks across Eurasia (Figures 3-4B, 3-5A) suggests a dynamic period for wolf populations over the last 20,000 years and that extinction of particular lineages is not inconceivable. Indeed, several external lines of evidence provide support for substantial turnover in wolf lineages. For example, ancient DNA, isotope, and morphologic evidence identify a divergent North American Late Pleistocene wolf [33] and in Eurasia, similarly distinct wolves exist in the early archaeological record in Northern Europe and Russia, 15–36kya [3–5]. Presumed changes in available prey (e.g. megafaunal extinctions) as habitats shrunk with the expansion of humans and agriculture also suggest the plausibility of wolf population declines and lineage turnover. A remaining alternative to our inferred population phylogeny is that the basal lineage was absorbed into the three lineages sampled. Such a hypothesis is questionable, though, as it requires there to be enough effective gene flow among the three wolf lineages such that no single lineage today serves best as a proxy for the basal lineage in our analysis. If true, the hypothesis that dogs were originally domesticated from a now-extinct wolf population suggests that ancient DNA studies will play a central role in advancing our understanding of the rapid transition from a large, aggressive carnivore to the omnivorous domestic companion that is a fixture of modern civilization.

Acknowledgments

We thank B. Chin, T. Toy, Z. Chen and the UCLA DNA Microarray Facility for library preparations and sequencing done at UCLA; D. Wegmann for initial development of the VcfAnnotator program used in analyses here; A. Platt for feedback on analyses and manuscript; R. Hefner and The National Collections of Natural History at Tel Aviv University for procuring and access to samples. We thank E. Randi, R. Godinho, and B. Yue for facilitating visits of MG, PMS, and ZF to the lab of RKW. The authors also wish to thank the Australian Native Dog Conservation Society and the Dingo Sanctuary Bargo for generously providing the dingo sample used in this study.

Author Contributions

Conceived and designed the experiments: AHF CDB TTH SFN EAO TMB RKW JN. Performed the experiments: RMS BLG HB OR CV KS ARB HGP CL VT. Analyzed the data: AHF RMS IG EH DODV PMS MG ZF PM BLG OR FH CA VT AS JN. Contributed reagents/ materials/analysis tools: CV EG JK EAO AW RKW. Wrote the paper: AHF RMS IG EH DODV PMS MG ZF PM HB AS TMB RKW JN.

References (style as published)

1. Levi T, Wilmers CC (2012) Wolves-coyotes-foxes: a cascade among carnivores. Ecology 93: 921–929.
2. Ripple WJ, Beschta RL (2012) Trophic cascades in Yellowstone: The first 15 years after wolf reintroduction. Biological Conserv 145: 205–213.
3. Ovodov ND, Crockford SJ, Kuzmin YV, Higham TFG, Hodgins GWL, et al. (2011) A 33,000-Year-Old Incipient Dog from the Altai Mountains of Siberia: Evidence of the Earliest Domestication Disrupted by the Last Glacial Maximum. PLoS ONE 6: e22821.
4. Germonpre M, Laznickova-Galetova M, Sablin MV (2012) Palaeolithic dog skulls at the Gravettian Predmosti site, the Czech Republic. J Archaeol Sci 39: 184–202.
5. Germonpre M, Sablin MV, Stevens RE, Hedges REM, Hofreiter M, et al. (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.

6. Larson G, Karlsson EK, Perri A, Webster MT, Ho SYW, et al. (2012) Rethinking dog domestication by integrating genetics, archeology, and biogeography. Proc Natl Acad Sci USA 109: 8878–8883.

7. Pang JF, Kluetsch C, Zou XJ, Zhang AB, Luo LY, et al. (2009) mtDNA Data Indicate a Single Origin for Dogs South of Yangtze River, Less Than 16,300 Years Ago, from Numerous Wolves. Mol Biol Evol 26: 2849–2864.

8. Pionnier-Capitan M, Bemilli C, Bodu P, Celerier G, Ferrie JG, et al. (2011) New evidence for Upper Palaeolithic small domestic dogs in South-Western Europe. J Archaeol Sci 38: 2123–2140.

9.Savolainen P, Zhang YP, Luo J, Lundeberg J, Leitner T (2002) Genetic evidence for an East Asian origin of domestic dogs. Science 298: 1610– 1613.

10. vonHoldt BM, Pollinger JP, Lohmueller KE, Han EJ, Parker HG, et al. (2010) Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. Nature 464: 898–902.

11. Boyko AR, Quignon P, Li L, Schoenebeck JJ, Degenhardt JD, et al. (2010) A Simple Genetic Architecture Underlies Morphological Variation in Dogs. PLoS Biol 8:e1000451.

12. Cadieu E, Neff MW, Quignon P, Walsh K, Chase K, et al. (2009) Coat Variation in the Domestic Dog Is Governed by Variants in Three Genes. Science 326: 150–153.

13. Karlsson EK, Baranowska I, Wade CM, Salmon Hillbertz NHC, Zody MC, et al. (2007) Efficient mapping of mendelian traits in dogs through genome-wide association. Nat Genet 39: 1321–1328.

14. Li Y, vonHoldt BM, Reynolds A, Boyko AR, Wayne RK, et al. (2013) Artificial selection on brain expressed genes during the domestication of dog. Mol Biol Evol. doi: 10.1093/molbev/mst088.

15. Axelsson E, Ratnakumar A, Arendt M-J, Maqbool K, Webster MT, et al. (2013) The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature 495: 360–364.

16. Miklósi A (2007) Dog behaviour, evolution, and cognition. Oxford ; New York: Oxford University Press. xiii, 274 p.

17. Savolainen P, Leitner T, Wilton AN, Matisoo-Smith E, Lundeberg J (2004) A detailed picture of the origin of the Australian dingo, obtained from the study of mitochondrial DNA. Proc Natl Acad Sci USA 101: 12387–12390.

18. Durand EY, Patterson N, Reich D, Slatkin M (2011) Testing for Ancient Admixture between Closely Related Populations. Mol Biol Evol 28: 2239–2252.

19. Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A (2011) Bayesian inference of ancient human demography from individual genome sequences. Nat Genet 43: 1031–1034.

20. Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature 475: 493–496.

21. Larson G, Burger J (2013) A population genetics view of animal domestication. Trends Genet 29: 197–205.

22. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, et al. (2005)
Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438: 803–819.

23. Gray MM, Granka JM, Bustamante CD, Sutter NB, Boyko AR, et al. (2009) Linkage Disequilibrium and Demographic History of Wild and Domestic Canids. Genetics 181: 1493–1505.

24. Wang G-D, Zhai WW, Yang H-C, Fan R-X, Cao X, et al. (2013) The genomics of selection in dogs and the parallel evolution between dogs and humans. Nature Commun 4:1860. DOI: 10.1038/ncomms2814.

25. Skoglund P, Gotherstrom A, Jakobsson M (2011) Estimation of Population Divergence Times from Non-Overlapping Genomic Sequences: Examples from Dogs and Wolves. Mol Biol Evol 28: 1505–1517.

26. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, et al. (2010) A Draft Sequence of the Neandertal Genome. Science 328: 710–722.

27. Perini FA, Russo CAM, Schrago CG (2010) The evolution of South American endemic canids: a history of rapid diversification and morphological parallelism. J Evol Biol 23: 311–322.

28. Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet 8: e1002967.

29. Vila C, Seddon J, Ellegren H (2005) Genes of domestic mammals augmented by backcrossing with wild ancestors. Trends Genet 21: 214–218.

30. Boyko AR, Boyko RH, Boyko CM, Parker HG, Castelhano M, et al. (2009) Complex population structure in African village dogs and its implications for inferring dog domestication history. Proc Natl Acad Sci USA 106: 13903–13908.

31. Kumar S, Subramanian S (2002) Mutation rates in mammalian genomes. Proc Natl Acad Sci USA 99: 803–808.

32. Sun JX, Helgason A, Masson G, Ebenesersdottir SS, Li H, et al. (2012) A direct characterization of human mutation based on microsatellites. Nat Genet 44: 1161-+.

33. Leonard JA, Vila C, Fox-Dobbs K, Koch PL, Wayne RK, et al. (2007) Megafaunal extinctions and the disappearance of a specialized wolf ecomorph. Curr Biol 17: 1146–1150.
34. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297–1303.
35. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, et al. (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet 41: 1061–1067.
36. Hormozdiari F, Hach F, Sahinalp SC, Eichler EE, Alkan C (2011) Sensitive and fast mapping of di-base encoded reads. Bioinformatics 27: 1915–1921.
37. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou MM, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034–1050.

Paper IV - Worldwide Patterns of Genomic Variation and Admixture in Gray Wolves

Zhenxin Fan,[1,2,13] **Pedro Silva**,[3,13] Ilan Gronau,[4] Shuoguo Wang,[5,14] Aitor Serres Armero,[6] Rena M. Schweizer,[2] Oscar Ramirez,[7] John Pollinger,[2] Marco Galaverni,[8] Diego Ortega Del-Vecchyo,[9] Lianming Du,[1] Wenping Zhang,[10] Zhihe Zhang,[10] Jinchuan Xing,[5,11] Carles Vilà,[12] Tomas Marques-Bonet,[7,12] Raquel Godinho,[3] Bisong Yue,[1] and Robert K. Wayne[2]

[1]Key Laboratory of Bioresources and Ecoenvironment (Ministry of Education), College of Life Sciences, Sichuan University, Chengdu 610064, People's Republic of China; [2]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, California 90095-1606, USA; [3]CIBIO-UP, University of Porto, Vairão, 4485-661, Portugal; [4]Efi Arazi School of Computer Science, the Herzliya Interdisciplinary Center (IDC), Herzliya 46150, Israel; [5]Department of Genetics, Rutgers, the State University of New Jersey, Piscataway, New Jersey 08854, USA; [6]Institute of Evolutionary Biology (UPF-CSIC), PRBB, 08003 Barcelona, Spain; [7]ICREA at Institute of Evolutionary Biology (UPF-CSIC), PRBB, 08003 Barcelona, Spain; [8]ISPRA, Ozzano dell'Emilia, 40064, Italy; [9]Interdepartmental Program in Bioinformatics, University of California, Los Angeles, California 90095-1606, USA; [10]Sichuan Key Laboratory of Conservation Biology on Endangered Wildlife, Chengdu Research Base of Giant Panda Breeding, Chengdu, Sichuan Province, People's Republic of China, 610081; [11]Human Genetics Institute of New Jersey, Rutgers, the State University of New Jersey, Piscataway, New Jersey 08854, USA; [12]Centro Nacional de Análisis Genómico (CNAG), Parc Científic de Barcelona, 08028 Barcelona, Spain

[13]These authors contributed equally to this work.

[14]Present address: Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA

Corresponding authors: rwayne@ucla.edu, bsyue@scu.edu.cn

Genome Research 26:1-11. doi: 10.1101/gr.197517.115
published online December 17, 2015

Abstract

The gray wolf (*Canis lupus*) is a widely distributed top predator and ancestor of the domestic dog. To address questions about wolf relationships to each other and dogs, we assembled and analyzed a data set of 34 canine genomes. The divergence between New and Old World wolves is the earliest branching event and is followed by the divergence of Old World wolves and dogs, confirming that the dog was domesticated in the Old World. However, no single wolf population is more closely related to dogs, supporting the hypothesis that dogs were derived from an extinct wolf population. All extant wolves have a surprisingly recent common ancestry and experienced a dramatic population decline beginning at least ~30 thousand years ago (kya). We suggest this crisis was related to the colonization of Eurasia by modern human hunter–gatherers, who competed with wolves for limited prey but also domesticated them, leading to a compensatory population expansion of dogs. We found extensive admixture between dogs and wolves, with up to 25% of Eurasian wolf genomes showing signs of dog ancestry. Dogs have influenced the recent history of wolves through admixture and vice versa, potentially enhancing adaptation. Simple scenarios of dog domestication are confounded by admixture, and studies that do not take admixture into account with specific demographic models are problematic.

Introduction

The gray wolf (*Canis lupus*) is a dominant large predator that exerts important top-down effects on biodiversity (Levi and Wilmers 2012; Ripple et al. 2014). The species is widely distributed throughout the Holarctic (including the Nearctic and Palearctic regions), and as many as 32 subspecies have been described (Aggarwal et al. 2003). Gray wolves have an ancient origin, first appearing about 500 thousand years ago (kya) in Eurasia and in North America soon thereafter (Nowak 1979; Kurten and Anderson 1980). Initial studies based on mitochondrial DNA (mtDNA) data suggested that the gray wolf had a complex evolutionary history without clear worldwide phylogeographic structure (e.g., Wayne et al. 1992; Vilà et al. 1999). However, subsequent studies found subpopulation structure related to local environmental characteristics (e.g., Carmichael et al. 2001; Geffen et al. 2004; Pilot et al. 2006, 2010, 2014; Musiani et al. 2007; vonHoldt et al. 2011). Genome-wide approaches using SNP genotyping arrays have confirmed these environmentally related genetic partitions and demonstrated extensive admixture with coyotes and, to a more limited extent, with domestic dogs (Pilot et al. 2010, 2014;

vonHoldt et al. 2010, 2011). Using complete genome sequence data of a wolf from Europe, Israel, and China, Freedman et al. (2014) found an unexpected recent coalescence of ~30 kya, suggesting that wolves existing before that time were phylogenetically distinct, a result supported by genetic, isotopic, and morphologic analyses (Leonard et al. 2007; Thalmann et al. 2013). The wolves from these three regions also suffered a substantial bottleneck that initiated ~15 kya, which was nearly coincident with the Wisconsin glacial maximum (Freedman et al. 2014). However, as inferred from genomic data, Zhang et al. (2014) found that Tibetan wolves experienced earlier and more dramatic population declines perhaps due to the extreme loss of wolf habitat with Late Pleistocene glaciations in the Tibetan Plateau. These findings suggest the recent worldwide history of wolves is complex and needs to be assessed with a fuller sample of genomes from throughout the historic range of the species.

The domestic dog (*Canis lupus familiaris*), a descendant of gray wolves, is the most widely abundant large carnivore (Vilà et al. 1999; Thalmann et al. 2013), but the specific region of origin is controversial. Previous genetic evidence suggested that dogs were domesticated either in the Middle East or East Asia (Savolainen et al. 2002; vonHoldt et al. 2010; Wang et al. 2013). However, a recent study based on ancient mtDNA analysis of dogs and wolves infers an origin in Europe from a now-extinct lineage of gray wolves (Thalmann et al. 2013). This result is consistent with whole-genome analysis, showing that none of the extant wolf lineages from putative domestication centers (Europe, Israel, and China) were more closely related to dogs (Freedman et al. 2014). Very recently, however, these conclusions were questioned by results from an extensive study of SNP genotypes in a worldwide sample of breed and village dogs, which concluded that dogs originated in Central Asia (Shannon et al. 2015). Consequently, we test for alternative regions of origin with a geographically broad sample of gray wolves.

The release of the boxer genome in 2005 (Lindblad-Toh et al. 2005) provided a high-quality dog reference for comparison to wolves and other canids (e.g., Wang et al. 2013; Freedman et al. 2014; Zhang et al. 2014; Koepfli et al. 2015). However, no studies have been performed to investigate population subdivision, demography, and relationships of gray wolves based on whole- genome sequences. In this study, we generate whole genomes of nine individual wolves, one coyote, and one golden jackal at 9–28× coverage using the Illumina HiSeq 2000 platform to geographically complement existing canine sequences. Combined with published genomes, we assemble a data set with 34 canid genomes to (1) assess relationship patterns across the entire geographic range of wolves; (2) affirm their recent demographic decline with a more geographically

extensive sample; (3) assess admixture be- tween dogs and wolves; and (4) explore the possibility of dog domestication outside the Middle East, Europe, and East Asia, which was not addressed in previous studies but is a possibility suggested by new findings (Shannon et al. 2015; Skoglund et al. 2015).

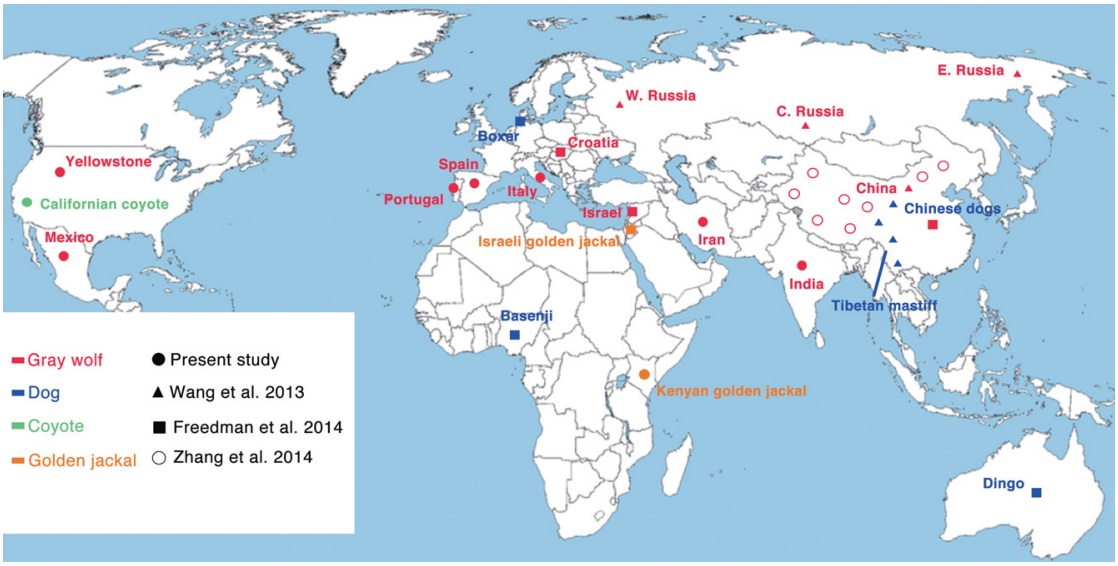


Fig. 3-7: Sample distribution. Solid circles are samples sequenced in this study. Open circles indicate sequences from Zhang et al. (2014). Triangles and boxes indicate sequences from Wang et al. (2013) and Freedman et al. (2014), respectively. Species memberships are indicated by color: gray wolf (red), domestic dog (blue), coyote (green), and golden jackal (yellow). The reference dog genome is from a boxer.

Materials and Methods

Samples and sequencing

We sequenced genomes of dogs, wolves, and other wild canids from Africa, Asia, Europe, the Middle East, and North America. Together with published canid genomes, we generated a final data set with 34 full genome sequences at 9–28× coverage with an average coverage of 29.8× (Fig. 3-7; see Supplemental Material).

Mapping short reads and genotyping

The 100-bp pair-end (PE) short reads of each sample were aligned to the dog genome (CanFam3.1) using Bowtie 2 (Langmead and Salzberg 2012) under the local alignment algorithm with very sensitive model and proper insert sizes of each sample. Default options were used for other parameters. Then, we applied Picard and GATK toolsets (DePristo et al. 2011) to process the alignments to SNP calls. The whole pipeline converted the short reads to BAM format alignment files, and then generated genotype calls in Variant Call Format (VCF). The pipeline is the same as used in our previous

studies (Fan et al. 2014; Freedman et al. 2014; Zhang et al. 2014). We applied a series of data quality filters to improve the quality of genotype calls (see Supplemental Material).

Phylogenetic tree and PCA

A ML tree from whole-genome SNP data was constructed using SNPhylo (Lee et al. 2014). SNPhylo transforms genotype data into a structured data array (Bioconductor gdsfmt) and then generates and aligns SNP sequences and constructs the phylogenetic trees. The program was run with 100 bootstrap repetitions, and only one outgroup was used (Israeli golden jackal) due to the software's internal limitations.

PCA was performed using the pairwise allele-sharing genetic distance. Following vonHoldt et al. (2011), sites exhibiting apparent strong local linkage disequilibrium ($R^2 > 0.5$) were filtered using the –indep option in PLINK (–indep 50 5 0.2) (Purcell et al. 2007). To improve resolution among wolves, the two golden jackals and coyote were removed from the PCA because they were too divergent from dogs and wolves, and their inclusion compressed the scatter among wolves on the first few PCs. The lower coverage genomes (<10-fold; Inner Mongolia wolf 2, Eastern Russian wolf, and Yellowstone wolf 3) were also removed due to their potential high genotype error (Supplemental Fig. S5). Additional PCA was also performed excluding one Tibetan wolf and one Qinghai wolf based on the observation that highland Chinese wolves were similar to one another but were highly divergent from all other wolves (Zhang et al. 2014). Finally, additional PCAs were performed with samples from specific geographic regions, such as Asia or Europe, and including only gray wolves. In both tree analyses and PCAs, we excluded the Yellowstone wolf 2 because it is the offspring of Yellowstone wolf 1 (mother) and Yellowstone wolf 3 (father).

Inference of population size changes through time with PSMC

We used PSMC (Li and Durbin 2011) to infer demographic history. The following parameters were used: numbers of iterations = 25; time interval=64×1; and generation time=3 (Freedman et al. 2014; Zhang et al. 2014). Our previous studies used a mutation rate of 1.0×10^{-8} per generation, a commonly applied value. However, one recent study based on ancient wolf genome sequences estimated that the mutation rate was only 0.4×10^{-8} per generation (Skoglund et al. 2015). Therefore, we used both mutation rates in our study to bracket estimates of divergence time and effective population size.

Runs of homozygosity analysis

ROHs were calculated with PLINK (Purcell et al. 2007). The input data here is the same data in PCA. We then searched for ROHs spanning at least 500 homozygous SNPs in 1 Mb nonoverlapping windows, allowing for a maximum of five heterozygous sites and 30 missing genotypes per window.

Detection of gene flow using the D-statistic

To test whether there was gene flow between wolves and dogs in each geographical region after their divergence, we applied the ABBA-BABA test (D-statistic) between closely related populations by detecting differences in allele sharing between two lineages (P1 and P2) with a third lineage (P3) (Durand et al. 2011; see Supplemental Material). Under the assumption of one gene flow event that is recent compared to the divergence of dogs and wolves, we further used the Durand et al. (2011) equation to estimate the proportion of dog ancestry in the wolf genomes (see Supplemental Material).

Demographic inference with G-PhoCS

The G-PhoCS method (Gronau et al. 2011) was used to estimate divergence times, population sizes, and migration rates. Considering the computational resources required, our analysis focused on a subset of 22 of the 33 genomes that represent all 11 wolf populations, four dog populations, and the Israeli golden jackal as outgroup (Supplemental Table S6). Alignments of the 22 genomes were done over the 13,647 neutral loci designed by Freedman et al. (2014) for use in demographic inference. We ran several analyses (see Supplemental Material) using the standard settings and assumed standard priors for model parameters, as described by Gronau et al. (2011).

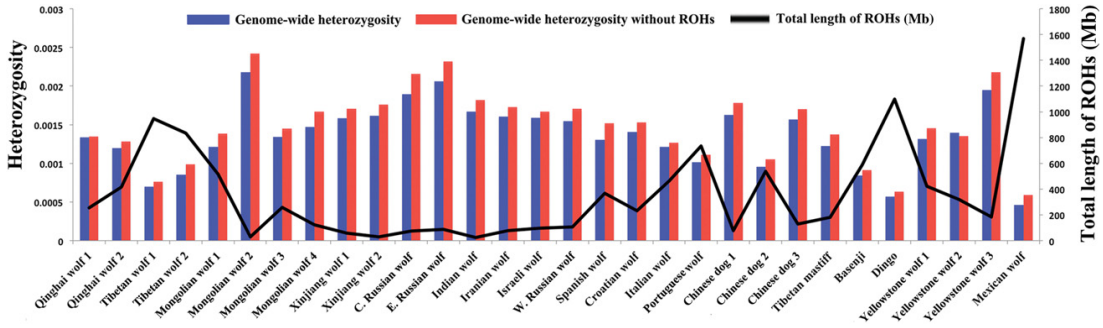


Fig. 3-8: Total length of runs of homozygosity (ROHs) and heterozygosity. The black line is the total length of ROHs (Mb) in each genome, and the blue and red bars are the genome-wide heterozygosity with and without ROHs, respectively.

Results

Genome data and heterozygosity

In this study, we amassed the full genome sequences of 24 wolves, seven dogs (including the reference genome), and three outgroups (Fig. 3-7; Supplemental Table S1). Eleven of the individuals were uniquely sequenced in this study using the HiSeq 2000 platform with the remaining sequences obtained from previous studies (Wang et al. 2013; Freedman et al. 2014; Zhang et al. 2014; Supplemental Material).

To quantify genome-wide heterozygosity, we calculated the number of heterozygous SNPs over all sites (Fig. 3-8; Supplemental Table S2). The Mexican wolf had the lowest autosomal heterozygosity (0.00046), and the two Tibetan wolves also showed very low heterozygosity (0.0007 and 0.00086). Within European wolves, the Portuguese wolf showed the lowest heterozygosity (0.00101). However, the SNP rate was similar in all wolves (Supplemental Table S2). Within dogs, the basenji and dingo had the lowest heterozygosity (<0.001), with dingo having the lowest value of 0.00057 (Supplemental Table S2). We further calculated the heterozygosity of 5 Mb nonoverlapping windows across the 38 autosomal chromosomes (Supplemental Figs. S1–S4). The result confirmed that the low heterozygosity of two Tibetan wolves was evident across their entire genomes (Supplemental Fig. S1). The Mexican wolf exhibited extremely low heterozygosity in about half the chromosomes (Supplemental Fig. S4), and the Portuguese wolf also had very low heterozygosity in more than 15 autosomes (Supplemental Fig. S2). In contrast, the Inner Mongolia wolf 2 had very high heterozygosity across all chromosomes, even higher than other Inner Mongolia wolves, which partly may reflect lower genome coverage and a higher fraction of miscalled sites (Supplemental Figs. S1, S5). We also calculated the heterozygosity of exons and neutral regions (Supplemental Fig. S6).

To avoid the effect of inbreeding in the calculation of heterozygosity, we removed runs of homozygosity (ROH, see below) and recalculated heterozygosity (Fig. 3-8). The results are similar to that of the full data set. For example, the inbred Mexican wolf still had the lowest heterozygosity within wolves, and two Tibetan wolves had higher values than the Mexican wolf but lower than other wolves.

Genome-wide phylogenetic tree and PCAs

Autosomal SNPs were used to construct a maximum likelihood (ML) tree (Fig. 3-9). The topology of ML trees with (Fig. 3-9) and without (Supplemental Fig. S7) the boxer reference genome is consistent with geographical proximity of populations and does not support any specific wolf population as more closely related and possibly ancestral to

domestic dogs. Specifically, all the dogs are monophyletic and define a sister taxon with Eurasian gray wolves that excludes a role for New World wolves in dog origins and suggests that the divergence of the modern Eurasian wolf population is nearly coincident with that of domestic dogs. Among gray wolves, East Asian wolves form a single clade, whereas European and Middle Eastern wolves (including Indian wolf) form a separate grouping. The Middle Eastern wolf is aligned with European rather than Asian wolf sequences (Fig. 3-9). For the New World wolves, two Yellowstone wolves cluster together, and then the Mexican wolf is grouped with them, but the divergence is large, suggesting an ancient separation of the two populations. However, the long branch of the Mexican wolf lineage may reflect the effect of small historic population size as the species went extinct in the wild, which was compounded by an extreme founding bottleneck in the captive population (Fredrickson et al. 2007). Both demographic events would tend to inflate genetic distance values. Nonetheless, this finding supports a previous hypothesis that Mexican wolves represented an early migration into North America (Leonard et al. 2005).

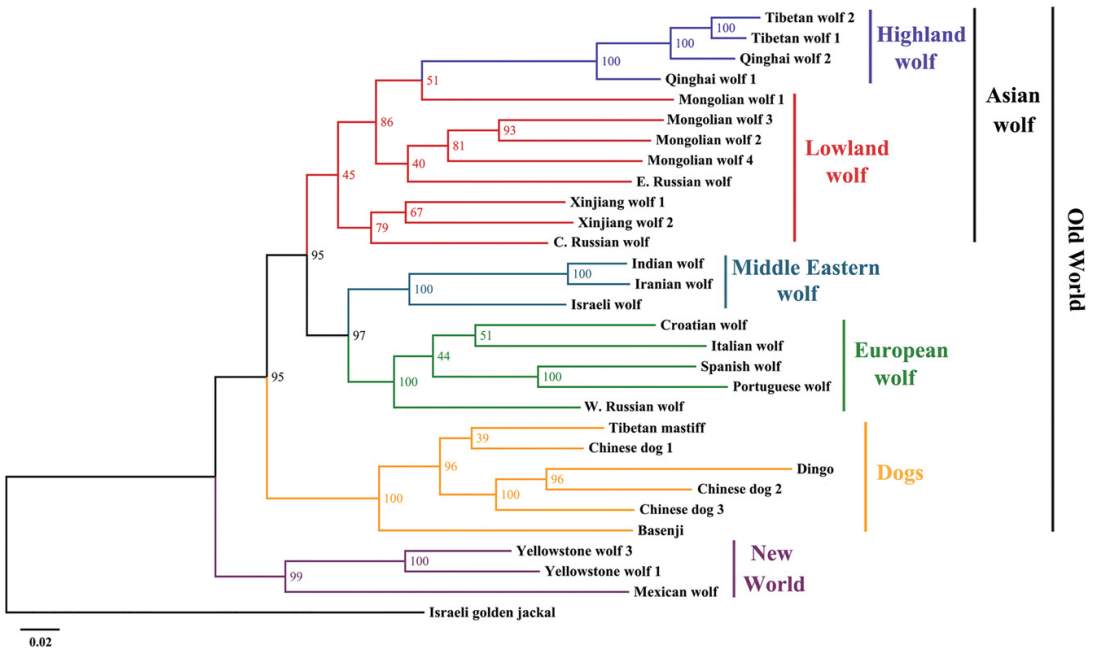


Fig. 3-9: The maximum likelihood tree of 30 sequences. Numbers represent node support inferred from 100 bootstrap repetitions. The reference genome boxer was not included. The Israeli golden jackal is the outgroup.

Principal component analysis (PCA) with LD-pruned data excluding the three outgroups and four wolf sequences (Inner Mongolia wolf 2, Eastern Russian wolf, Yellowstone wolf 2, and Yellowstone wolf 3) showed that PC1 (20.2% of variation) divided the samples into three clusters: domestic dogs; highland Chinese wolves; and other gray wolves (Fig. 3-10A). Further, when two outlier highland wolves were removed

(Tibet wolf 1 and Qinghai wolf 1), dogs were more tightly clustered and separated from all wolves on PC1, whereas PC2 distinguished high altitude wolves and the Central Russian wolf from all other wolves (27.7% of variation of both axes combined) (Fig. 3-10B). PC3 and PC4 of both data sets separate Old and New World wolves, with the Mexican wolf showing the greatest distinction (Fig. 4C,D). The results with only wolves and dogs excluded showed a similar pattern with regard to clusters of wolves (Supplemental Fig. S8). Critically, we found no support for a closer association of Chinese wolves with domestic dogs as suggested by previous studies (Savolainen et al. 2002; Pang et al. 2009; Ding et al. 2012; Wang et al. 2013). We also ran PCAs for different geographical regions (Supplemental Figs. S9–S11). These results are consistent with the tree-based analysis in Figure 3-9, but do not take into account rate variation between lineages that can bias inferences about the actual amount of divergence. Moreover, PCA is a graphical approach that highlights genetic clusters in the data and should not be used to infer genealogical relationships (Novembre and Stephens 2008).

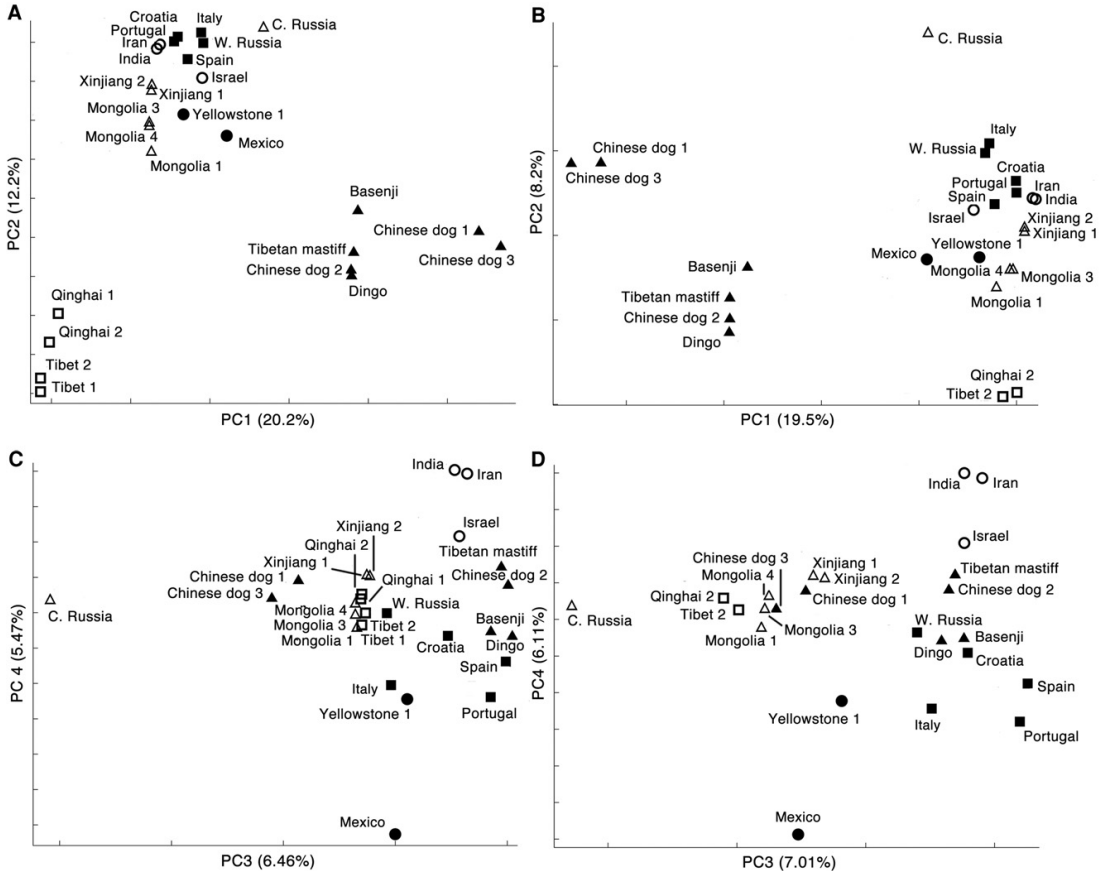


Fig. 3-10: Principal component analyses. (A) PC1 and PC2 of dogs and 20 wolves; (B) PC1 and PC2 of dogs and 18 wolves, excluding the Tibetan wolf 1 and Qinghai wolf 1; (C) PC3 and PC4 of dogs and 20 wolves; (D) PC3 and PC4 of dogs and 18 wolves, excluding the Tibetan wolf 1 and Qinghai wolf 1. (□) Highland Asian wolves; (▵) lowland Asian wolves; (○) Middle Eastern wolves; (■) European wolves; (▲) dogs; (●) North American wolves.

PSMC

The pairwise sequentially Markovian coalescent model (PSMC) was applied to investigate the timing of population-specific demography. Here, we report only the results for the higher mutation rate (Fig. 3-11; Supplemental Fig. S12) to be consistent with Freedman et al. (2014), but consider the results from both rates (1.0×10^{-8} and 0.4×10^{-8} per generation) in the Discussion, as effective size and the timing of population size changes should differ by a factor of approximately 2.5 (Supplemental Figs. S13, S14). All the wolves exhibited similar demographic trajectories until ~80 kya; thereafter, the four highland Chinese wolves showed very different trajectories when compared with all other wolves (Fig. 3-11A). The Tibetan wolves experienced a continuous population decline beginning ~25 to 55 kya and did not experience further population growth; whereas the Qinghai wolf experienced population growth at the same time as the Tibetan wolf bottleneck (Fig. 3-11A). However, caution needs to be used in interpreting these results because they might be explained by ancestral population structure or reflect smoothing across time intervals (Freedman et al. 2014). Other wolves experienced population growth or stag- nation from 25 to 55 kya, which overlaps the Greatest Lake Period (25–40 kya) (Li and Zhu 2001).

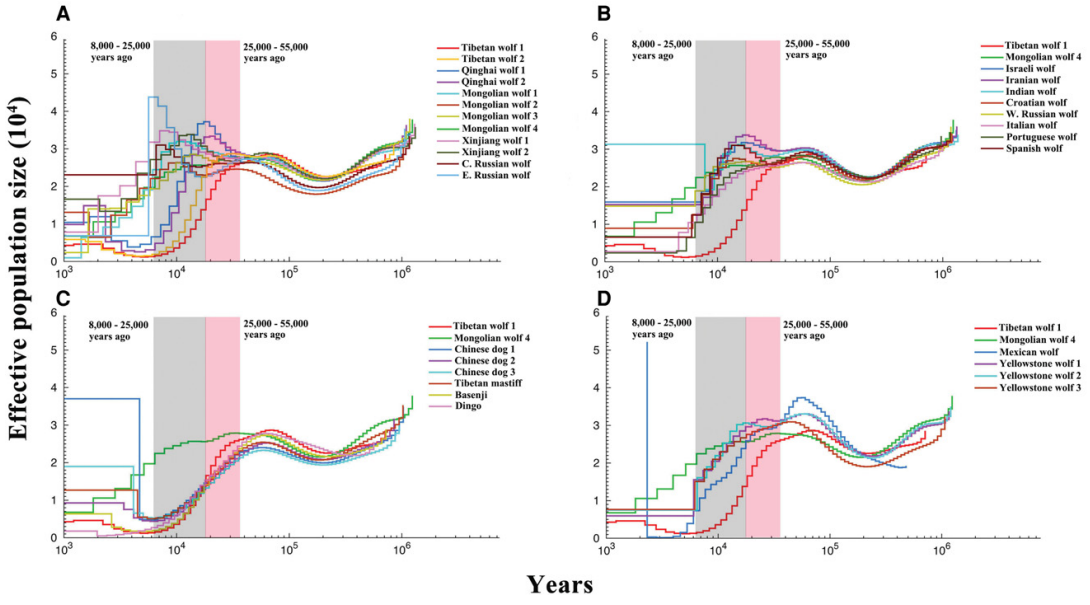


Fig. 3-11: Demographic history inferred using PSMC. Following Freedman et al. (2014) and Zhang et al. (2014), we used a generation time = 3 and a mutation rate = 1.0 × 10−8 per generation. The Tibetan wolf 1 and Inner Mongolia wolf 4 are shown in all the plots for comparison purposes. (A) All the Asian wolves; (B) all the European wolves, Middle Eastern wolves, and Indian wolf; (C) dogs; (D) Mexican wolf and Yellowstone wolves.

PSMC projections showed that the remaining wolves suffered a worldwide decline of two- to threefold beginning ~8 to 25 kya (Fig. 3-11), which is associated with

the end of the last glacial period (10.5–25 kya) and megafaunal extinctions. The Chinese wolves showed the most divergent trajectories with the Tibetan wolves, demonstrating a sharp decline beginning >25 kya and followed by a less precipitous decline in Qinghai wolves. In contrast, the lowland Chinese wolf populations do not initiate a decline until ~10 kya (Fig. 3-11A). The Middle Eastern wolves (Israeli, Iranian, and Indian wolves) and European wolves exhibited slightly different demographic trajectories between 10 and 80 kya (Fig. 3-11B). All these wolves show evidence of a population decline beginning 25 kya. Domestic dogs had similar trajectories and experienced a population decline and demographic divergence from wolves beginning ~50 kya (Fig. 3-11C). The three Yellowstone wolves had concordant trajectories (Fig. 3-11D). However, the Mexican wolf experienced a more severe bottleneck, which may reflect both a recent history of decline and demographic smoothing across the last 10,000-yr interval (see Hedrick et al. 1997; Freedman et al. 2014). The Israeli golden jackal had higher N_e than the Kenya golden jackal, and the California coyote exhibited a different trajectory as expected given its status as an independent lineage (Supplemental Fig. S12). In conclusion, a consistent result across all these trajectories is a decline in population sizes during the period of 8 to 25 kya, coincident with the expansion of modern humans worldwide and the development of technology for capturing large game (Van Valkenburgh et al. 2015).

Autozygosity segments

To assess the history of inbreeding, we quantified genome-wide ROH using PLINK (Fig. 3-8; Supplemental Fig. S15; Purcell et al. 2007). The Mexican wolf had the longest ROH with a total length of 1,569,600 kb (Fig. 3-8), which was consistent with a founding bottleneck and subsequent inbreeding (Hedrick et al. 1997; Fredrickson et al. 2007). In fact, the distribution of ROH in the Mexican wolf was distinct from that of all other wolves, and showed the highest fraction of autozygous long segments (Supplemental Fig. S15d), which suggests very recent inbreeding (e.g., Boyko et al. 2010). The two Tibetan wolves had the longest total length of ROH (947,844 kb and 835,018 kb) and the highest fraction of autozygous segments in Old World wolves, especially at small segment size. This result indicates ancient inbreeding in the Tibetan wolf population (Fig. 3-8; Supplemental Fig. S15a). The Italian wolf had the highest fraction of autozygous segments at smaller ROH sizes among European wolves, whereas the Portuguese wolf had more segments at longer sizes (Supplemental Fig. S15b). This contrasting pattern is consistent with previous genetic analysis, suggesting an ancient population decline in Italian wolves (Lucchini et al. 2004; Pilot et al. 2014) and historical records showing a very recent population decline in Portuguese wolves (Sastre

et al. 2011). Within dogs, dingo and basenji had the greatest ROH (dingo: 1,097,810 kb; basenji: 589,502 kb). They also had a higher fraction of autozygous segments, especially in the size range <4 Mb than the Tibetan mastiff and three Chinese indigenous dogs (Fig. 3-8; Supplemental Fig. S15c), suggesting more ancient inbreeding perhaps in the founding population of dingo that arrived to Australia (>4 kya) and in the origin of the basenji, an ancient breed of domestic dog. These results show that novel demographic insights into population-specific demography are provided by PSMC and ROH analyses, which are consistent with known recent history and past environmental events.

ABBA-BABA

Multiple runs of the ABBA-BABA test were performed to assess gene flow between Old World wolves and dogs (Supplemental Table S3). The results showed that all the European wolves and the Israeli wolf had significant gene flow with basenji and boxer. For the Asian wolves, the two Russian wolves and all the lowland Chinese wolves had significant gene flow with all the Chinese indigenous dogs, Tibetan mastiff, and dingo, whereas the two Tibetan wolves did not show any significant admixture with any dogs (Supplemental Table S3). However, Qinghai wolf 1 showed significant gene flow with two of the three Chinese indigenous dogs, and both Qinghai wolves had gene flow with dingo. The Mexican wolf and Yellowstone wolf did not show any admixture signal with boxer, dingo, or Chinese indigenous dogs (Supplemental Table S3). We note that where admixture is detected from multiple dog samples in one or more wolf populations, it may suggest that gene flow actually occurred from the common ancestor of these dog into a specific wolf population or one that was ancestral to multiple wolf populations.

We estimated the proportion of Chinese indigenous dog ancestry in Asian wolves that had evidence for significant admixture and for which more than one dog defined the comparison pool (Supplemental Table S3, see above; Green et al. 2010; Durand et al. 2011). The proportion of Chinese indigenous dog ancestry in the two Russian wolves varied from 15.3% to 19.52%. The proportion of dog ancestry in the two Xinjiang wolves varied from 9.28% to 11.3%. The average proportion of the dog ancestry in four Inner Mongolia wolves was 10.86%, 12.06%, 13.16%, and 21.59% (Supplemental Table S4). These results suggest substantial dog ancestry in wolf populations worldwide, which is conceivable given the long coexistence of dogs and wild wolf populations (Thalmann et al. 2013; Freedman et al. 2014). The only Old World population not showing any dog ancestry is the Tibetan wolf, which is also the most divergent population in the PCA (Fig. 3-10), suggesting the dog component of wolf genomes may influence patterns of relationships. The high altitude wolf populations also have a very recent history of

exposure to aboriginal dog populations considering that the area was only permanently colonized by humans ~7 kya (Brantingham et al. 2010; Chen et al. 2015).

Regarding the European and Middle Eastern wolves, we used basenji and boxer to estimate the dog ancestry in these wolf genomes (Supplemental Table S4). The proportion of dog ancestry in Israeli wolf, Western Russian wolf, and Spanish wolf is >20%. Of the others, the Portuguese wolf had the smallest proportion at 7.97%, whereas the Croatian wolf had the largest at 13.76% dog ancestry (Supplemental Table S4). These findings indicate a highly variable but substantial dog ancestry in most all extant wolf populations.

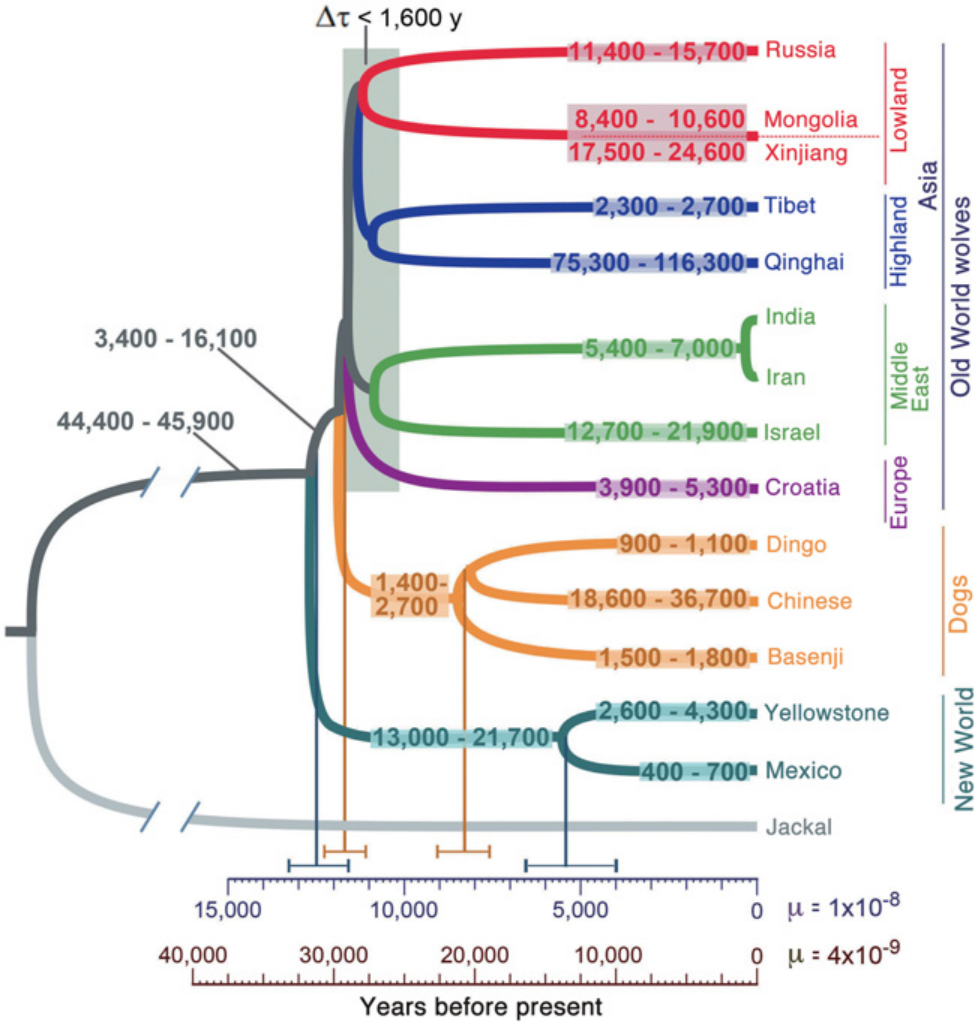


Fig. 3-12: Demographic model inferred using G-PhoCS. Estimates of divergence times and effective population sizes (Ne) inferred by applying a Bayesian demography inference method (G-PhoCS) to sequence data from 13,647 putative neutral loci in a subset of 22 canid genomes (because of limitations in computational power). Estimates were obtained in four separate analyses (Methods; Supplemental Table 6). Ranges of Ne are shown and correspond to 95% Bayesian credible intervals. Estimates are calibrated by assuming a per-generation mutation rate of μ = 10−8. Mean estimates (vertical lines) and ranges corresponding to 95% Bayesian credible intervals are provided at select nodes. Scales are given in units of years by assuming an average generation time of 3 yr and two different mutation rates: μ = 10−8 (dark blue) and μ = 4 × 10−9 (brown). The model also considered gene flow between different population groups (see Table 1).

Demographic inference with G-PhoCS

We used the Generalized Phylogenetic Coalescent Sampler method (G-PhoCS) to infer the demographic history of wolves and dogs, including ancestral population sizes, divergence times, and rates of gene flow (Fig. 3-12). The analysis shows that wolf populations diverged over a relatively short period of time from ~11,000 to 13,000 yr ago (ya), assuming a per-generation mutation rate of $\mu = 1.0 \times 10^{-8}$ and an average generation time of 3 yr (Fig. 3-12). If a slower mutation rate of $\mu = 0.4 \times 10^{-8}$ is used as suggested by Skoglund et al. (2015), this period of time is increased by a factor of 2.5 to 27,500–32,500 ya (see Discussion). The divergence of New and Old World wolves is the oldest of these events at 12,500 ya, followed by divergence of Eastern and Western Eurasian wolves at 11,700 ya. The divergence times between sequences from Europe, the Middle East, and Asia fall within a relatively short period of time of ~1600 yr. New World wolves show an intermediate divergence time of ~5400 ya. We infer dogs diverged from wolves just before the Eurasian wolf population splits (11,700 ya; CI: 11,100–12,300 ya). This divergence time is only 285–1565 yr (95% CI) more recent than the divergence of New World wolves. The tree implies a considerable preancestry of extant dogs of 1400– 2700 yr and a substantial level of divergence among existing dog lineages (dingo, Chinese indigenous dog, and basenji). In contrast, gray wolves have a much more recent common ancestry than expected from their fossil record, and all the population diverged over a narrow time period consistent with a bottleneck followed by a rapid population expansion across Eurasia.

The population ancestral to Old and New World wolves was estimated to have a relatively large effective size of 45,100 individuals (CI: 44,400–45,900), when assuming a per-generation mutation rate of $\mu=1.0\times10^{-8}$. After divergence of Old and New World wolves, both populations experience decline, to 8000 individuals in Old World wolves and 17,300 individuals in New World wolves (Fig. 3-12; Supplemental Table S5). The Tibetan wolf had the smallest N_e within the Old World wolves (2500 individuals), whereas the lowland Chinese wolves had nearly fourfold larger N_e (Fig. 3-12; Supplemental Table S5). The ancestral population of New World wolves had a relatively large effective size of 17,300 individuals, implying a fairly modest bottleneck in the founding of the North American population. However, the two sampled populations have much lower inferred sizes of 3500 individuals for the Yellowstone wolf and 600 individuals for the Mexican wolf. The latter likely reflects a history of decline and extinction in the wild (see Discussion).

G-PhoCS models migration bands allowing a test of admixture from D-statistics. We infer relatively high rates of gene flow (5%–21%; aggregated 95% Bayesian credible intervals) from Chinese indigenous dogs to all Asian wolf populations and significant gene flow (2.4%–7.2%) in the opposite direction only for the lowland Chinese wolves (Table 3-1). Conversely, we find little evidence of admixture with the ancestors of the dingo (migration rates <3% in both directions). For the two highland Chinese wolf populations, G-PhoCS found relatively high rates of gene flow from Chinese indigenous dogs (Tibetan wolf: 5.5%–8.8%; Qinghai wolf: 14.1%–18.9%), although the ABBA-BABA test did not find evidence of admixture. In Western Eurasia, we observe high rates of gene flow between the Israeli wolf and basenji in both directions (4.3%–13.6%) and somewhat lower rates from Croatian wolf to basenji (1.3%–6.4%). Our findings of admixture between Chinese dogs and Asian wolves and Israeli wolf and basenji suggests admixture between the lineages ancestral to these breeds and wolf populations, as their geographic overlap is currently very limited. Consistent with previous results are the high rates of gene flow inferred from the population ancestral to all wolves and dogs into the golden jackal population (11.3%–13.6%) and much lower rates from several sampled wolf populations into the golden jackal population (up to 2.8%) (Freedman et al. 2014). The higher observed value of admixture between golden jackal and the common ancestor of modern wolves suggests an ancient admixture event. Finally, we infer low, but significant, levels of gene flow from the basenji into the Mexican wolf population (1.2%–3.2%), suggesting that like other wolf populations, the New World wolves also experienced admixture with dogs or share a common ancestor with Old World wolves that experienced admixture.

Table 3-1: Migration events detected by G-PhoCS

Wolf population	→Chinese dog	Chinese dog→	→Dingo	Dingo→	→Basenji	Basenji→	→Jackal	Jackal →
Inner Mongolian	5.9 (4.8–7.2)[a]	15.3 (12.1–16.9)[a]	0 (0–0.2)[a]	0.1 (0–0.7)[a]	—	—	0 (0–0.4)[a]	0.1 (0–0.3)[a]
Xinjiang	3.1 (2.4–4.0)[b]	10 (8.6–11.1)[b]	0.3 (0–1.7)[b]	1.4 (0.8–2.2)[b]	—	—	2 (1–2.9)[b]	0.2 (0.1–0.4)[b]
Tibetan	0 (0–0.2)[a]	6.8 (5.5–8.8)[a]	1.8 (0–2.8)[a]	0 (0–0)[a]	—	—	0 (0–0.1)[a]	0.3 (0.2–0.5)[a]
Qinghai	0 (0–0.1)[b]	16.1 (14.1–18.9)[a]	0.3 (0–3.0)[a]	0 (0–0.1)[a]	—	—	0.1 (0–0.6)[a]	0.4 (0.2–0.6)[a]
Russian	0.9 (0–2.8)[a]	18.7 (15.7–20.8)[a]	0.2 (0–1.4)[a]	0.1 (0–0.7)[a]	—	—	1.6 (0.6–2.8)[a]	0.2 (0.1–0.4)[a]
Croatian	—	—	0.3 (0–1.9)[c]	1.2 (0–2.5)[c]	3.7 (1.3–6.4)[c]	0.2 (0–1.2)[c]	0.1 (0–1.3)[c]	0 (0–0.1)[c]
Israeli	—	—	0.1 (0–1)[c]	0.1 (0–0.8)[c]	8 (4.3–13.1)[c]	11.2 (8.7–13.6)[c]	0.7 (0–2.5)[c]	0.2 (0–0.7)[c]
Iranian	—	—	0.1 (0–0.6)[c]	0 (0–0.2)[c]	0.1 (0–0.8)[c]	0 (0–0.5)[c]	0 (0–0.5)[c]	0 (0–0.2)[c]
Indian	—	—	0.1 (0–1.5)[c]	0 (0–0.7)[c]	0.1 (0–0.7)[c]	0.2 (0–2.2)[c]	0.1 (0–0.7)[c]	0.9 (0–4.4)[c]
Mexican	—	—	—	0.4 (0–1.5)[d]	—	2.2 (1.2–3.2)[d]	0 (0–0.2)[d]	0.2 (0–0.5)[d]
Yellowstone	—	—	—	0.1 (0–0.6)[d]	—	0 (0–0.2)[d]	0 (0–0.3)[d]	0.4 (0.2–0.6)[d]
Other migration bands								
Israeli → Croatian				0.1 (0–0.8)[d]				
Croatian → Israeli				0 (0–0.4)[d]				
Tibetan → Inner Mongolian				5 (3.1–6.8)[d]				
Inner Mongolian → Tibetan				0.3 (0–1.3)[d]				
Dog/Wolf ancestor → Jackal				11.9 (11.3–13.6)[d]				
Jackal → Dog/Wolf ancestor				0 (0-0)[d]				

Numbers are the migration rates (total rate %).
[a]Estimated in "Asian" run of G-PhoCS with Inner Mongolian wolf.
[b]Estimated in "Asian" run of G-PhoCS with Xinjiang wolf.
[c]Estimated in "European" run of G-PhoCS.
[d]Estimated in "Global" run of G-PhoCS.

Discussion

Genetic diversity and relationships of Old and New World wolves

Analysis of complete genome sequence data adds considerable resolution to the evolutionary relationships of gray wolves and domestic dogs. First, the genome-wide phylogenetic tree shows that the earliest split was between New and Old World wolves, which was followed by divergence between Old World wolves and dogs (Fig. 3-9). This result confirms dogs were domesticated in the Old World. In addition, the finding that no single wolf population is more closely clustered with domestic dogs supports the hypothesis that dogs were derived from a now extinct population of Late Pleistocene wolf (Thalmann et al. 2013; Freedman et al. 2014). However, the divergence time suggested by G-PhoCS (11,700 ya; CI: 11,100–12,300 ya) is more recent than estimates based on ancient DNA analysis of early dogs and wolves (27,000 ya) (Thalmann et al. 2013). These differences might be caused by inflated mutation rates in the neutral regions used in this study, undetected admixture with dogs, or other assumptions of the underlying G-PhoCS model. The existence of dog fossils older than this recent divergence date, and confirmed by mtDNA sequence data, supports a more ancient origination (Thalmann et al. 2013; Skoglund et al. 2015). In fact, if the mutation rates associated with Skoglund et al. (2015) are used, the divergence time increases to ~29 kya, a value close to their estimate of 27 kya (Fig. 3-12). Finally, within the Old World clade, wolf and dog represent sister taxa. Therefore, suggestions that the dog or dingo are a separate species (*Canis familiaris*) (e.g., Crowther et al. 2014) would cause gray wolves to be a polyphetic taxon; and consequently, our results support dogs as a divergent subspecies of the wolf. This result has societal significance as legislation in some countries and regional governments consider wolves and dogs as distinct species restricting the possession, interbreeding, or the use of vaccines and medications in wolves or dog–wolf hybrids if they have only been approved for use in dogs. In this sense, analysis of evolutionary history informs law and veterinary practice, as dog lineages are nearly as distinct from one another as wolves are from dogs, and the justification for treating dogs and wolves differently is questionable.

The evolutionary tree (Fig. 3-9) and PCA (Fig. 3-10) show that the Mexican wolf is a divergent form of gray wolf, suggesting it is a remnant of an early invasion into North America (García-Moreno et al. 1996; Leonard et al. 2005; vonHoldt et al. 2011) and contradicting suggestions that it is not a distinct subspecies (Cronin et al. 2015). The ROH and genome-wide heterozygosity results (Fig. 3-8) also showed that the Mexican wolf is a highly inbred population (vonHoldt et al. 2011). The subspecies had the smallest

effective population size of only 600 individuals in the sampled wolf populations (Fig. 3-12; Supplemental Table S5). Further, the high long-range ROH in the Mexican wolf implies a long-term decline, followed by a small founding population and inbreeding in the captive population (Hedrick et al. 1997; Fredrickson et al. 2007). These results justify immediate conservation actions to protect this endangered and distinct wolf lineage. Further, population numbers should be increased through captive breeding and in situ conservation to prevent additional genetic erosion. Currently, such efforts have been hindered by the lack of an informed management plan (Wayne and Hedrick 2011; Hendricks et al. 2016).

The Tibetan wolf was found to be the most highly divergent Old World wolf, given its distinct position in the phylogenetic tree (Fig. 3-9) and the PCA plot (Fig. 3-10). It also exhibited extremely low heterozygosity (Fig. 3-8; Supplemental Figs. S1–S4), suggesting that it experienced a historical bottleneck, and only recently recolonized much of the Tibetan Plateau. Indeed, PSMC revealed that the Tibetan wolf suffered a substantial population bottleneck that began ~55 kya (mutation rate 1.0×10^{-8}) or >100 kya, assuming a slower mutation rate, and then declined to the present day (Fig. 3-11A). Notably, all other wolves showed evidence of growth during the Greatest Lake Period from ~25 to 55 kya (Fig. 3-11). The severe habitat loss during glaciations probably contributed to the dramatic population decline of the Tibetan wolf between 10 and 55 kya (Xu and Shen 1995; Yi et al. 2005; Clark et al. 2009; Chevalier et al. 2011; Heyman 2014; Zhang et al. 2014). In addition, both archaeological and genetic analysis suggest that the first colonization might be as early as 30 kya (Aldenderfer 2011), and little evidence exists for permanent human occupation before 7 kya (Brantingham et al. 2010; Chen et al. 2015). Therefore, the appearance of human settlements may have contributed to the decline of Tibetan wolf population but did not initiate the population bottleneck more than 50 kya. Finally, Tibetan wolves had the longest total length of ROHs of the Old World wolves (Fig. 3-8), and a large proportion of their ROHs are in relatively short segments (Supplemental Fig. S15), which suggests that it experienced ancient inbreeding. Moreover, the ABBA-BABA test did not detect substantial gene flow between Tibetan wolf and dogs, suggesting dog admixture did not contribute to ROH (see discussion below). In summary, we suggest that the unique high altitude environment and history of the Tibetan Plateau made wolves there more susceptible to habitat loss, genetic isolation, and allowed for local adaptation. Consequently, these conditions resulted in the evolution of the most distinct wolf population in the Old World.

Geographical structure is evident within Old World wolves (Fig. 3-9). Previously, analysis of Eurasian wolves with mtDNA control region sequences did not reveal any

distinct genetic partitions and suggested modern wolves originated over 250 kya (Vilà et al. 1999). However, European and Middle Eastern partitions were apparent in genome-wide SNP data (vonHoldt et al. 2011). The one-million-year divergence time between wolves and coyotes used previously was based on fossil occurrence data; and given the dynamics of morphological turnover in the coyote lineage (Meachen and Samuels 2012), first occurrence of coyote-like specimens may not accurately reveal the ancestry of modern forms. Our results and those from previous studies (Freedman et al. 2014; Koepfli et al. 2015; Skoglund et al. 2015) suggest the one-million-year divergence time may be inflated by a factor of 20 or more, and modern Eurasian wolves coalesce ~13 kya or ~32.5 kya, the latter using the slower mutation rate from Skoglund et al. (2015). Importantly, the slower rate leads to divergence dates more consistent with the presence of ancient dog fossils well before 15,000 yr ago. Nonetheless, the Skoglund et al. rate needs additional confirmation because it is based on a single fossil specimen with only onefold sequencing depth and used only a subset of DNA sequence to calculate the rate (Skoglund et al. 2015). Thus, until more direct measurements of mutation rates become available, fossil calibration will remain the main source of uncertainty in the timing of key events in canid evolutionary history.

The PSMC results revealed that all wolves shared a similar trajectory before ~100–125 kya (mutation rate 0.4×10^{-8}) or ~30–50 kya (mutation rate 1.0×10^{-8}). In combination, the dating and PSMC results suggested that over the last million years, numerous wolf-like forms existed but that turnover was high, and modern wolves were not the lineal ancestors of dogs (Leonard et al. 2005; Thalmann et al. 2013; Freedman et al. 2014). Indeed, the population size of the Croatian wolf reduced about 10-fold compared to the wolf ancestor, and Yellowstone wolf and Mexican wolf also reduced five- to 28-fold (Fig. 3-12; Supplemental Table S5). This pattern of population reduction and turnover also is supported by recent mtDNA sequence analysis of modern and ancient wolves from the Last Glacial Maximum (Leonard et al. 2007; Pilot et al. 2010; Thalmann et al. 2013) and the dynamic pattern of turnover in other large carnivores such as brown and polar bears, hyenas, and lions as inferred from genetic data (Miller et al. 2012; Cho et al. 2013).

Finally, even assuming a slower mutation rate, our results imply a remarkably recent coalescence of extant wolves several hundred thousand years after the appearance of wolf-like canids (Wayne and Ostrander 2007). Both slow and fast mutation rate estimates are consistent with the possibility that modern humans impacted the demography of gray wolves as they colonized Eurasia, encountered wolves, domesticated some, and possibly caused the decline of others. Humans are the most

effective competitor of large carnivores and could have readily removed them from ecosystems as they do today. Additionally, the presence of large domestic dogs may have accelerated the rate of decline of carnivores that competed with humans (e.g., Shipman 2015). Our results imply that the effect of humans on large predators may have preceded the megafaunal extinctions ~10 kya and may represent one of the earliest anthropogenic causes of decline in animal populations.

Admixture and relationships to domestic dogs

None of our wolf sequences cluster exclusively with domestic dogs, supporting the hypothesis based on only three wolf genomes (Freedman et al. 2014) and ancient DNA (Thalmann et al. 2013) that the immediate gray wolf ancestor of dogs is now extinct. Nonetheless, modern gray wolves have likely influenced the re- cent history of domestic dogs through admixture. Both the ABBA-BABA tests and G-PhoCS support the notion of extensive ad- mixture between dogs and wolves, with up to 20% of the genome of East Asian wolves showing signs of dog ancestry. We also detected that the genomes of European and Middle Eastern wolves had ~7%–25% dog ancestry. Most of the observed gene flow events have not been reported previously. Interestingly, the two highland wolf populations of the Tibetan Plateau showed no evidence of ad- mixture in the ABBA-BABA tests, but G-PhoCS did infer elevated migration rates from Chinese indigenous dogs into these populations (Tibetan wolf: 5.5%–8.8%; Qinghai wolf: 14.1%–18.9%). Conceivably, this finding may be a result of gene flow from dogs into the population ancestral to all modern wolves, which influenced the distribution of coalescent times but cannot be detected using D-statistics because it similarly affected all wolf populations.

For comparison, using ABBA-BABA tests, it was found that modern humans admixed with Neanderthals over 40 kya, but no more than 5% of the modern human genome could be attributed to admixture (Green et al. 2010), suggesting that wolves and dogs have more extensive and regionally based admixture. As in Neanderthals, admixture may have enhanced adaptation in wolves. For example, admixture of pre-Columbian dogs and wolves in North America transferred the black coat color locus to wolves, conferring greater longevity and resulting in a continent- wide selective sweep (Anderson et al. 2009; Coulson et al. 2011). However, in the North American wolves sampled, no apparent trace of admixture remains elsewhere in the genome (Anderson et al. 2009). The persistent admixture between dogs and wolves suggests a unique mode of evolution in which mutations that occur independently in dogs and wolves, under dramatically different selective regimes, can be shared and potentially accelerate the

process of evolution. Such coupled evolutionary histories may exist in other vertebrate species as well, such as in wild and domestic pigs and in brown bears and polar bears (Groenen et al. 2012; Miller et al. 2012).

Finally, admixture can have a confounding effect on inferences about dog domestication history. Specifically, past inferences about dog origins based on private SNPs shared with dogs vonHoldt et al. 2010), greater genome-wide similarity between Chinese wolves and dogs (Wang et al. 2013), or lower LD (Shannon et al. 2015) may reflect regional admixture with wolves and gene flow among dog populations rather than the geographic origin of domestication. Similarly, highly divergent breeds may have more admixture and wolf ancestry retained in their genome. Potentially, this bias might be removed by applying analytical approaches that excise dog segments from wolf genomes. However, direct tracking of genetic changes in wolves and dogs through ancient DNA analysis may be a more robust approach (Grimm 2015).

Data access

The data generated from this study have been submitted to the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra/) under accession number SRP044399.

Acknowledgments

This work was supported by National Science Foundation (NSF) grant EF-1021397 (R.K.W., R.M.S.), the National Key Technology R&D Program of China 2012BAC01B06 (Z.F., B.S.Y.), ICREA, EMBO YIP 2013 and MICINN BFU2014-55090-P (T.M.B.), National Human Genome Research Institute (NHGRI) grant R00HG005846 (J.X.), UC MEXUS-CONACYT doctoral fellowship 213627 (D.O.D.V.), the Chengdu Giant Panda Breeding Research Foundation CPF-yan-2012-10 (W.Z., Z.Z.), and the grant PRIC from Fundació Barcelona Zoo and Ajuntament de Barcelona (O.R.). We thank Prof. Jeffrey Brantingham at the Department of Anthropology, UCLA, and Dr. Xiaoming Wang at the Natural History Museum of Los Angeles County, Los Angeles, for valuable discussions on the Tibetan Plateau.

References

Aggarwal RK, Ramadevi J, Singh L. 2003. Ancient origin and evolution of the Indian wolf: evidence from mitochondrial DNA typing of wolves from Trans-Himalayan region and Pennisular India. Genome Biol 4: P6.

Aldenderfer M. 2011. Peopling the Tibetan plateau: insights from archaeology. High Alt Med Biol 12: 141–147.

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.

Boyko AR, Quignon P, Li L, Schoenebeck J, Degenhardt JD, Lohmueller KE, Zhao K, Brisbin A, Parker HG, vonHoldt BM, et al. 2010. A simple genetic architecture underlies morphological variation in dogs. PLoS Biol 8: e1000451.

Brantingham PJ, Rhode D, Madsen DB. 2010. Archaeology augments Tibet's genetic history. Science 329: 1467.

Carmichael LE, Nagy JA, Larter NC, Strobeck C. 2001. Prey specialization may influence patterns of gene flow in wolves of the Canadian Northwest. Mol Ecol 10: 2787–2798.

Chen FH, Dong GH, Zhang DJ, Liu XY, Jia X, An CB, Ma MM, Xie YW, Barton L, Ren XY, et al. 2015. Agriculture facilitated permanent human occupation of the Tibetan Plateau after 3600 B.P. Science 347: 248–250.

Chevalier ML, Hilley G, Tapponnier P, Van der Woerd J, Jing LZ, Finkel RC, Ryerson FJ, Li HB, Liu XH. 2011. Constraints on the late Quaternary glaciations in Tibet from cosmogenic exposure ages of moraine surfaces. Quat Sci Rev 30: 528–554.

Cho YS,Hu L,Hou H,Lee H,Xu J,Kwon S,Oh S,Kim HM,Jho S,Kim S,et al. 2013. The tiger genome and comparative analysis with lion and snow leopard genomes. Nat Commun 4: 2433.

Clark PU, Dyke AS, Shakun JD, Carlson AE, Clark J, Wohlfarth B, Mitrovica JX, Hostetler SW, McCabe AM. 2009. The last glacial maximum. Science 325: 710–714.

Coulson T, MacNulty DR, Stahler DR, vonHoldt B, Wayne RK, Smith DW. 2011. Modeling effects of environmental change on wolf population dynamics, trait evolution, and life history. Science 334: 1275–1278.

Cronin MA, Cánovas A, Bannasch DL, Oberbauer AM, Medrano JF. 2015. Wolf subspecies: reply to Weckworth, et al. and Fredrickson, et al. J Hered 106: 417–419.

Crowther MS, Fillios M, Colman N, Letnic M. 2014. An updated description of the Australian dingo (Canis dingo Meyer, 1793). J Zool 293: 192–203. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A frame- work for variation discovery and genotyping using next-generation

DNA sequencing data. Nat Genet 43: 491–498. Ding ZL, Oskarsson M, Ardalan A, Angleby H, Dahlgren LG, Tepeli C,

Kirkness E, Savolainen P, Zhang YP. 2012. Origins of domestic dog in southern East Asia is supported by analysis of Y-chromosome DNA. Heredity 108: 507–514.

Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for ancient ad- mixture between closely related populations. Mol Biol Evol 28: 2239–2252.

Fan Z, Zhao G, Li P, Osada N, Xing J, Yi Y, Du L, Silva P, Wang H, Sakate R, et al. 2014. Whole genome sequencing of Tibetan macaque (Macaca thi- betana) provides new insight into the macaque evolutionary history. Mol Biol Evol 31: 1475–1489.

Fredrickson RJ, Siminski P, Woolf M, Hedrick PW. 2007. Genetic rescue and inbreeding depression in Mexican wolves. Proc Biol Sci 274: 2365–2371. 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.

García-Moreno J, Matocq MD, Roy MS, Geffen E, Wayne RK. 1996. Relationships and genetic purity of the endangered Mexican wolf based on analysis of microsatellite loci. Conserv Biol 10: 376–389.

Geffen E, Anderson MJ, Wayne RK. 2004. Climate and habitat barriers to dispersal in the highly mobile grey wolf. Mol Ecol 13: 2481–2490.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, et al. 2010. A draft sequence of the Neandertal genome. Science 328: 710–722. Grimm D. 2015. How the wolf became the dog. Science 348: 277.

Groenen MA, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, Rogel-Gaillard C, Park C, Milan D, Megens HJ, et al. 2012. Analyses of pig genomes provide insight into porcine demography and evolu- tion. Nature 491: 393–398.

Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A. 2011. Bayesian inference of ancient human demography from individual genome sequences. Nat Genet 43: 1031–1034.

Hedrick PW, Miller PS, Geffen E, Wayne RK. 1997. Genetic evaluation of the three captive Mexican wolf lineages. Zoo Biology 16: 47–69.

Hendricks SA, Clee PR, Harrigan RJ, Pollinger JP, Freedman AH, Callas R, Figura PJ, Wayne RK. 2016. Re-defining historical geographic range in species with sparse records: implications for the Mexican wolf reintro- duction program. Biol Conserv 194: 48–57.

Heyman J. 2014. Paleoglaciation of the Tibetan Plateau and surrounding mountains based on exposure ages and ELA depression estimates. Quat Sci Rev 91: 30–41.

Koepfli KP, Pollinger J, Godinho R, Robinson J, Lea A, Hendricks S, Schweizer RM, Thalmann O, Silva P, Fan Z, et al. 2015. Genome-wide evidence reveals that African and Eurasian golden jackals are distinct species. Curr Biol 25: 2158–2165.

Kurten B, Anderson E. 1980. Pleistocene mammals of North America. Columbia University Press, New York.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357–359.

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.

Leonard JA, Vilà C, Wayne RK. 2005. Legacy lost: genetic variability and population size of extirpated US grey wolves (Canis lupus). Mol Ecol 14: 9–17.

Leonard JA, Vilà C, Fox-Dobbs K, Koch PL, Wayne RK, Van Valkenburgh B. 2007. Megafaunal extinctions and the disappearance of a specialized wolf ecomorph. Curr Biol 17: 1146–1150.

Levi T, Wilmers CC. 2012. Wolves–coyotes–foxes: a cascade among carni- vores. Ecology 93: 921–929.

Li H, Durbin R. 2011. Inference of human population history from individ- ual whole-genome sequences. Nature 475: 493–496.

Li B, Zhu Li. 2001. “Greatest lake period” and its palaeo-environment on the Tibetan Plateau. J Geogr Sci 11: 34–42.

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

Lucchini V, Galov A, Randi E. 2004. Evidence of genetic distinction and long-term population decline in wolves (Canis lupus) in the Italian Apennines. Mol Ecol 13: 523–536.

Meachen JA, Samuels JX. 2012. Evolution in coyotes (Canis latrans) in re- sponse to the megafaunal extinctions. Proc Natl Acad Sci 109: 4191–4196.

Miller W, Schuster SC, Welch AJ, Ratan A, Bedoya-Reina OC, Zhao F, Kim HL, Burhans RC, Drautz DI, Wittekindt NE, et al. 2012. Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. Proc Natl Acad Sci 109: E2382–E2390.

Musiani M, Leonard JA, Cluff HD, Gates CC, Mariani S, Paquet PC, Vilà C, Wayne RK. 2007. Differentiation of tundra/taiga and boreal coniferous forest wolves: genetics, coat colour and association with migratory car- ibou. Mol Ecol 16: 4149–4170.

Novembre J, Stephens M. 2008. Interpreting principal component analyses of spatial population genetic variation. Nat Genet 40: 646–649.

Nowak RM. 1979. North American Quaternary Canis. Monograph of the Museum of Natural History. University of Kansas, Lawrence, KS.

Pang JF, Kluetsch C, Zou XJ, Zhang AB, Luo LY, Angleby H, Ardalan A, Ekström C, Sköllermo A, Lundeberg J, et al. 2009. mtDNA data indicate a single origin for dogs south of Yangtze River, less than 16,300 years ago, from numerous wolves. Mol Biol Evol 26: 2849–2864.

Pilot M, Je ̨drzejewski W, Branicki W, Sidorovich VE, Jedrzejewska B, Stachura K, Funk SM. 2006. Ecological factors influence population ge- netic structure of European grey wolves. Mol Ecol 15: 4533–4553.

Pilot M, Branicki W, Je ̨drzejewski W, Goszczyn´ski J, Je ̨drzejewska 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, Je ̨drzejewska B, Randi E, Je ̨drzejewski W, Sidorovich VE, Ostrander EA, Wayne RK. 2014. Genome-wide signa- tures of population bottlenecks and diversifying selection in European wolves. Heredity 112: 428–442.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. 2007. PLINK: a tool set for whole-ge- nome association and population-based linkage analyses. Am J Hum Genet 81: 559–575.

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 eco- logical effects of the world's largest carnivores. Science 343: 1241484.

Sastre N, Vilà C, Salinas M, Bologov VV, Urios V, Sánchez A, Francino O, Ramírez O. 2011. Signatures of demographic bottlenecks in European wolf populations. Conserv Genet 12: 701–712.

Savolainen P, Zhang YP, Luo J, Lundeberg J, Leitner T. 2002. Genetic evi- dence for an East Asian origin of domestic dogs. Science 298: 1610–1613. Shannon LM, Boyko RH, Castelhano M, Corey E, Hayward JJ, McLean C, White ME, Abi Said M, Anita BA, Bondjengo NI, et al. 2015. Genetic structure in village dogs reveals a Central Asian domestication origin. Proc Natl Acad Sci 112: 13639–13644.

Shipman P. 2015. The invaders: how humans and their dogs drove Neanderthals to extinction. Belknap Press of Harvard University Press, Cambridge, MA.

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.

Thalmann O, Shapiro B, Cui P, Schuenemann VJ, Sawyer SK, Greenfield DL, Germonpré MB, Sablin MV, 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.

Van Valkenburgh B, Hayward MW, Ripple WJ, Meloro C, Roth VL. 2015. The impact of large terrestrial carnivores on Pleistocene ecosystems. Proc Natl Acad Sci. doi: 10.1073/pnas.1502554112.

Vilà C, Amorim IR, Leonard JA, Posada D, Castroviejo J, Petrucci-Fonseca F, Crandall KA, Ellegren H, Wayne RK. 1999. Mitochondrial DNA phylo- geography and population history of the grey wolf Canis lupus. Mol Ecol 8: 2089–2103.

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.

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 12: 1294–1305.

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

Wayne RK, Hedrick P. 2011. Genetics and wolf conservation in the American West: lessons and challenges. Heredity 107: 16–19.

Wayne RK, Ostrander EA. 2007. Lessons learned from the dog genome. Trends Genet 23: 557–567.

Wayne RK, Lehman N, Allard MW, Honeycutt RL. 1992. Mitochondrial DNA variability of the gray wolf: genetic consequences of population decline and habitat fragmentation. Conserv Biol 6: 559–569.

Xu DM, Shen YP. 1995. On ancient ice-sheet and ice age in the Tibetan pla- teau. J Glaciol Geocryol 17: 213–229.

Yi CL, Cui ZJ, Xiong HG. 2005. Numerical period of Quaternary glaciations in China. Quatern Sci 25: 609–611.

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.

Paper V - The Effects of Population Structure and Sampling Scheme on Demographic Inferences from Microsatellite Data: an Empirical Test on the Iberian Wolf Population

Silva P et al.

(manuscript in preparation for submission)

Abstract

Most methods demographic parameter estimation from microsatellite data, such as effective population sizes (N_e), population size changes or more complex demographic histories, assume that samples originate from an isolated and unstructured population. However, this assumption is violated in many natural populations, possibly leading to biased inferences. The Iberian wolf population has been found to present significant levels of genetic population structure, in addition to having a recent history of decline and fragmentation. In this study, we use the Iberian wolf population as a test case to empirically assess the limitations of some widely-used demographic inference methods: N_e estimation by the LD-N_e method, methods to detect past population size change using summary statistics (heterozygosity excess and M ratio), and a full-likelihood method (*MSVAR*) to infer long-term demographic history. Additionally, we use of data simulated under different demographic and sampling scenarios. We find that N_e estimates of the total population varied significantly depending if population structure is ignored. Significant signs of bottlenecks were only found in some of the subpopulations, which might be a consequence of the low power of these tests in structured populations. Demographic history inference using *MSVAR* reveals an old onset of population decline (504 ya: 95% CI 166-1637), suggesting that the Iberian wolf population might have suffered impacts previous to the last two centuries, although these estimates could also be inflated due to the population structure.

Introduction

The patterns of genetic diversity of natural populations are extensively shaped by demography. Understanding demographic processes therefore elucidates the evolutionary history and helps in conceptualizing effective conservation measures for endangered populations. A particularly important parameter in this respect is the effective population size, N_e, defined as the size of an ideal population that has the same rate of genetic change as the observed population (Wright 1931). N_e, in interaction with other forces such as mutation, selection, migration and recombination, determines the genetic variability of a population, which can influence its capacity to survive and adapt to environmental changes (Frankham 2005; Charlesworth 2009). Directly investigating the demography of natural populations requires census data that can be very difficult and time-consuming to obtain. Furthermore, demographic trends may take a long time to become apparent, or changes in population size may have occurred sometime in the past without any accurate census data having been collected. Due to variance in reproductive success, the effective population size is usually also much lower than the census size (Frankham, 1995). Population genetic approaches are therefore powerful alternatives that allow inferences on current and past demographic parameters of interest from the present distribution of genetic variation in the population (Wang 2005; Schwartz et al. 2007; Palstra and Fraser 2012).

Several estimators for contemporary N_e have been developed from different measures of genetic change, the most common of which are the 'inbreeding N_e', related to the common ancestry of alleles in a population with a limited number of breeders, and the 'variance N_e', related to the rate of change in allele frequencies through time (Luikart et al., 2010). The variance Ne is generally more sensitive to population size changes whereas the inbreeding Ne does not change until inbreeding accumulates (Allendorf and Luikart 2007). It is also possible to estimate a 'long-term N_e', based on measures of gene diversity and mutational parameters assumed for a population at equilibrium, and that can differ substantially from contemporary N_e because it reflects the effects of evolutionary forces over very long periods of time (Wang 2005; Charlesworth 2009).

Drastic changes in N_e, namely severe reductions ('population bottlenecks'), can have lasting effects on the survival chance of a natural population: during a bottleneck the rate of inbreeding and loss of genetic variation increase while the effectiveness of selection is decreased, potentially leading to the fixation of deleterious alleles and reducing its adaptive potential to future environmental changes (Frankham 2005). The most widely used genetic bottleneck detection methods are based on quantifying

transitory deviations from expected equilibrium values of various statistics (such as the number of alleles, heterozygosity, or allele size distribution) that occur when the size of a population is severely reduced (Cornuet and Luikart 1996; Luikart and Cornuet 1998; Garza and Williamson 2001). While these methods are relatively simple to implement and have been widely used, they are rather limited in the amount of information they provide. For example, they do not estimate past population sizes before size changes (allowing to evaluate its severity), and are limited to detecting recent size changes without precise estimates about their timing. To estimate these parameters, likelihood-based methods have been developed (e.g. (Beaumont 1999; Storz and Beaumont 2002; Leblois et al. 2014). While more powerful, because they consider all information available in the data as opposed to a single summary statistic, these methods are mathematically more complex and computationally demanding (Marjoram and Tavaré 2006).

The N_e estimation and population size change detection methods described above have similar limitations in the sense that their results can be biased by departures from their respective model assumptions. Populations in these models are usually assumed to be isolated and to conform to a simple Wright-Fisher (WF) population model with constant size, random mating, equal sex ratio, discrete generations and randomly variable reproductive success. The genetic markers used are also usually assumed to mutate according to a given known mutational model. However, natural populations rarely conform to these idealized conditions: inter-population gene flow and intra-population subdivision have been observed in many species, and these assumption violations in particular are known to confound demographic inferences (Wakeley 1999; Nielsen and Beaumont 2009; Mazet et al. 2016). In particular, this can lead to biased N_e estimates (Waples and England 2011; Neel et al. 2013; Gilbert and Whitlock 2015), the inference of spurious bottlenecks or expansions (Williamson-Natesan 2005; Broquet et al. 2010; Paz-Vinas et al. 2013) and misleading demographic histories by likelihood-based methods (Chikhi et al., 2010; Girod et al., 2011). These limitations are especially important in the study of endangered populations, for which genetic-based demographic parameters and trends are often used to inform conservation or management decisions.

In this study, we use the Iberian wolf population as a test case to empirically assess the limitations of some widely-used demographic methods. Iberian wolves (*Canis lupus signatus*) constitute the largest remaining wolf population in Western Europe and are currently isolated from other European populations. At the beginning of the 20th century, wolves were still abundant over the entire Iberian Peninsula, but their distribution was greatly reduced due to direct human persecution and changes in prey abundance, with an all-time low of ~500 individuals in the 1970s (Garzón 1979). The

implementation of legal protection measures led to an increase in range and population size in recent decades, and currently it is estimated to include >2000 individuals in more than 300 packs, distributed mainly in the northwestern part of the peninsula (Álvares et al. 2005; Chapron et al. 2014). Non-withstanding the ability of wolves to disperse over long distances, dispersal rates in this population appear to be relatively low, and a substantial genetic population structure can be discerned (Silva et al. in prep - Paper I). It therefore does not conform to the WF-model assumptions that are made by N_e estimation and population size change methods, as described above, and provides an exceptional opportunity to test the performance of these methods in a real setting. We use microsatellite genotype data from an extensive sample of Iberian wolves, covering most of its distribution area, and information about genetic population structure inferred from Bayesian clustering methods in a companion study (Silva et al. in prep - Paper I) to test the effects of population subdivision, gene flow among subpopulations and sampling scheme on the performance of i) Linkage disequilibrium based N_e estimation methods; ii) the detection of past population size changes using summary statistic based methods (heterozygosity excess and *M* ratio); and iii) the estimation of long-term population trends inferred by a full-likelihood method (*MSVAR*). Additionally, we support our conclusions with the use of data simulated under different demographic and sampling scenarios.

Materials and Methods

Samples and genotypes

Our sample consists of 218 wolves from Portugal and Spain, genotyped at 46 microsatellite loci, as in Silva *et al.*, in prep - Paper I. The sampling spans the whole distribution area of this species in the Iberian Peninsula. Close familiar relationships were avoided by allowing a maximum of two individuals from the same pack.

Silva et al. used Bayesian clustering methods to identify genetic population clusters based on microsatellite genotypes. Taking into account the sample location information, geographically meaningful subpopulations were identified at K=4 and K=11 (K being the number of clusters), that are interpreted to summarize the genetic structure of this population at two different hierarchical levels. Differentiation between subpopulations was moderately high: mean pairwise F_{ST} at K=4 was 0.10 (0.08-0.14.), and at K=11, 0.13 (0.03-0.25). Here we consider the same genetic partitioning scheme, analyzing both K=4 and K=11. Individuals termed ‘dispersants’ in Silva et al. in prep - Paper I correspond to individuals with clear genetic membership to a subpopulation different than the one where they were sampled. Some of these had no clear final

subpopulation. In this study, dispersants are included in their genetic subpopulation of origin, regardless of being sampled a defined final subpopulation or not. Individuals without clear genetic and geographic membership were excluded (3 individuals at K=4, 9 individuals at K=11).

Estimation of contemporary effective population size

We used the linkage disequilibrium (LD) method (Waples and Do 2008), an estimator based on inbreeding N_e, as implemented in *NeEstimator* v2 (Do et al. 2014) to estimate current effective population sizes. This method is based on the deviations of pairwise locus association frequencies from those expected under random mating (gametic/linkage equilibrium) that arise when the sampled individuals descend from a finite number of parents. Alleles with very low frequencies were excluded from all analysis (P_{crit} = 0.01), as recommended by Waples & Do (2010), and a monogamous mating system was assumed. We separately applied the method to each of the 4 or 11 identified subpopulations, as well as to a pooled sample of all 218 individuals.

NeEstimator v2 can also perform the heterozygote excess method of Zhdanova and Pudovkin (2008) and the coancestry method of Nomura (2008), but these methods are more appropriate for very small effective population sizes (< ~30) (Zhdanova and Pudovkin 2008; Luikart et al. 2010), and generally have poorer accuracy (Gilbert and Whitlock 2015). As such, our estimates with these methods gave indeterminate (infinite size) or very low values (<10) with both the local and pooled sampling schemes (results not shown).

Moment-based methods for detecting population bottlenecks

To detect putative signatures of recent population size contractions we used the heterozygote excess method (Cornuet and Luikart 1996; Luikart and Cornuet 1998), implemented in *Bottleneck* 1.2.02 (Piry et al. 1999), and the *M* ratio method (Garza and Williamson 2001). According to population genetic theory, the allele number and frequency distribution for selectively neutral loci result from the balance between mutation and genetic drift. These two bottleneck detection methods are based on the expected transient effects of a significant reduction in effective population size, which increases the loss of alleles by drift. The heterozygote excess method refers to the relatively lower allelic diversity at a locus than that expected from the observed heterozygosity in a population at equilibrium, given that in a bottleneck low frequency alleles tend to be lost at a faster rate than heterozygosity. The *M* ratio method relates to the ratio between the number of alleles and the range of allele sizes: during a population

bottleneck, the number of alleles at a locus is reduced faster than their size range, since the latter is only reduced if the lost allele is the smallest or the largest. The *M* ratio is expected to continue to decrease after the size reduction if the population continues to be small, and therefore this method is expected to be informative about size reductions that occurred longer ago than other methods.

In *Bottleneck*, a two phase model of mutation (TPM) was assumed, with 90% of the changes being single steps and a variance among multiple steps of 12. Significance of the results was based on the Wilcoxon sign ranked test (a one-tailed test for heterozygosity excess), with p=0.05. We calculated the *M* ratio for each subpopulation using the program *M_P_val* (Garza and Williamson 2001) and compared it with critical values (M_C) that would be expected in populations at mutation-drift equilibrium; a population size contraction is suggested if $M < M_C$. M_C values were calculated with *Critical_M* (Garza and Williamson 2001) assuming the suggested mutational parameters: a proportion of single-step mutations of 90% (p_g=0.1) with an average size (Δ_g) of 3.5 and a mutation rate (μ) of 5×10^{-4}/locus/generation. Additionally we assumed varying values of θ ($=4N_e\mu$) for the equilibrium populations of 0.02, 0.2 and 2 (corresponding to N_e values of 10, 100 and 1000 for the mentioned mutation rate, respectively). Both bottleneck detection methods were applied to the total pooled sample and to each of the K=4 and K=11 subpopulation.

Likelihood-based method for detecting population size changes

We used *MSVAR* 1.3 (Beaumont 1999; Storz and Beaumont 2002) to detect past population size changes and to estimate related demographic parameters of interest. *MSVAR* is a Bayesian likelihood-based method that assumes a model of a population that has undergone a linear or exponential size change at some point in the past, and estimates the posterior distribution of several demographic parameters such as the current and past population sizes (*N0* and *N1*, respectively), and the time of the onset of population size change (*xa*). All model parameters are allowed to vary between loci, including the mutation rate (μ), and are drawn from log-normal prior distributions. Mean values for the prior distributions are specified by normal (hyperprior) distributions with means α, and standard deviations (SDs) σ; similarly, SDs of prior distributions follow distributions with means β, and SDs τ. For our analyses, we set (all values are presented log-transformed): $\alpha_{N0} = \alpha_{N1} = 3$ (no prior information whether the population declined or expanded), $\alpha_{xa} = 2$, $\alpha_{\mu} = -3.3$ (corresponding to a mutation rate of 5×10^{-4}/locus/generation), $\beta_{N0} = \beta_{N1} = \beta_{xa} = \beta_{\mu} = 0$, $\sigma_{N0} = \sigma_{N1} = \sigma_{xa} = \sigma_{\mu} = 0.5$, $\tau_{N0} = \tau_{N1} = \tau_{xa} = 0.5$,

and $\tau_\mu = 2$. The average generation time was assumed to be 3 years (Mech and Boitani 2003).

We ran 5 independent MCMC chains for each of the two population size change models implemented in *MSVAR* (exponential and linear size change) for 1×10^9 steps and a thinning interval of 10 thousand steps. The mixing of each chain was assessed visually in *Tracer* v1.6.0 (Rambaut et al. 2014), and the convergence of the independent chains was evaluated by calculating the Brooks, Gelman & Rubin statistic in *BOA* v1.1 (Smith 2007), with values <1.1 considered as representing good convergence. Final parameter estimate means and 95% high posterior density (HPD) intervals were obtained with *Tracer* by combining the last 50% of data points of the 5 chains for each model (total of 25,000 data points for each of the two models). Due to the high computational requirements of this method, we only used the pooled sample of all individuals, and local samples of four subpopulations for K=11: Alto Minho (n=12), S Douro (n=6), Castilla y Léon (n=14) and W Galicia (n=54).

Simulations

All the N_e estimation and population size change methods previously described (LD-N_e, heterozygote excess test, *M* ratio test, and *MSVAR*) assume that samples are derived from a single isolated WF population. Our pooled sample of 218 individuals likely does not meet this condition, given the described genetic population structure of the Iberian wolf population (Silva et al., unpublished); on the other hand, local samples of individual subpopulations also depart from the idealized condition because some level of gene flow with each other does exist. We therefore tested these methods by repeating the analyses on simulated data. Conditions similar to the ones found in the Iberian wolf population were replicated by simulating 10 populations of 50 diploid individuals each, with equal sex distribution, within an island model of migration using *EASYPOP* v2.01 (Balloux 2001). We used three different migration rates (proportion of migrants per generation, *m*), corresponding to the range of differentiation values found between our real subpopulations: m=0.1, 0.03 and 0.015, which correspond approximately to F_{ST} values of 0.05, 0.14 and 0.25, respectively, according to the relation $F_{ST} \approx 1/(1+4Nm)$. As a control, we also simulated isolated subpopulations (m=0). A monogamous mating system was assumed, with no extra pair matings. Forty-six loci were simulated under a step-wise mutation model (SSM) with a mutation rate of 5×10^{-4}/locus/generation and a maximum of 10 alleles. Populations were initialized with maximal variability, and simulation proceeded during 10,000 generations, after which all individuals were

sampled, and pairwise population F_{ST} values were verified to have stabilized around the expected values. Each simulation was replicated 5 times.

To simulate sampling schemes similar to the ones used for the real data, 20 individuals were randomly chosen from each population (local sampling); for the pooled sample, these individuals were simply pooled together. Each of the 5 simulation replicates for each migration rate therefore yielded 10 local samples of 20 individuals each, and one pooled sample of 200 individuals. We also assessed the effect of sampling by using the full sample of each simulation replicate (10 local samples of 50 individuals each, and one pooled sample of 500 individuals).

Given that loci were simulated under the SSM, this mutation model was used in the respective tests in *Bottleneck*, instead of the TPM used for the real data. Likewise, M_C values were calculated using p_g=0 (only single-step mutations) and the known effective population size used in the simulations (50). Given the high computational time burden for running *MSVAR*, we ran 3 independent MCMC chains for one simulated dataset of 200 pooled individuals for each of the three migration rates.

Results

Current effective population sizes

Current effective population size estimates for the Iberian wolf population as a whole varied significantly depending on the scale of sampling used (table 3-2). Pooling all 218 samples, the total N_e was estimated at 102 individuals (95% CI: 99-106). Effective size estimates at K=4 vary between 55 individuals (95% CI: 49.9-60.1) for the Asturias subpopulation, and 112 individuals (95% CI: 98.8-127.7) for the Galician subpopulation. At K=11, a great variability in N_e was found between subpopulations. The smallest subpopulation are E Asturias and S Douro (17 individuals, 95% CIs: 14.3-19.2 and 10.4-33.7, respectively), and the largest is W Galicia (101 individuals, 95% CI: 91.2-113.3).

Our simulations showed that pooling together a sample derived from populations with low gene flow (and therefore, moderately differentiated) can lead to erroneous inferences (figure 3-13a). For the lowest migration rate tested (m=0.015), the estimated N_e is closer to the N_e of each individual population (50), independently of the sampling scheme used. N_e values are still significantly under-estimated with m=0.1 (mean N_e of 322 and 351 for a total and partial samples, respectively).

Table 3-2: Current effective population size estimates at different sampling scales: total Iberian Peninsula and local subpopulation samples at K=4 and K=11. Estimates were performed using the LD method in NeEstimator v2.

Population	**N**	**N_e (95% CI)**
total Iberian Peninsula	218	102.3 (99.0-105.9)
4 subpopulations		
Asturias	52	54.7 (49.9-60.1)
Portugal	48	93.0 (81.7-107.1)
Castilla y León	55	71.6 (63.8-81.0)
Galicia	60	111.8 (98.8-127.7)
11 subpopulations		
Alto Minho	12	59.2 (39.9-105.9)
W Trás-os-Montes	11	53.2 (37.2-87.8
E Asturias	18	16.5 (14.3-19.2)
E Trás-os-Montes	18	71.4 (56.2-95.9)
SE Asturias	23	55.6 (47.2-66.7)
W Asturias	12	46.3 (33.4-71.8)
Castilla y León	14	42.7 (32.8-59.3)
S Douro	6	17.0 (10.4-33.7)
W Galicia	54	101.3 (91.2-113.3)
C Asturias	31	39.9 (35.8-44.9)
E Galicia	10	36.9 (26.7-56.8)

On the other hand, local N_e estimates are only slightly affected when the assumption of isolation (i.e., no gene flow with external populations) is violated (figure 3-13b), at least for the migration rates tested here. For the three migration rates tested (*m*=0.015, 0.03 and 0.1), corresponding to differentiation values from low to high (Fst=0.05, 0.14 and 0.25), similar to the ones measures between our real subpopulations (Silva et al. in prep - Paper I), N_e estimates were always close to the simulated value of 50, with the exception of the case when a full sample was used for the highest migration rate (mean N_e of 77).

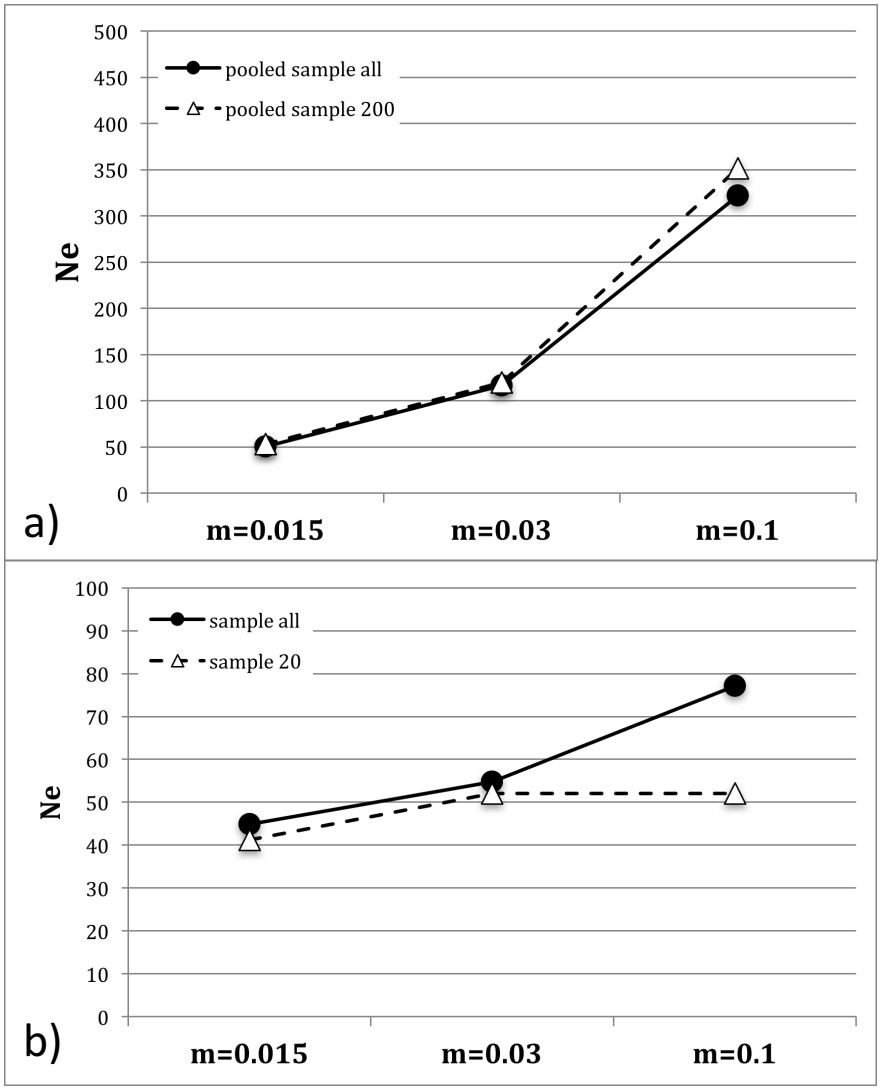


Fig. 3-13: Effect of the migration rate and sampling scheme on the effective population size estimates of a) a pooled sample and b) local samples of 10 populations simulated under an island model with varying degrees of migration (*m*). Each of the 10 populations was simulated with *N*=50. Values refer to the harmonic mean of a total of 5 (a) or 50 (b) estimates across 5 simulation replicates. Sampling schemes correspond to using all individuals ('sample all') or only 40% of individuals ('sample 200' or 'sample 20').

Signals of bottleneck

When using all samples of the Iberian wolf population, no significant signature of recent size change was detected with either the heterozygote excess or the *M* ratio method (table 3-3). Signals of bottlenecks were detected however in four subpopulations using the heterozygote excess method when the local sampling scheme of K=11 was used: W Asturias, E Galicia, Castilla y León and E Trás-os-Montes. Subpopulations from Alto Minho and S Douro showed the lowest *M* ratios, but these values were not smaller than the respective critical thresholds.

No population size change was modeled in our simulated datasets, however some of these populations were still found to depart from equilibrium conditions according to *Bottleneck* (Supplementary Table S1). In local samples, deviations towards either heterozygote excess or deficiency were found, meaning that under naïve conditions, population declines and expansions, respectively, would be inferred for these populations. The proportion of false bottleneck signals appears to decrease with increasing migration rates (20% for m=0.015 vs. 4-6% for m=0.1), while false expansion signals seem to be relatively rarer (~2-4%) and not directly related to the magnitude of migration within the tested range of our simulations. It is also worth noting that even in the case of completely isolated populations (m=0), i.e. when the no migration assumption is not violated, some of the simulated populations presented signals of population declines (~12%) or expansions (~4%). Pooling samples together appears to skew the test towards heterozygote deficiency, with a majority of samples (3 out of 5), except for m=0.1 (with 1 out of 5), presenting signals of population expansion (Supplementary Table S1).

Likewise, migration between the simulated stationary populations skewed M ratio values, which could lead to erroneous inferences of population bottlenecks (Supplementary Table S2). Even very low migration (m=0.015) leads to a decrease in M ratios of local samples, that in most cases (>82%) are smaller than expected (M_C) values. On the other hand, pooling together samples from a structured population would only lead to such an erroneous inference in the total absence of migration, since no significant departure from equilibrium M values were found for any of the other tested migration rates.

Table 3-3: Signals of bottlenecks at different sampling scales: total Iberian Peninsula and local subpopulation samples at K=4 and K=11. Estimates were performed using moment-based methods (Heterozygote excess and M ratio). Statistically significant values ($p<0.05$) are marked with *.

population	n	Wilcoxon sign ranked test p-value (heterozygote excess)	M ratio	M_c (θ=0.02, Ne=10)	M_c (θ=0.2, Ne=100)	M_c (θ=2, Ne=1000)
total Iberian Peninsula	218	0.65	1.7000	0.9073	0.8901	0.8124
4 subpopulations						
Asturias	52	0.14	1.3577	0.9019	0.8836	0.7986
Portugal	48	0.23	1.3871	0.9000	0.8829	0.7982
Castilla y León	55	0.23	1.3759	0.9007	0.8812	0.7987
Galicia	60	0.56	1.3438	0.9026	0.8857	0.8028
11 subpopulations						
Alto Minho	12	0.37	0.9473	0.9074	0.8864	0.7849
W Trás-os-Montes	11	0.13	1.1781	0.9062	0.8865	0.7812
E Asturias	18	0.10	1.0979	0.9053	0.8867	0.7908
E Trás-os-Montes	18	0.04*	1.3630	0.9064	0.8883	0.7916
SE Asturias	23	0.29	1.1903	0.9068	0.8877	0.7969
W Asturias	12	0.02*	1.0380	0.9057	0.8843	0.7823
Castilla y León	14	0.04*	1.0793	0.9052	0.8861	0.7856
S Douro	6	0.14	0.9311	0.9063	0.8846	0.7665
W Galicia	54	0.60	1.2657	0.9034	0.8858	0.8026
C Asturias	31	0.11	1.2508	0.9035	0.8854	0.7965
E Galicia	10	0.03*	1.1252	0.9059	0.8862	0.7798

Past demographic trends with *MSVAR*

The *MSVAR* method for assessing and dating population size changes identified the most recent demographic event as a severe decline, the onset dates of which depend on the model used (Supplementary Table S3). For both size change models, all 5 MCMC chains presented good mixing and converged to similar final values, as measured by the Brooks, Gelman and Rubin statistic. Density distributions of the final demographic parameter estimates were smooth and unimodal. The exponential model inferred a decline from an ancestral effective population size (*N1*) of 14 348 (4 375-44 525) to a current one (*N0*) of 55 (18-179), starting 504 (166-1 637) years ago. Modeling the size change as linear, the ancestral population is estimated to have decreased from 8 128 (2 704-26 375) to 23 (7-73) individuals, starting 3 599 (1 248-11 995) years ago. Both models therefore infer a similarly severe population size reduction (r = *N0/N1* = 0.0038 and 0.0028 for the exponential and linear models, respectively).

Parameters estimated by *MSVAR* for three (Alto Minho, S Douro and Castilla y León) of the four tested subpopulations had very large 95% HPD intervals, leading to overlapping past and current N_e values, while poor chain mixing was observed for the W Galicia population (Supplementary Table S4). Evidence for a population size change can therefore not be inferred with confidence from this data.

In the absence of population size changes in the simulated dataset, *MSVAR* infers parameter estimates with large 95% HPD intervals for all tested migration scenarios, with past and current N_e estimates partially overlapping (Supplementary Table S5). Based on mean values, the population appears to grow by a factor of 2x (r = 1.6-2.9), although past N_e values (*N1*) and their 95% HPD intervals are completely contained with the 95% HPD intervals of current N_e values (*N0*). Similarly, HPD intervals for the dates of size change are very broad (6 to >500 years), with a peak at very recent times (40-60 years ago).

Discussion

Demographic parameter inference methods, including N_e estimators and tests to detect population size changes, usually assume a simple Wright-Fisher model where the population is isolated (no migration with other populations), and that genetic variation within the population is unstructured. In terms of classical population genetics models, the existence of moderately differentiated subpopulations implies a greater resemblance to the island model of migration (Wright 1931). For such a model, the total metapopulation could only be considered approximately panmictic if migration was high.

In this study we tested the effects of population structure, gene flow and sampling scale in estimating N_e and detecting past size changes using the Iberian wolf population, in which these assumptions are mostly violated. As a whole, the Iberian wolf population meets the assumption of isolation, since it is currently isolated from the other European wolf populations, but several moderately differentiated subpopulations within Iberia have been described (Silva et al. in prep - Paper I). Silva *et al.* in prep - Paper I found pairwise F_{ST} values of ~0.13 (0.03-0.25) between the 11 subpopulations, implying less than 2 effective migrants per generation on average (*Nm*, where *N* is the effective population size and *m* the migration rate per generation). Our pooled sample of 218 individuals cannot, therefore, be considered as representing a single unstructured population. On the other hand, each wolf subpopulation within the Iberian Peninsula cannot be considered to conform to a strict WF model either, since some gene flow between subpopulations does exist.

Current effective population size of the Iberian wolf population

LD-based N_e estimators have been shown to be biased when some of the underlying model assumptions are violated (Waples and England 2011; Neel et al. 2013; Ryman et al. 2014; Gilbert and Whitlock 2015). Relevant to our study case are the interactions between the sampling scale and genetic population structure, namely when the sample is not derived from a single isolated and unstructured population. If samples from moderately differentiated populations are combined as one sample, the total N_e is expected to be underestimated because of a type of Wahlund effect where LD is generated due to the combination of offspring from genetically differentiated parents (mixture LD) (Neel et al. 2013). While this sampling effect has been demonstrated in a simulated continuously distributed population with a pattern of isolation-by-distance (Neel et al. 2013), a similar trend is observed when a structured population is modeled as connected populations in an island model (Gilbert & Whitlock, 2015; this study). Our simulations show that for migration rates of the order implied by the differentiation values between Iberian wolf subpopulations, this effect can be very severe, with the global metapopulation N_e being underestimated by up to 5-10x (figure 3-13a), and being closer to the N_e of each single subpopulation. On the other hand, when trying to estimate local N_e in a non-isolated population with unaccounted flow with other populations, values can be over-estimated if migration is sufficiently high, because of the inclusion of foreign individuals in the sample that appear as additional parents (Waples and England 2011; Neel et al. 2013). With increasing migration rates, local N_e will then instead approach the global N_e of the metapopulation, since the local sample will resemble a random sample

from a panmictic population. Our simulations suggest that the range of migration rates found in the Iberian subpopulations are not sufficiently high to produce this over-estimation bias, and the method still produces reasonably accurate values for local N_e (figure 3-13b).

Taking into account the mentioned biases and the results from our simulations, we conclude that the effective size of the Iberian wolf population derived from the total pooled sample (102.3, 95%CI: 99-105.9) is probably underestimated. Interpreting N_e as the number of effective breeders that produced the sample (Waples 2005), a tentative approximation to the the total current N_e for the Iberian wolf population could be placed around 500 individuals, assuming our estimates for the subpopulations are reasonably accurate, as seems to be supported by our simulations. Assuming a census size (N_c) of >2000 individuals (Álvares et al. 2005; Blanco and Cortés 2012), this would imply a N_e/N_c ratio of around 0.25 or less. Although there are no objective criteria to establish a 'typical' N_e/N_c ratio and considerable variability exists in these values between organisms (Palstra and Fraser 2012), our value falls within the range presented by other mammalian carnivore populations (reviewed in Frankham 1995 and Palstra and Fraser 2012), e.g.: 0.20 in Ethiopian wolves (Randall et al. 2009); 0.02-0.21 in African wild dogs (Marsden et al. 2012); 0.33-0.42 in feral domestic cats (Kaeuffer et al. 2004); 0.1-0.27 in brown bears (Miller and Waits 2003; Tallmon et al. 2004); 0.41 in tigers (Smith and McDougal 1991). Furthermore, these results would mean that ¼ of individuals effectively reproduce. Since there is usually a single breeding pair per wolf pack, this would correspond to an average pack size of 8 individuals, which is close to values estimated for the Iberian wolf population: 8.31 (Barrientos 2000) and 6-7 (Fernández-Gil et al. 2010).

Sastre et al. (2011) estimated the N_e of the entire Iberian wolf population at around 50 individuals (two estimates, 53.8 and 43.2), suggesting an overestimation of the census size or a very strong bottleneck effect. We propose instead that this estimate is severely biased downward due to a limited sampling, since only 47 samples were used, 35 of which from the single region of Castilla y León. Due to the described effect of pooling samples from a structured population, the inclusion of 12 individuals from various other regions probably had little effect in improving the estimate, and the presented value might instead be closer to the local N_e of Castilla y León. Based on 14 samples from this region we arrive at a similar estimate (N_e=42.7, 95% CI: 32.8-59.3).

Signatures of bottlenecks in the Iberian wolf population

During the 20th century, wolves were eradicated from most of the Iberian Peninsula, with populations persisting mainly in the northwestern part of the peninsula, including the Cantabrian mountain range, Galicia and Northern-Central Portugal (Valverde 1971; Petrucci-Fonseca 1990). In the 1970s, the population is estimated to have numbered ca. 500 individuals (Garzón 1979). However, it should be noted that this estimate is only an informed approximation, since no rigorous census was conducted at the time. Since the introduction of legal conservation policies in the subsequent decades, the Iberian wolf population has been expanding from this persisting area (Álvares et al. 2005; Chapron et al. 2014). If some kind of population structure already existed at the beginning of the last century, the described population size reduction probably involved the extinction of many local subpopulations, mainly in the southern regions, and size reductions of variable severity of the surviving ones. Considering the whole population, this type of population decline would deviate substantially from the size reduction of a single WF population assumed by common bottleneck detection methods.

The performance of moment-based bottleneck detection methods on a global sampling of a structured population has previously been tested by simulations in a spatially structured population with an IBD pattern (Leblois et al. 2006). The heterozygote excess method of Cornuet and Luikart (1996) was found to have similar performances in WF and IBD populations, although large-scaled samples were found to reduce the power to detect population size reductions and increase the rate of false expansion signals, suggesting that a local sampling scheme should be preferentially used. While in our simulations we did not include any population size changes, we found that in a pooled sample from differentiated populations the method indeed inferred spurious population expansions. In the presence of such a sample from a population that actually experienced a population decline, the power of the heterozygote excess method might therefore be lower. When applied to our pooled sample of 218 Iberian wolves, no significant bottleneck signal was detected. This might either be a result of the mentioned lower power of the test in a structured population, or a result of the complex recent demographic history comprising both declines and expansions, as described above. Furthermore, the test is expected to be sensitive only to recent size reductions, but the Iberian wolf population has been expanding in recent decades (>10 generations), which might make a bottleneck signal even more difficult to detect.

When applied to local samples, four subpopulations presented signals of bottlenecks with the heterozygote excess method: E Galicia, E Trás-os-Montes, W Asturias and Castilla y León. These groups are mostly located in the area where wolves

are thought to have been more abundant during the all-time population low, with the exception of Castilla Y León, whose southernmost distribution represents a recent expansion. They might therefore correspond to the surviving populations that still carry a genetic signature of a recent size reduction. These results should be interpreted with caution however, since the heterozygote excess method has previously been described to be affected by gene flow in non-isolated populations: Pope et al. (2000) found, by simulating populations under a linear stepping-stone model, that very small migration rates (m<~0.003) could produce false signals of population expansion, while higher rates resulted in the detection of population declines. The migration rates simulated in our island population model fall within this 'false population bottleneck' range, and indeed up to 20% of our simulated populations presented this bias. It is therefore possible that the bottleneck signals detected in some of these four Iberian subpopulations result from this effect.

Contrary to the heterozygote excess method, the *M* ratio method of Garza and Williamson 2001 has been found to be strongly affected by departure from the WF model assumptions both at a global and local sampling scale (Leblois et al. 2006; this study). In the case of IBD structured populations tested in the study of Leblois et al. (2006), equilibrium *M* values were found to be lower than corresponding values in WF populations, highlighting the difficulty in calculating sensible M_C values in these cases. Furthermore, *M* values after a population size reduction remained lower than equilibrium values during a short time window in structured populations and then became larger than initial equilibrium values, effectively reducing the power to detect past population declines. Ironically, the *M* ratio method is generally expected to be more informative about older size reductions than the heterozygote excess test, given that the number of alleles is expected to recover faster than the allele size range (Garza and Williamson 2001), but the effects of population structure might actually counteract this advantage. In our simulations we also found a decrease in *M* values of stationary populations due to genetic population structure, although the effects appear to be of different severity for local and pooled samples with the parameter combinations we tested. As such, *M* values of non-isolated populations connected by gene flow of the same order of magnitude as that inferred for Iberian wolf subpopulations appear to be significantly reduced. While this would mean that spurious bottlenecks would more easily be inferred, no significant reduction of *M* values was actually observed in our local samples. Since these populations have recently been expanding, the lack of size reduction signals might be a result from the post-bottleneck inflation effect described by Leblois et al. (2006).

Within the parameter range tested in our simulations, a pooled sample from a stationary structured population appears to only result in significantly reduced *M* values when differentiation is very high. It seems therefore unlikely that false bottleneck signals in the Iberian wolf population with this method could result from this effect alone. Sastre et al. (2011) has previously found significant bottleneck signals with the *M* ratio in the Iberian wolf population (but not with the heterozygote excess method), while our test was not significant, which might be due to differences in sampling and loci used.

Demographic history inference for the Iberian wolf population

The results from the demographic inference using *MSVAR* seem to suggest a very strong population decline of the Iberian wolf population starting several hundred, possibly thousands, of years ago, which cannot be explained solely by the recorded population reduction during the 20th century. Effective population sizes after the size reduction are estimated at ca. 50-70 individuals, while ancestral sizes are in the order of 10 000, implying a 200x size reduction. Even the more recent inferred starting date for the population reduction (~500 years) is incompatible with a single bottleneck in the last century. If these results are taken at face value, they suggest the influence of much older, unknown, events on the genetic diversity of the current population.

As with the other methods mentioned in this study, these results should be interpreted with caution: *MSVAR* assumes the sample is derived from an isolated, unstructured population, a condition that is unlikely to be met by our pooled sample of Iberian wolves. Recent theoretical work (Mazet et al. 2016) has shown that demographic inference methods tend to interpret population structure as population size changes, leading to erroneous demographic explanations that could also be caused by changes in the connectivity of subpopulations. These biases parallel the trends described above for bottleneck detection methods: gene flow in supposedly isolated populations leads to inferences of population declines and overestimation of effective sizes; while a sampling of several subpopulations (demes) that is assumed to come from a panmictic population can counteract this effect, it can sometimes result in false population expansions. With increasing migration values, the artificial bottlenecks appear older and ancestral effective sizes larger in local populations, while false expansions in pooled samples from structured populations appear more recent, and population sizes smaller. Our *MSVAR* analyses on simulated data correspond to a situation of pooled samples from several stationary subpopulations, and suggest that no false size changes are inferred due to population structure alone with the parameter combinations tested here. While mean parameter estimates could be interpreted as slight population expansions (mean r=2.9,

1.6 and 2.2 for *m*=0.015, 0.03 and 0.1, respectively), current N_e values (*N1*) and respective HPD intervals are almost fully contained within HPD estimates of the ancestral N_e values (*N0*), and as such no strong evidence for population size change can be inferred.

Given the mentioned limitations, it is possible that the severity and date of the population decline inferred by *MSVAR* for the Iberian wolf population has been overestimated. Theoretically, our pooled sampling scheme should minimize the effects of population structure, and our simulations support the notion that the levels of gene flow observed between wolf subpopulations should not bias towards population declines. However, in the simulation study of Chikhi *et al.*, 2010 population size reductions of 200x (log(N0/N1) = -2) were still observed in pooled samples from demes with F_{ST}=0.25. Further analyses with different sampling combinations, or methods that take population structure into account, are needed to clarify if our results correspond to an overestimation of bottleneck parameters, or to a much older population decline.

References

Allendorf FW, Luikart G. 2007. Conservation and the Genetics of Populations. Blackwell Publishing, Oxford.

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: Conference ‘‘Frontiers of Wolf Recovery: Southwestern US and the World. p. 76–77.

Balloux F. 2001. EASYPOP (version 1.7): a computer program for population genetics simulations. J. Hered. 92:301–302.

Barrientos LM. 2000. Tamaño y composición de diferentes grupos de lobos en Castilla y León. Galemys 12:249–256.

Beaumont M a. 1999. Detecting population expansion and decline using microsatellites. Genetics 153:2013–29.

Blanco JC, Cortés Y. 2012. Surveying wolves without snow: A critical review of the methods used in Spain. Hystrix 23:35–48.

Broquet T, Angelone S, Jaquiery J, Joly P, Lena JP, Lengagne T, Plenet S, Luquet E, Perrin N. 2010. Genetic bottlenecks driven by population disconnection. Conserv. Biol. 24:1596–1605.

Chapron G, Kaczensky P, Linnell JDC, von Arx M, Huber D, Andren H, Lopez-Bao J V., Adamec M, Alvares F, Anders O, et al. 2014. Recovery of large carnivores in

Europe's modern human-dominated landscapes. Science (80-.). 346:1517–1519.

Charlesworth B. 2009. Effective population size and patterns of molecular evolution and variation. Nat. Rev. Genet. 10:195–205.

Chikhi L, Sousa VC, Luisi P, Goossens B, Beaumont M a. 2010. The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. Genetics 186:983–95.

Cornuet J, Luikart G. 1996. Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. Genetics 144:2001–2014.

Do C, Waples RS, Peel D, Macbeth GM, Tillett BJ, Ovenden JR. 2014. NeEstimator v2: Re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. Mol. Ecol. Resour. 14:209–214.

Fernández-Gil A, Barrientos LM, Nuño Á. 2010. Como Estimar Estacionalmente el Tamaño Médio de Grupo en Poblaciones Ibéricas de Lobos. In: Fernández-Gil A, Álvares F, Vilà C, Ordiz A, editors. Los lobos de la Península Ibérica. Propuestas para el diagnóstico de sus poblaciones. ASCEL, Palencia. p. 208.

Frankham R. 1995. Conservation genetics. Annu. Rev. Genet. 29:305–27.

Frankham R. 2005. Genetics and extinction. Biol. Conserv. 126:131–140.

Garza JC, Williamson EG. 2001. Detection of reduction in population size using data from microsatellite loci. Mol. Ecol. 10:305–18.

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

Gilbert KJ, Whitlock MC. 2015. Evaluating methods for estimating local effective population size with and without migration. Evolution (N. Y).:n/a–n/a.

Girod C, Vitalis R, Leblois R, Fréville H. 2011. Inferring population decline and expansion from microsatellite data: a simulation-based evaluation of the Msvar method. Genetics 188:165–179.

Kaeuffer R, Pontier D, Devillard S, Perrin N, Kaeuffer R. 2004. Effective size of two feral domestic cat populations (Felis catus L.): Effect of the mating system. Mol. Ecol. 13:483–490.

Leblois R, Estoup A, Streiff R. 2006. Genetics of recent habitat contraction and reduction in population size: Does isolation by distance matter? Mol. Ecol. 15:3601–3615.

Leblois R, Pudlo P, Néron J, Bertaux F, Reddy Beeravolu C, Vitalis R, Rousset F. 2014. Maximum-Likelihood Inference of Population Size Contractions from Microsatellite Data. Mol. Biol. Evol. 31:2805–2823.

Luikart G, Cornuet J-M. 1998. Empirical Evaluation of a Test for Identifying Recently Bottlenecked Populations from Allele Frequency Data. Conserv. Biol. 12:228–237.

Luikart G, Ryman N, Tallmon D a., Schwartz MK, Allendorf FW. 2010. Estimation of census and effective population sizes: The increasing usefulness of DNA-based approaches. Conserv. Genet. 11:355–373.

Marjoram P, Tavaré S. 2006. Modern computational approaches for analysing molecular genetic variation data. Nat. Rev. Genet. 7:759–770.

Marsden CD, Woodroffe R, Mills MGL, McNutt JW, Creel S, Groom R, Emmanuel M, Cleaveland S, Kat P, Rasmussen GSA, et al. 2012. Spatial and temporal patterns of neutral and adaptive genetic variation in the endangered African wild dog (Lycaon pictus). Mol. Ecol. 21:1379–1393.

Mazet O, Rodríguez W, Grusea S, Boitard S, Chikhi L. 2016. On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference? Heredity (Edinb). 116:362–371.

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

Miller CR, Waits LP. 2003. The history of effective population size and genetic diversity in the Yellowstone grizzly (Ursus arctos): Implications for conservation. Proc. Natl. Acad. Sci. U. S. A. 100:4334–4339.

Neel MC, McKelvey K, Ryman N, Lloyd MW, Short Bull R, Allendorf FW, Schwartz MK, Waples RS. 2013. Estimation of effective population size in continuously distributed populations: there goes the neighborhood. Heredity (Edinb). 111:189–99.

Nielsen R, Beaumont MA. 2009. Statistical inferences in phylogeography. Mol. Ecol. 18:1034–1047.

Nomura T. 2008. Estimation of effective number of breeders from molecular coancestry of single cohort sample. Evol. Appl. 1:462–474.

Palstra FP, Fraser DJ. 2012. Effective/census population size ratio estimation: a compendium and appraisal. Ecol. Evol. 2:2357–65.

Paz-Vinas I, Quéméré E, Chikhi L, Loot G, Blanchet S. 2013. The demographic history of populations experiencing asymmetric gene flow: Combining simulated and empirical data. Mol. Ecol. 22:3279–3291.

Petrucci-Fonseca F. 1990. O lobo ibérico (Canis lupus signatus Cabrera, 1907) em Portugal. PhD Thesis, Faculdade de Ciencias da Universidade de Lisboa, Lisbon, Portugal.

Piry S, Luikart G, Cornuet JM. 1999. BOTTLENECK: A computer program for detecting recent reductions in the effective population size using allele frequency data. J. Hered. 90:502–503.

Pope LC, Estoup a., Moritz C. 2000. Phylogeography and population structure of an ecotonal marsupial, Bettongia tropica, determined using mtDNA and microsatellites. Mol. Ecol. 9:2041–2053.

Rambaut A, Suchard MA, Xie D, Drummond AJ. 2014. Tracer v1. 6. Comput. Progr. Doc. Distrib. by author, website http//beast.bio.ed.ac.uk/Tracer.

Randall D, Pollinger JP, Argaw K, Macdonald DW, Wayne R. 2009. Fine-scale genetic structure in Ethiopian wolves imposed by sociality, migration, and population bottlenecks. Conserv. Genet. 11:89–101.

Ryman N, Allendorf FW, Jorde PE, Laikre L, Hössjer O. 2014. Samples from subdivided populations yield biased estimates of effective size that overestimate the rate of loss of genetic variation. Mol. Ecol. Resour. 14:87–99.

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.

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

Smith BJ. 2007. boa : An R Package for MCMC Output Convergence Assessment and Posterior Inference. J. Stat. Softw. 21.

Smith JLD, McDougal C. 1991. The Contribution of Variance in Lifetime Reproduction to Effective Population Size in Tigers. Conserv. Biol. 5:484–490.

Storz JF, Beaumont M a. 2002. Testing for genetic evidence of population expansion and contraction: an empirical analysis of microsatellite DNA variation using a hierarchical Bayesian model. Evolution (N. Y). 56:154–166.

Tallmon D a, Bellemain E, Taberlet P, Swenson JE. 2004. Genetic monitoring of Scandinavian brown bear effective population size and immigration. J. Wildl. Manage. 68:960–965.

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

Wakeley J. 1999. Nonequilibrium migration in human history. Genetics 153:1863–1871.

Wang J. 2005. Estimation of effective population sizes from data on genetic markers. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 360:1395–409.

Waples RS. 2005. Genetic estimates of contemporary effective population size: to what time periods do the estimates apply? Mol. Ecol. 14:3335–52.

Waples RS, Do C. 2008. LDNE: A program for estimating effective population size from data on linkage disequilibrium. Mol. Ecol. Resour. 8:753–756.

Waples RS, Do C. 2010. Linkage disequilibrium estimates of contemporary Ne using highly variable genetic markers: A largely untapped resource for applied conservation and evolution. Evol. Appl. 3:244–262.

Waples RS, England PR. 2011. Estimating contemporary effective population size on the basis of linkage disequilibrium in the face of migration. Genetics 189:633–644.

Williamson-Natesan EG. 2005. Comparison of methods for detecting bottlenecks from microsatellite loci. Conserv. Genet. 6:551–562.

Wright S. 1931. Evolution in Mendelian populations. Genetics 16:97–159.

Zhdanova OL, Pudovkin AI. 2008. Nb_HetEx: A program to estimate the effective number of breeders. J. Hered. 99:694–695.

Chapter 4 - GENERAL DISCUSSION

4.1 - Insights into the evolutionary history of gray wolves and domestic dogs from whole genome sequence data

The use of genomic data in Papers II, III and IV allowed an unprecedented view into the ancient evolutionary history of gray wolves. These studies are representative of the increasing use of genomic resources to uncover ancient demographic histories of many species (e.g. Gronau et al. 2011; Miller et al. 2012; Zhao et al. 2013; Fan et al. 2014). In wolves, our genomic studies benefit largely from the existing resources originally developed for the domestic dog. These include a high-quality reference genome (Lindblad-Toh et al. 2005), for which a reasonably complete annotation exists, as well as commercially available SNP genotyping platforms. These resources were very important for the data quality control and the definition of neutral genomic regions appropriated for the attempted demographic inferences.

Wolves were found to have experienced a dramatic bottleneck in the last 30-50 ky, implying that current wolf populations expanded after the end of the Pleistocene and possess only a fraction of their ancestral genetic diversity. This contrasts with information gleaned from the fossil record, in which no significant reduction on the geographical distribution is apparent during the Pleistocene glacial periods (Sommer and Benecke 2005). The loss of Pleistocene wolf lineages is also supported by the comparison between modern and ancient mtDNA sequences of wolf samples (Leonard et al. 2007; Pilot et al. 2010; Thalmann et al. 2013). Many contemporary large mammalian species, including other large carnivores, have been found to present a reduced genetic diversity when compared with their Pleistocene counterparts (Hofreiter and Barnes 2010). For example, brown and polar bears, hyenas and lions show a dynamic pattern of turnover (Miller et al. 2012; Cho et al. 2013). It is therefore likely that wolves had a similar history, and several wolf-like forms existed over the last hundreds of thousands of years, possibly representing specifically-adapted ecomorphs, and which are not direct ancestors of wolf populations that exist today. This turnover might have been precipitated by changes in environmental conditions, via effects of specific prey, and by the expansion of modern humans in Eurasia (Koch and Barnosky 2006; Leonard et al. 2007; Thalmann et al. 2013).

The inferred evolutionary history supports the ancient divergence of several wolf lineages within the surviving wolf genetic diversity. Several contemporary wolf populations are currently geographically isolated due to the extinction of intermediate populations caused by human persecution and anthropogenic impacts (Kaczensky et al. 2012; Chapron et al. 2014). However, the relationships between worldwide wolf

populations inferred in Papers II and IV show that some of this geographical structure might be older, and that their distinctiveness might have been exacerbated by long-term isolation and inbreeding, as well as local adaptations. In particular, Mexican and Tibetan wolves appear as highly divergent wolf lineages in the New and Old World, respectively (Paper IV).

In Europe, Italian and Iberian wolves, which are isolated in southern peninsulas, were found to have an old divergence (2.4-7.4 kya) that precedes largely the extirpation of wolves in Central Europe during the late 19th century (Paper II). However, unlike other species whose phylogeographic patterns reflect contractions to Southern European refugia during the Pleistocene ice ages (Hewitt 2000), the ancient divergence dates of Italian and Iberian wolves suggest that relatively more recent events are to blame. The ancient isolation of Italian wolves had previously been described based on microsatellite data (Lucchini et al. 2004), and similar divergence dates (3.2-5.6 kya) had been described from SNP data (Pilot et al. 2014). Using an independent approach, our results corroborate these findings, and suggest a similar history for the Iberian wolf population. Furthermore, given the very similar separation dates and demographic trajectories, it is possible that the same events precipitated the divergence of these two populations.

The inferred demographic history of wolves has also implications in regarding the domestication of dogs. None of the sampled wolf populations in Papers III and IV are genetically closest to domestic dogs, supporting other studies based on ancient mtDNA (Thalmann et al. 2013) proposing that the ancestral wolf population from which dogs were first domesticated is probably extinct. Additionally, significant admixture between dogs and wolves was found in several wolf populations, supporting previous hypothesis of prolonged interbreeding between early forms of dogs and wild wolves (Larson et al. 2012) at least at regional levels. This means that some of the previous strategies used to infer the date and location of dog domestication, such as SNP haplotype sharing (VonHoldt et al. 2010) or genome-wide similarity (Wang et al. 2013) between dogs and wolves may not be informative regarding those questions, and might indeed be confounded by regional post-domestication admixture. The extinction of the wolf population ancestral to dogs means that studies investigating the location and date of domestication will need to increasingly rely on ancient DNA data to address these questions.

Wolf-dog divergence dates inferred from the genomic studies (Papers II-IV) are estimated between ca. 10-16 kya or ca. 25-30 kya, depending on the assumed mutation rate. The first set of dates are much more recent than estimates based on ancient DNA analysis of early dogs and wolves (ca. 27kya, Thalmann et al. 2013). These differences

might be caused by inflated mutation rates in the genomic regions used, undetected admixture with dogs or other unknown biases of the employed methods. However, the largest source of uncertainty is the mutation rate. A first step towards solving this problem was the sequencing of an ancient wolf genome (Skoglund et al. 2015). Using the mutation rate inferred from this individual leads to congruent domestication dates between the diverse studies cited. However, further verification is needed given that this estimate results from a single individual sequenced at low coverage.

The use of a relatively small number of genomes sequenced at high coverage allowed the use of methods capable of taking into account some of the greatest challenges in demographic history inference in very recently diverged populations, such as dogs and wolves, or the European wolf populations: the effects of incomplete lineage sorting (ILS) and post-divergence gene flow. ILS refers to the persistence of ancestral genetic polymorphisms during species divergence events, leading to a much older coalescence time than the divergence ('deep coalescence'). Both ILS and the admixture resulting from post-divergence gene flow can lead to significant differences in the genealogical history of distinct loci within the same genome (Cutter 2013). The sequencing of full genomes maximizes the number of loci obtained, and thus the use of methods that take these phenomena explicitly into account lead to more robust inferences (Cutter 2013).

A more complete picture of the demographic history of dogs and wolves is also useful for the study of genomic regions under selection during dog domestication. Identifying regions under selection is challenging because demographic changes can result in patterns that mimic selective sweeps (Schraiber and Akey 2015). The evolutionary history inferred in Paper III has allowed the performance of the first analysis of selection in dogs that explicitly incorporates a demographic model (Freedman et al. 2016). This study uses the demographic inferences to calibrate false discovery rates of selection detection methods, allowing for the more robust identification of targets of selection. This study identified genomic regions that likely experienced positive selection in dogs, that include loci related to behavior, neurological functions and lipid metabolism, suggesting that both behavior and dietary adaptations were important during the early stages of the domestication process.

4.2 - Genetic population structure of Iberian wolves and implications for demography

The Iberian wolf population has been subjected to centuries of persecution and human-induced environmental changes. These have led to the extirpation of wolves from most of the peninsula during the 20th century (Valverde 1971; Garzón 1979; Petrucci-Fonseca 1990), following the disappearance of wolves from Central Europe in the 19th century (Valière et al. 2003). The implementation of conservation measures in recent decades have been successful in lowering the mortality of wolves, and have allowed the population to expand (Álvares et al. 2005; Chapron et al. 2014). It is therefore timely to evaluate in the present moment what consequences this demographic history had on the genetic diversity of the population.

Knowledge about the genetic structure and patterns of gene flow of a population can reveal unknown ecological features, which are important for the effective implementation of conservation measures. Additionally, the definition of population structure serves as a foundation for further genetic and ecological studies for which the definition of population units is essential. Paper I explored the genetic structure and patterns of gene flow of the Iberian wolf population, establishing this foundation, and Paper V explores its demographic history in this context, as well as the limitations of the associated methods.

Within the relatively small area of the Iberian Peninsula unexpected levels of population structure were found (Paper I), wherein up to 11 population clusters can be discerned. These groups appear in a remarkably reticulated pattern, many of them in very small areas. They are characterized by relatively high genetic differentiation and very few dispersant individuals. The low number of dispersing individuals, low levels of admixture between groups and the overlap with individual tracking data suggests that Iberian wolves do not disperse very long distances frequently, despite the species' capability to do so. These results are consistent with other studies that show significant population structure in wolf populations (e.g. Pilot et al. 2006; Carmichael et al. 2007; Scandura et al. 2011; Jansson et al. 2012; Hindrikson et al. 2013), although the Iberian Peninsula seems to represent a remarkable case of a very fine-scale pattern in a relatively small region.

Genetic population structure can result from diverse factors, such as ecological traits, geographical obstacles to gene flow or historical events. It is possible that the observed structure of the Iberian wolf population is a consequence of the demographic decline and fragmentation that it underwent. However, the low number of dispersant

individuals found suggests that possibly other factors might have some influence as well. Clarifying what environmental determinants contribute to the population structure might benefit from future studies that integrate individual movement data and establish explicit correlations to environmental factors such as climate, habitat type, diet, human disturbance, etc. Additionally, if the identified population structure results from specific traits of Iberian wolves it should predate the anthropogenic disturbances of the 20th century, and the use of old tissue samples might be informative in this regard.

The partition of genetic diversity in the Iberian Peninsula results in a population dynamic that resembles a meta-population, however most demographic methods for inferring effective population sizes, population size changes or more complex demographic histories assume that samples originate from a simple isolated population with no genetic structure. The consequences of violating these assumptions when inferring demographic parameters for the Iberian wolf population were explored in Paper V. The effective size of the population was found to be within the expected values given the known census size estimates when population structure is taken into account, contrary to previous studies that claimed inflated census sizes or a more severe bottleneck than expected (Sastre et al. 2011). Significant signals of population reductions were found in only some of the identified subpopulations, which might be a consequence of the low power of these tests in structured populations. Furthermore, the decline of the Iberian wolf population might not fit the expected models assumed by these tests if population structure was already present before the bottleneck. While demographic inferences based on likelihood methods suggest that the decline of the Iberian wolf population might have started much earlier than assumed so far, these inferences might also be biased by departures from model assumptions. Further exploration of these questions using methods that explicitly take population structure into account are needed.

4.3 - Concluding remarks

The work that constitutes the present thesis demonstrates the utility of both traditional and emerging genetic variation markers in the inference of evolutionary history. The use of both full genome sequences and microsatellite data allowed for the exploration of the demographic history and genetic population structure of wolves at different geographical and time scales. One of the used approaches was the implementation of recently developed demographic inference methods that leverage the evolutionary information contained in full genomes to make powerful inferences about

the past, while taking into account the confounding effects of ILS and post-divergence admixture. This approach relied on a relatively low number of individual genomes, but because each genome can be viewed as mosaic of genetic fragments with different coalescent histories, they can be very informative about ancient demographic events. The obtained results contributed to increase our knowledge about the evolutionary history of gray wolves and dogs. As a complementary approach, the genotyping of microsatellite markers represented a more cost-effective sampling of genetic variation from a larger number of individuals, at the regional scale of the Iberian Peninsula. Due to their high variability, this extensive sampling can be very informative about more recent demographic events and current ecological traits. This knowledge is important both from a historical perspective and to inform effective conservation and management decisions.

4.4 - References

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: Conference ‘‘Frontiers of Wolf Recovery: Southwestern US and the World. p. 76–77.

Carmichael LE, Krizan J, Nagy J a., Fuglei E, Dumond M, Johnson D, Veitch a., Berteaux D, Strobeck C. 2007. Historical and ecological determinants of genetic structure in arctic canids. Molecular Ecology 16:3466–3483.

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

Cho YS, Hu L, Hou H, Lee H, Xu J, Kwon S, Oh S, Kim H-M, Jho S, Kim S, et al. 2013. The tiger genome and comparative analysis with lion and snow leopard genomes. Nature Communications 4.

Cutter AD. 2013. Integrating phylogenetics, phylogeography and population genetics through genomes and evolutionary theory. Molecular Phylogenetics and Evolution 69:1172–1185.

Fan Z, Zhao G, Li P, Osada N, Xing J, Yi Y, Du L, Silva P, Wang H, Sakate R, et al. 2014. Whole-Genome Sequencing of Tibetan Macaque (Macaca thibetana) Provides New Insight into the Macaque Evolutionary History. Molecular biology and evolution:1–15.

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 Genetics 12:e1005851.

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

Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A. 2011. Bayesian inference of ancient human demography from individual genome sequences. Nature Genetics 43:1031–1034.

Hewitt G. 2000. The genetic legacy of the Quaternary ice ages. Nature 405:907–13.

Hindrikson M, Remm J, Männil P, Ozolins J, Tammeleht E, Saarma U. 2013. Spatial Genetic Analyses Reveal Cryptic Population Structure and Migration Patterns in a Continuously Harvested Grey Wolf (Canis lupus) Population in North-Eastern Europe. PLoS ONE 8:e75765.

Hofreiter M, Barnes I. 2010. Diversity lost: are all Holarctic large mammal species just relict populations? BMC biology 8:46.

Jansson E, Ruokonen M, Kojola I, Aspi J. 2012. Rise and fall of a wolf population: genetic diversity and structure during recovery, rapid expansion and drastic decline. Molecular Ecology 21:5178–5193.

Kaczensky P, Chapron G, von Arx M, Huber D, Andrén H, Linnell JDC. 2012. Status, Management and Distribution of Large Canivores - Bear, Lynx, Wolf & Wolverine - in Europe.

Koch PL, Barnosky AD. 2006. Late Quaternary Extinctions: State of the Debate. Annual Review of Ecology, Evolution, and Systematics 37:215–250.

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. Proceedings of the National Academy of Sciences of the United States of America 109:8878–83.

Leonard J a., Vilà C, Fox-Dobbs K, Koch PL, Wayne RK, Van Valkenburgh B. 2007. Megafaunal extinctions and the disappearance of a specialized wolf ecomorph. Current Biology 17:1146–1150.

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

Lucchini V, Galov A, Randi E. 2004. Evidence of genetic distinction and long-term population decline in wolves (Canis lupus) in the Italian Apennines. Molecular Ecology 13:523 – 536.

Miller W, Schuster SC, Welch AJ, Ratan A, Bedoya-Reina OC, Zhao F, Kim HL, Burhans RC, Drautz DI, Wittekindt NE, et al. 2012. Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. Proceedings of the National Academy of Sciences of the United States of America 109:1–9.

Petrucci-Fonseca F. 1990. O lobo ibérico (Canis lupus signatus Cabrera, 1907) em Portugal. PhD Thesis, Faculdade de Ciencias da Universidade de Lisboa, Lisbon, Portugal.

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

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

Pilot M, Jędrzejewski W, Branicki W, Sidorovich VE, Jędrzejewska B, Stachura K, Funk SM. 2006. Ecological factors influence population genetic structure of European grey wolves. Molecular Ecology 15:4533–53.

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. Conservation Genetics 12:701–712.

Scandura M, Iacolina L, Capitani C, Gazzola A, Mattioli L, Apollonio M. 2011. Fine-scale genetic structure suggests low levels of short-range gene flow in a wolf population of the Italian Apennines. European Journal of Wildlife Research 57:949–958.

Schraiber JG, Akey JM. 2015. Methods and models for unravelling human evolutionary history. Nature Reviews Genetics 16:727–740.

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. Current Biology:1–5.

Sommer R, Benecke N. 2005. Late-Pleistocene and early Holocene history of the canid fauna of Europe (Canidae). Mammalian Biology - Zeitschrift für Säugetierkunde 70:227–241.

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 (New York, N.Y.) 342:871–4.

Valière N, Fumagalli L, Gielly L, Miquel C, Lequette B, Poulle M-L, Weber J-M, Arlettaz R, Taberlet P. 2003. Long-distance wolf recolonization of France and Switzerland inferred from non-invasive genetic sampling over a period of 10 years. Animal Conservation 6:83–92.

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

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.

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

Zhao S, Zheng P, Dong S, Zhan X, Wu Q, Guo X, Hu Y, He W, Zhang S, Fan W, et al. 2013. Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. Nature Genetics 45:67–71.

APPENDIX

Supplementary Material for Paper I - Cryptic Population Structure and Evidence of Low Dispersal in the Iberian wolf

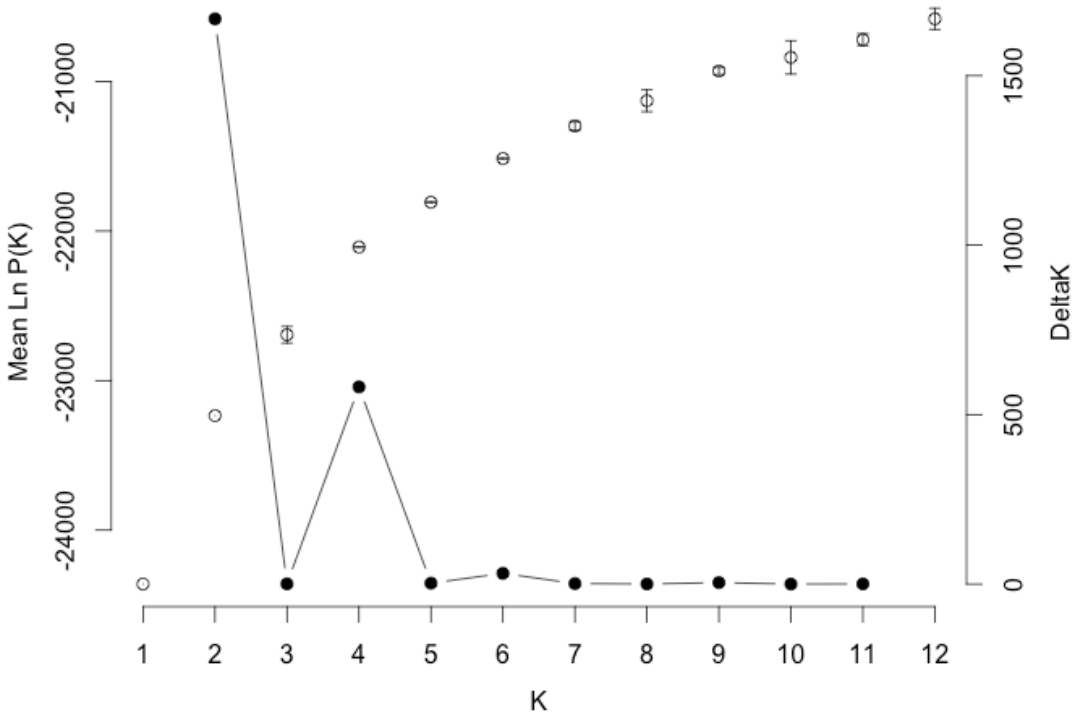


Fig. S1: Mean posterior likelihoods (open circles) and ΔK values (full circles) values of Structure runs for K=1 to K=12 across 20 independent runs for each K.

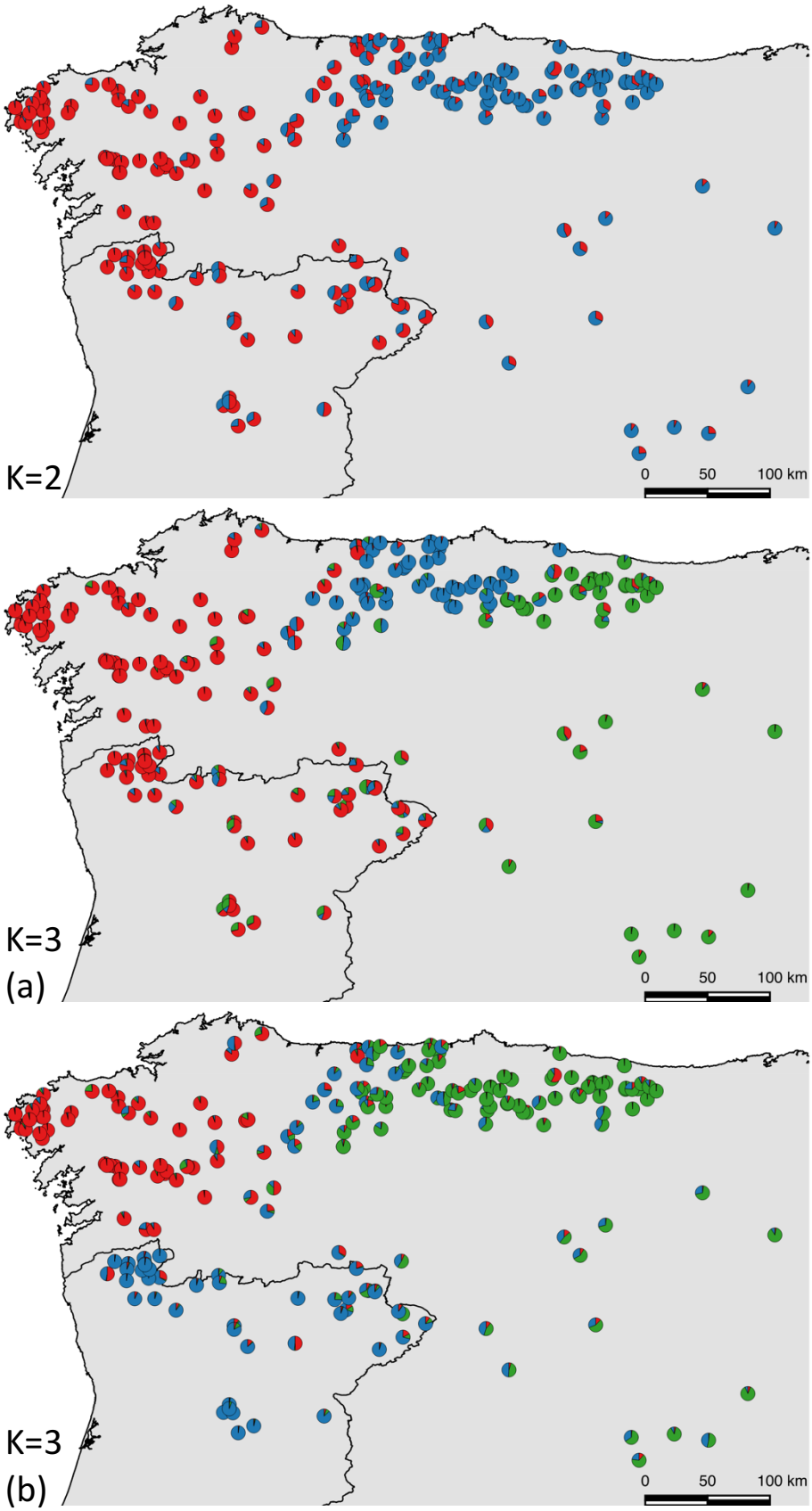


Fig. S2: Spatial projection of Structure results for K=2 to K=12. Each cluster is represented as a different color. Individuals are represented as circles with colors proportional to individual membership proportions to each cluster. Some K values presented more than one partition scheme with high posterior likelihood; in these cases, the alternatives are identified with (a), (b), etc.

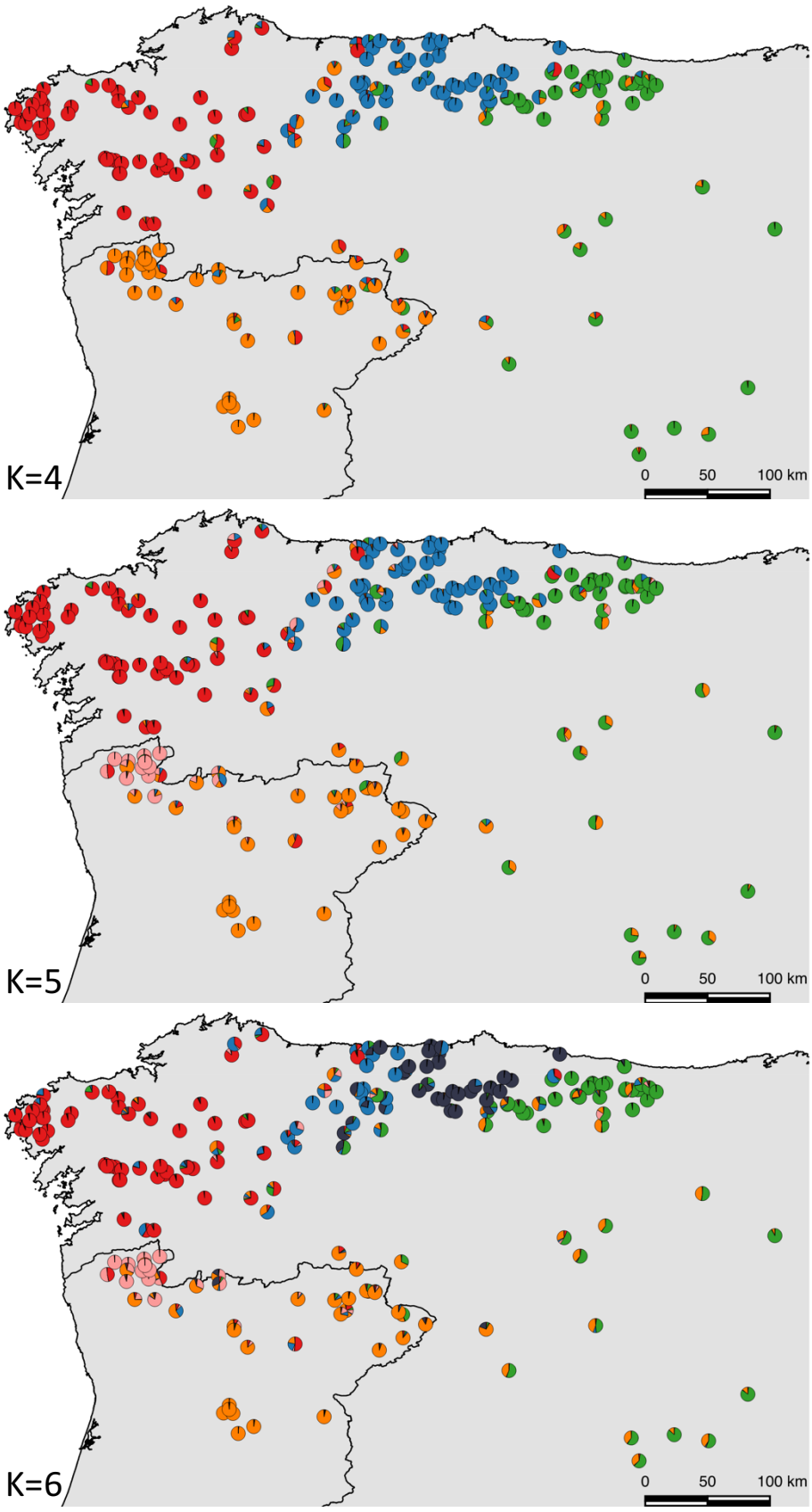


Fig. S2 (continued)

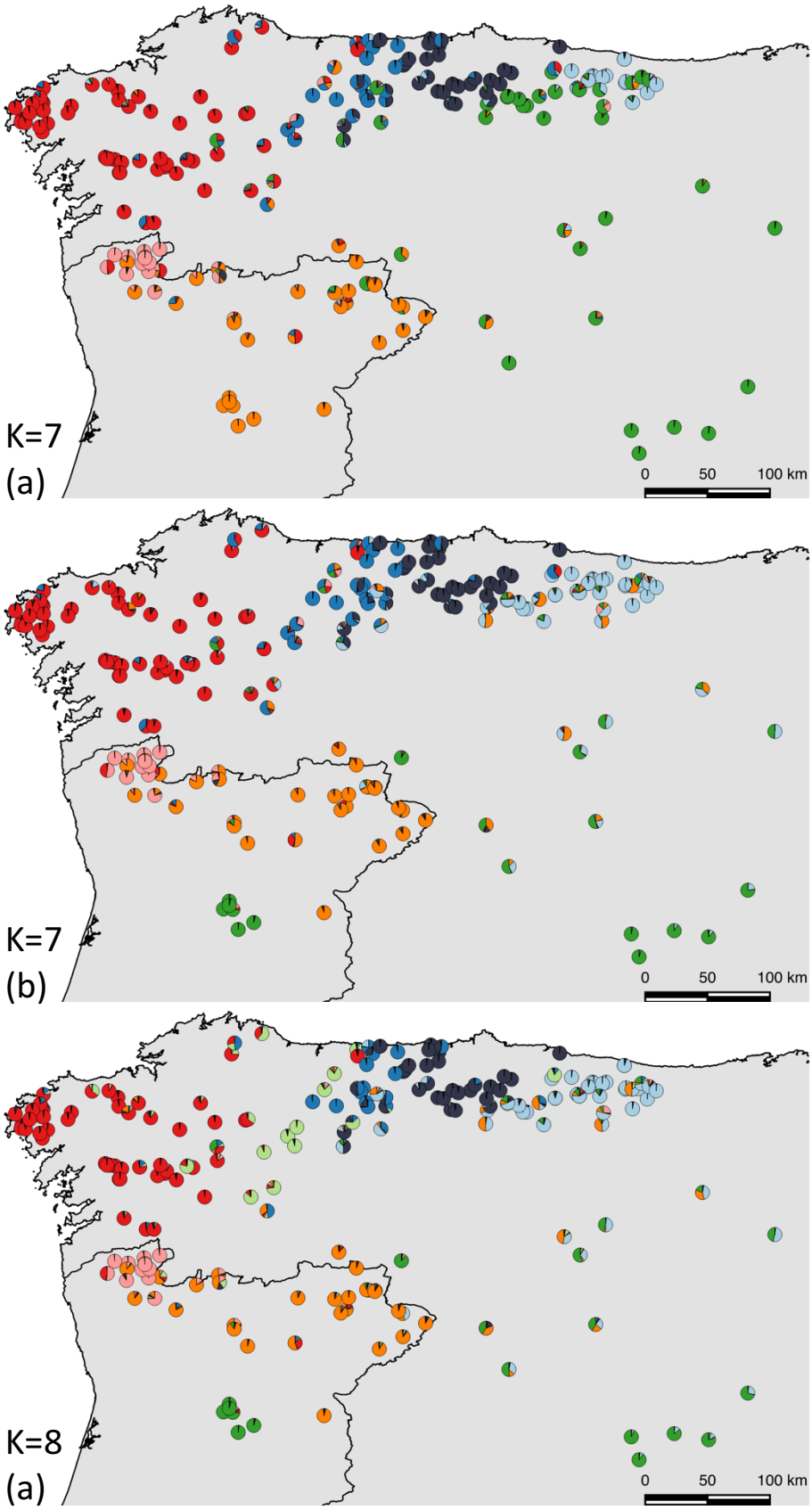


Fig. S2 (continued)

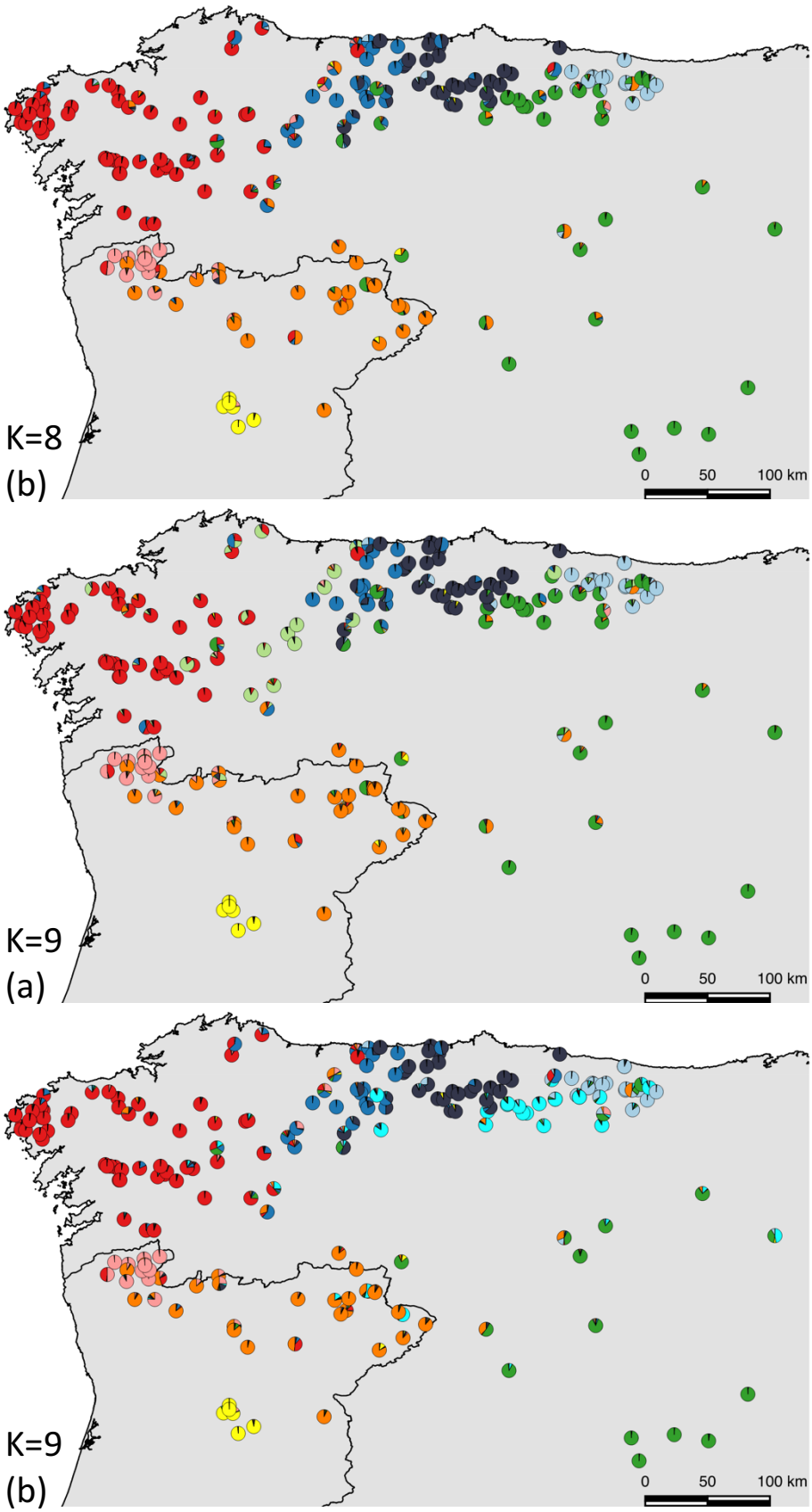


Fig. S2 (continued)

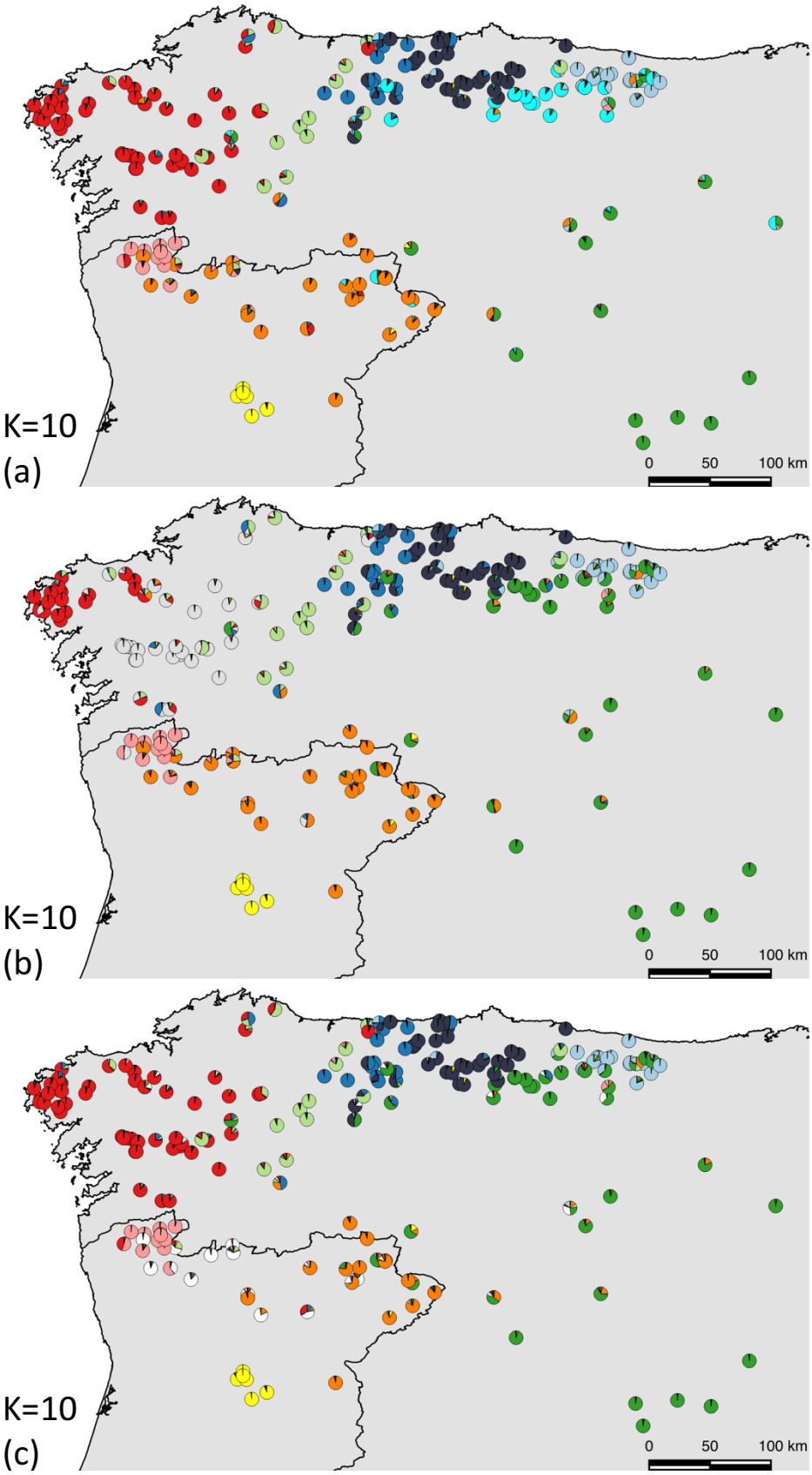


Fig. S2 (continued)

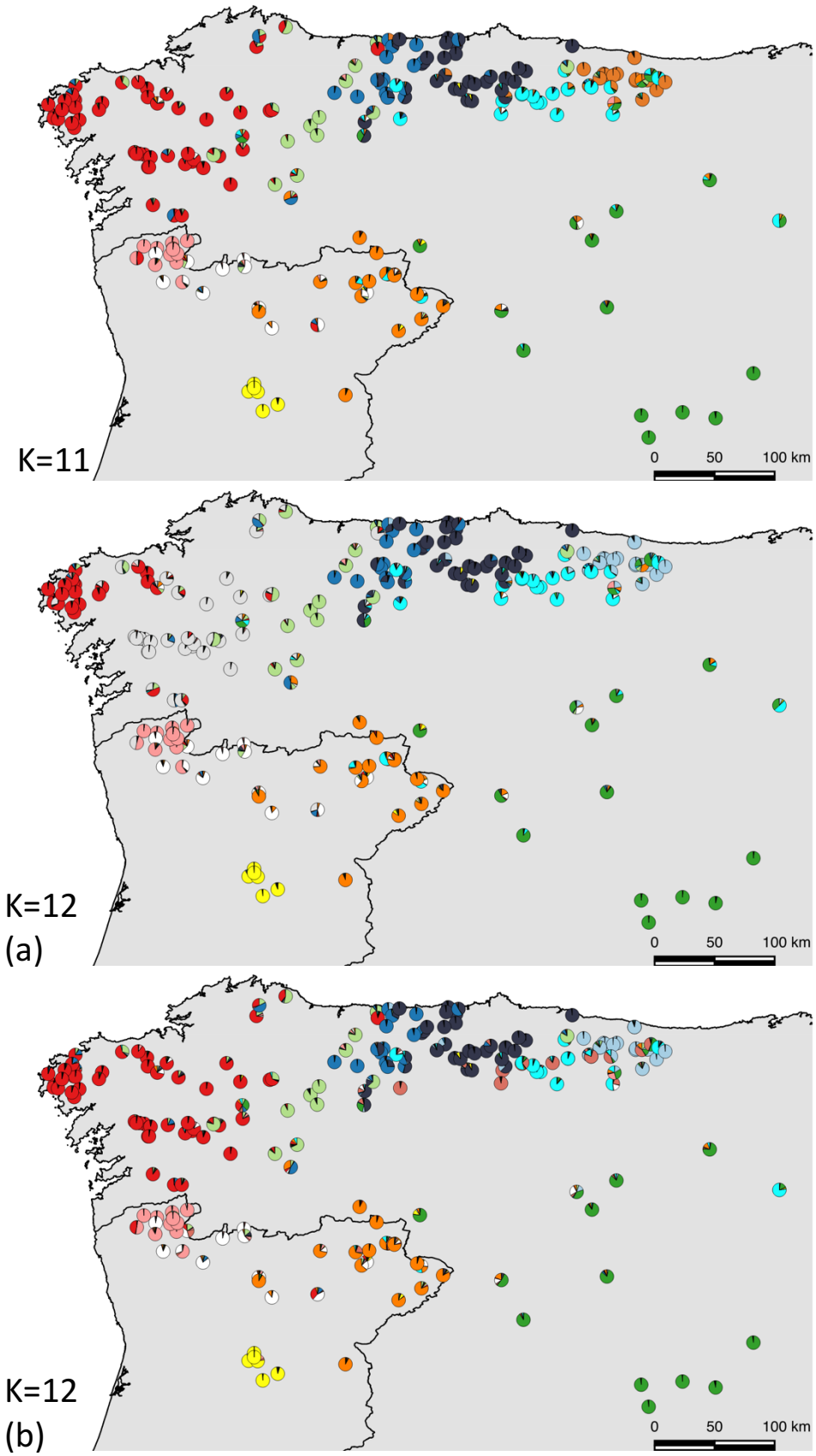


Fig. S2 (continued)

Supplementary Table 1: Membership proportion cutoff value in each genetic cluster used to classify individuals as pure or admixed.

cluster	cutoff
K=4	
Asturias	0.77
Portugal	0.82
Castilla y León	0.86
Galicia	0.87
K=11	
Alto Minho	0.88
E Trás-os-Montes	0.91
SE Asturias	0.84
W Asturias	0.95
Castilla y León	0.81
S Douro	0.9
W Galicia	0.87
C Asturias	0.87
E Galicia	0.87
W Trás-os-Montes	0.85
E Asturias	0.92

Supplementary Table 2: Microsatellite loci used in this study with corresponding repeat motif, allele size range, multiplex and reference. PCR amplification conditions are described in Godinho et al. 2011 and Godinho et al. 2015.

Locus name	Repeat type	Allele Range	Multiplex	Reference
AHT103	Di	71-89	MS2	Holmes et al. 1995
AHT111	Di	72-92	MS2	Holmes et al. 1993
AHT121	Di	74-118	Finnzymes	Holmes et al. 1995
AHT132	Di	170-182	MS1	by N. Holmes
AHT137	Di	128-160	Finnzymes	Holmes et al. 1995
AHTh171	Di	216-240	Finnzymes	Breen et al. 2001
AHTh260	Compond	229-265	Finnzymes	Breen et al. 2001
AHTk211	Di	82-98	Finnzymes	Lingaas et al. 1997
AHTk253	Di	280-300	Finnzymes	Lingaas et al. 1997
C04.140	Di	132-160	MS2	Ostrander et al. 1993
C08.410	Di	95-125	MS3	Ostrander et al. 1995
C08.618	Di	188-208	MS3	Ostrander et al. 1995
C09.173	Di	100-118	MS2	Ostrander et al. 1993
C09.474	Di	111-133	MS3	Ostrander et al. 1995
C13.758	Di	220-244	MS2	Mellersh et al. 1997
C14.866	Di	221-257	MS2	Mellersh et al. 1997
C20.253	Di	95-125	MS2	Ostrander et al. 1993
C20.446	Di	173-201	MS3	Ostrander et al. 1995
C22.279	Di	108-132	Finnzymes	Ostrander et al. 1993
C27.442	Di	158-172	MS1	Ostrander et al. 1995
CPH02	Di	87-113	MS3	Fredholm and Winterø 1995
CPH05	Di	95-131	MS3	Fredholm and Winterø 1995
CPH09	Di	133-163	MS3	Fredholm and Winterø 1995
CPH14	Di	185-205	MS2	Fredholm and Winterø 1995
CXX.459	Di	141-167	MS3	Ostrander et al. 1995
Dbar2	Di	163-169	MS4	Kerns et al. 2004
FH2001	Tetra	123-155	MS2	Francisco et al. 1996
FH2010	Tetra	216-240	MS1	Francisco et al. 1996
FH2054	Tetra	121-181	Finnzymes	Francisco et al. 1996
FH2079	Tetra	246-292	MS1	Francisco et al. 1996
FH2161	Tetra	228-260	MS3	Francisco et al. 1996
FH2848	Di	224-244	Finnzymes	Breen et al. 2001
INRA21	Di	87-103	Finnzymes	Mariat et al. 1996
INU005	Di	104-136	Finnzymes	Finnzymes, Inc
INU030	Di	136-156	Finnzymes	Finnzymes, Inc
INU055	Di	196-210	Finnzymes	Finnzymes, Inc
PEZ1	Tetra	99-131	MS1	Neff et al. 1999
PEZ3	Tri	106-147	MS1	Neff et al. 1999
PEZ5	Tetra	95-119	MS1	Neff et al. 1999
REN162C04	Di	189-215	Finnzymes	Guyon et al. 2003
REN169D01	Di	192-220	Finnzymes	Guyon et al. 2003
REN169O18	Di	145-171	Finnzymes	Guyon et al. 2003

REN247M23	Di	263-283	Finnzymes	Guyon et al. 2003
REN54P11	Di	223-245	Finnzymes	Guyon et al. 2003
REN64E19	Di	132-181	MS3	Breen et al. 2001
VWF	Hexa	138-192	MS2	Shibuya et al. 1993

References

Breen M, Jouquand S, Renier C, Mellersh CS, Hitte C, Holmes NG, Chéron A, Suter N, Vignaux F, Bristow AE, et al. 2001. Chromosome-specific single-locus FISH probes allow anchorage of an 1800-marker integrated radiation-hybrid/linkage map of the domestic dog genome to all chromosomes. *Genome Research* 11:1784–1795.

Francisco L V, Langston a a, Mellersh CS, Neal CL, Ostrander E a. 1996. A class of highly polymorphic tetranucleotide repeats for canine genetic mapping. *Mammalian genome : official journal of the International Mammalian Genome Society* 7:359–362.

Fredholm M, Winterø a K. 1995. Variation of short tandem repeats within and between species belonging to the Canidae family. *Mammalian genome : official journal of the International Mammalian Genome Society* 6:11–18.

Godinho R, Llaneza L, Blanco JC, Lopes S, Alvares F, García EJ, Palacios V, Cortés Y, Talegón J, Ferrand N, et al. 2011. Genetic evidence for multiple events of hybridization between wolves and domestic dogs in the Iberian Peninsula. *Molecular ecology* 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 non-invasive samples with ancestry informative markers. *Molecular Ecology Resources*:317–328.

Guyon R, Lorentzen TD, Hitte C, Kim L, Cadieu E, Parker HG, Quignon P, Lowe JK, Renier C, Gelfenbeyn B, et al. 2003. A 1-Mb resolution radiation hybrid map of the canine genome. *Proceedings of the National Academy of Sciences of the United States of America* 100:5296–5301.

Holmes NG, Dickens HF, Parker HL, Binns MM, Mellersh CS, Sampson J. 1995. Eighteen canine microsatellites. *Animal genetics* 26:132–133.

Holmes NG, Mellersh CS, Humphreys SJ, Binns MM, Holliman a, Curtis R, Sampson J. 1993. Isolation and characterization of microsatellites from the canine genome. *Animal genetics* 24:289–292.

Kerns JA, Newton JM, Berryere TG, Rubin EM, Cheng J-F, Schmutz SM, Barsh GS. 2004. Characterization of the dog Agouti gene and a nonagoutimutation in German Shepherd Dogs. *Mammalian Genome* 15:798–808.

Lingaas F, Sorensen A, Juneja RK, Johansson S, Fredholm M, Winterø AK, Sampson J, Mellersh C, Curzon A, Holmes NG, et al. 1997. Towards construction of a canine linkage map: establishment of 16 linkage groups. *Mammalian genome* 8:218–221.

Mariat D, Kessler JL, Vaiman D, Panthier JJ. 1996. Polymorphism characterization of five canine microsatellites. *Animal genetics* 27:434–435.

Mellersh CS, Mellersh CS, Langston a a, Langston a a, Acland GM, Acland GM, Fleming M a, Fleming M a, Ray K, Ray K, et al. 1997. A linkage map of the canine genome. *Genomics* 46:326–336.

Neff MW, Broman KW, Mellersh CS, Ray K, Acland GM, Aguirre GD, Ziegle JS, Ostrander E a, Rine J. 1999. A second-generation genetic linkage map of the domestic dog, Canis familiaris. *Genetics* 151:803–820.

Ostrander E a., Mapa F a., Yee M, Rine J. 1995. One hundred and one new simple sequence repeat-based markers for the canine genome. *Mammalian Genome* 6:192–195.

Ostrander EA, Sprague Jr. GF, Rine J. 1993. Identification and characterization of dinucleotide repeat CAn markers for genetic mapping in dog. *Genomics* 16:207–213.

Shibuya H, Collins B, Huang TH-M, Johnson G. 1993. A polymorphic AGGAAT, tandem repeat in an intron of the canine von Willebrand factor gene. *Animal genetics* 25:122.

Supplementary Table 3: Summary of the spatial behavior information for 85 wolves collared in the Iberian Peninsula from 1982 to 2015. "-" = Information not available.

ID	Full monitoring MCP area (km^2)	Period	Sampling period (days)	Number of locations recorded	Sex	Age (yrs)	Collar type	Reference
1	67	1982	228	-	M	>2	VHF	Pereira et al. 1985
2	35	1982-1983	276	-	M	<2	VHF	Pereira et al. 1985
3	65	1991	44	-	F	>2	VHF	Moreira 1992
4	91	1991-1992	165	-	F	<2	VHF	Moreira 1992
5	309	1996-1997	534	-	M	>2	VHF	Pimenta 1998
6	288	1996-1997	534	-	F	>2	VHF	Pimenta 1998
7	123	1997	169	-	M	>2	VHF	Pimenta 1998
8	228	1997	159	-	M	<2	VHF	Pimenta 1998
9	155	1997-1998	135	37	F	>2	VHF	
10	950	1997-2002	2121	242	M	<2	VHF	
11	530	1997-2002	1734	282	F	>2	VHF	
12	880	1997-2003	2101	247	F	<2	VHF	
13	2810	1997-2003	2129	417	F	<2	VHF	
14	504	1997-1999	487	-	M	<2	VHF	Grilo et al. 2002
15	1040	1998	229	47	M	>2	VHF	
16	132	1998-1999	119	-	M	<2	VHF	Grilo et al. 2002
17	169	1998-1999	150	-	F	>2	VHF	Grilo et al. 2002
18	695	1998-1999	563	327	M	>2	VHF	
19	1230	1998-2001	1029	96	F	>2	VHF	
20	2030	1998-2002	1645	188	M	>2	VHF	
21	670	1998-2003	2047	183	M	>2	VHF	
22	56	1999	102	89	F	<2	VHF	
23	23	1999	58	23	M	<2	VHF	
24	225	1999	142	55	M	<2	VHF	
25	270	1999-2000	293	56	F	>2	VHF	
26	398	1999-2002	784	313	M	<2	VHF	
27	890	2000-2003	820	135	M	>2	VHF	

28	640	2002-2004	638	312	F	<2	VHF	
29	233	2004	89	153	M	>2	GPS	
30	350	2004-2005	388	42	M	<2	VHF	
31	80	2005-2006	536	15	F	<2	VHF	
32	648	2005-2007	632	1853	F	>2	GPS	
33	238	2006	154	3484	M	>2	GPS	
34	168	2006	52	1133	M	<2	GPS	
35	35	2006	23	247	M	<2	GPS	
36	570	2006	95	971	M	>2	GPS	
37	80	2007	251	2684	F	<2	GPS	
38	80	2007	95	1091	M	<2	GPS	
39	189	2007	69	755	M	>2	GPS	
40	51	2007-2008	75	861	F	>2	GPS	
41	61	2008	103	1011	M	<2	GPS	
42	151	2008-2010	526	5836	F	>2	GPS	
43	391	2009	111	3377	F	>2	GPS	
44	281	2009	344	7146	M	>2	GPS	
45	429	2009-2010	315	10181	F	>2	GPS	
46	831	2009-2010	241	7252	F	>2	GPS	
47	51	2009-2010	112	2595	M	<2	GPS	
48	191	2010-2011	129	4357	M	<2	GPS	
49	205	2010-2011	206	5069	M	>2	GPS	
50	211	2010-2011	128	2498	F	<2	GPS	
51	1169	2010-2011	232	1948	F	>2	GPS	
52	850	2010-2011	399	-	M	<2	GPS	Roque et al. 2011
53	98	2010-2011	368	3005	F	>2	GPS	
54	210	2010-2011	256	2083	M	>2	GPS	
55	290	2011	143	3187	F	<2	GPS	
56	235	2011	138	2989	F	<2	GPS	
57	274	2011	139	2889	F	<2	GPS	
58	70	2011	91	2094	F	>2	GPS	
59	165	2011	121	2621	F	>2	GPS	
60	1031	2011-2012	269	6433	M	>2	GPS	
61	115	2011-2012	488	9499	F	<2	GPS	
62	89	2011-2012	146	3486	F	>2	GPS	
63	30	2011-2012	91	2120	M	<2	GPS	
64	530	2011-2012	163	3920	M	>2	GPS	
65	237	2011-2012	284	4866	M	<2	GPS	

66	911	2012	138	1214	M	<2	GPS
67	291	2012	58	600	F	>2	GPS
68	89	2012-2013	259	2855	M	<2	GPS
69	32	2012-2013	71	254	M	<2	GPS
70	800	2012-2013	342	2977	F	>2	GPS
71	113	2012-2013	386	4116	F	>2	GPS
72	1427	2012-2014	606	6394	M	<2	GPS
73	55	2013	115	682	F	<2	GPS
74	96	2013	303	1993	F	>2	GPS
75	1017	2013-2014	466	4051	F	>2	GPS
76	225	2013-2014	401	4457	F	>2	GPS
77	312	2013-2014	162	7306	M	<2	GPS
78	410	2013-2014	307	13493	M	<2	GPS
79	197	2013-2014	327	13709	M	<2	GPS
80	314	2014	285	4792	F	>2	GPS
81	1725	2014	286	7605	M	<2	GPS
82	83	2012	75	962	M	<2	GPS
83	14	2012	59	948	F	<2	GPS
84	48	2014	16	315	M	>2	GPS
85	124	2015	43	542	M	>2	GPS

References

Grilo C, Moço G, Cândido A, Alexandre S, Petrucci-Fonseca F. 2002. Bases para a definição de corredores ecológicos na conservação de uma população marginal e fragmentada: o caso da população lupina a sul do rio Douro – 1ª Fase. Relatório Técnico PRAXIS XXI.

Moreira L. 1992. Contribuição para o estudo da ecologia do lobo (Canis lupus signatus Cabrera 1907) no Parque Natural de Montesinho. Relatório de estágio.

Pereira M, Fonseca F, Magalhães C. 1985. Wolf ecology in Portugal. In: Symposium Predateurs, Lisbonne 29/31.3.1985. Lisboa. p. 122–167.

Pimenta V. 1998. Estudo comparativo de duas alcateias no nordeste do distrito de Bragança: Utilização do espaço e do tempo e hábitos alimentares. Relatório de estágio.

Roque S, Godinho R, Cadete D, Pinto S, Pedro A, Bernardo J, Petrucci-Fonseca F, Álvares F. 2011. Plano de Monitorização do Lobo Ibérico nas áreas dos Projetos Eólicos das Serras de Montemuro, Freita, Arada e Leomil – Ano IV e Análise Integrativa dos Resultados (2006–2011). Relatório Final.

Supplementary Table 4: Pairwise Jost's D distance matrix between geographical populations (below diagonal) and genetic clusters (above diagonal) at K=4

	Asturias	**Portugal**	**Castilla y León**	**Galicia**
Asturias	--	0.14	0.14	0.16
Portugal	0.11	--	0.19	0.13
Castilla y León	0.11	0.14	--	0.20
Galicia	0.14	0.10	0.16	--

Supplementary Table 5: Pairwise Jost's D distance matrix between geographical populations (below diagonal) and genetic clusters (above diagonal) at K=11

	Alto Minho	**E Trás-os-Montes**	**SE Asturias**	**W Asturias**	**Castilla Y León**	**S Douro**	**W Galicia**	**C Asturias**	**E Galicia**	**W Trás-os-Montes**	**E Asturias**
Alto Minho	--	0.30	0.29	0.30	0.38	0.37	0.25	0.32	0.24	0.19	0.40
E Trás-os-Montes	0.19	--	0.19	0.23	0.23	0.29	0.18	0.24	0.15	0.18	0.33
SE Asturias	0.26	0.12	--	0.23	0.16	0.31	0.18	0.16	0.22	0.20	0.17
W Asturias	0.24	0.12	0.15	--	0.31	0.40	0.24	0.24	0.16	0.27	0.33
Castilla Y León	0.32	0.16	0.11	0.23	--	0.33	0.24	0.27	0.26	0.29	0.24
S Douro	0.32	0.21	0.28	0.29	0.29	--	0.33	0.31	0.30	0.32	0.40
W Galicia	0.20	0.12	0.16	0.18	0.20	0.29	--	0.21	0.13	0.17	0.30
C Asturias	0.26	0.16	0.13	0.12	0.21	0.28	0.18	--	0.19	0.20	0.24
E Galicia	0.17	0.11	0.18	0.10	0.22	0.26	0.10	0.17	--	0.21	0.32
W Trás-os-Montes	0.10	0.05	0.14	0.13	0.19	0.25	0.10	0.14	0.10	--	0.34
E Asturias	0.31	0.20	0.08	0.20	0.16	0.34	0.23	0.17	0.22	0.22	--

Supplementary Material for Paper II - Historic Demography and Divergence of European Wolf Populations

Command lines used for simulated data in the G-PhoCS analyses

Command Line 1: G-PhoCS model using the inferred demographic parameters

```
./ms 17 15000 -r 920 1000 -I 9 1 2 2 2 2 2 2 2 -n 1 0.0002920 -
n 2 0.0000770 -n 3 0.0000350 -n 4 0.0008340 -n 5 0.0002000 -n 6
0.0000720 -n 7 0.0002450 -n 8 0.0000990 -n 9 0.0006970 -m 4 2 0.0
-m 4 1 0.0 -m 7 1 6060.61 -m 7 2 606.06 -m 7 3 303.03 -m 3 4 303.03
-m 6 8 0.0 -m 5 9 0.0 -ej 0.000008 7 8 -em 0.000008 7 1 0.0 -em
0.000008 7 2 0.0 -em 0.000008 7 3 0.0 -em 0.000008 6 8 0.0 -en
0.000008 8 0.0000160 -ej 0.000009 6 8 -en 0.000009 8 0.0001070 -
ej 0.00001 5 8 -em 0.00001 5 9 0.0 -en 0.00001 8 0.0006530 -ej
0.000032 4 8 -en 0.000032 8 0.0000630 -ej 0.000019 2 1 -em 0.000019
4 2 0.0 -em 0.000019 4 1 0.0 -en 0.000019 1 0.0000020 -ej 0.000019
3 1 -em 0.000019 3 4 0.0 -en 0.000019 1 0.0001790 -ej 0.000033 8
1 -em 0.000033 9 1 1132.78 -en 0.000033 1 0.0017500 -ej 0.001366
9 1 -em 0.001366 9 1 0.0 -en 0.001366 1 0.0005580 -T -seeds 1 2 3
-p 10 | tail +4 | grep -v // >treefile
```

Command Line 2: Example of how we create simulated data from a set of trees inside each loci using seq-gen

```
./seq-gen -mHKY -l 1000 -s 1 -fe -t 0.5 -p 1000 < TemporalTrees.txt
> LociData.txt
```

Where TemporalTrees.txt has a collection of trees inside each of the loci, an example of the data for TemporalTrees.txt is shown below:

```
[414]((((1:0.000022,(4:0.000006,5:0.000006):0.000016):0.000035,(
(2:0.000002,3:0.000002):0.000047,(9:0.000027,((10:0.000000,11:0.
000000):0.000015,(15:0.000003,(12:0.000003,13:0.000003):0.000000
):0.000012):0.000011):0.000022):0.000009):0.000164,(8:0.000087,1
4:0.000087):0.000134):0.001056,((16:0.000015,17:0.000015):0.0003
75,(6:0.000307,7:0.000307):0.000082):0.000887);
[14]((((1:0.000022,(4:0.000006,5:0.000006):0.000016):0.000035,((
2:0.000002,3:0.000002):0.000047,(9:0.000027,((10:0.000000,11:0.0
00000):0.000015,(15:0.000003,(12:0.000003,13:0.000003):0.000000)
:0.000012):0.000011):0.000022):0.000009):0.000164,(8:0.000087,14
:0.000087):0.000134):0.001056,((16:0.000015,17:0.000015):0.00037
5,(6:0.000307,7:0.000307):0.000082):0.000887);
[63]((16:0.000015,17:0.000015):0.001756,((((1:0.000022,(4:0.0000
06,5:0.000006):0.000016):0.000035,((2:0.000002,3:0.000002):0.000
047,(9:0.000027,((10:0.000000,11:0.000000):0.000015,(15:0.000003
```

```
,(12:0.000003,13:0.000003):0.000000):0.000012):0.000011):0.00002
2):0.000009):0.000164,(8:0.000087,14:0.000087):0.000134):0.00105
6,(6:0.000307,7:0.000307):0.000969):0.000494);
[174]((16:0.000015,17:0.000015):0.001653,((((1:0.000022,(4:0.000
006,5:0.000006):0.000016):0.000035,((2:0.000002,3:0.000002):0.00
0047,(9:0.000027,((10:0.000000,11:0.000000):0.000015,(15:0.00000
3,(12:0.000003,13:0.000003):0.000000):0.000012):0.000011):0.0000
22):0.000009):0.000164,(8:0.000087,14:0.000087):0.000134):0.0010
56,(6:0.000307,7:0.000307):0.000969):0.000391);
[59]((16:0.000015,17:0.000015):0.001653,((((1:0.000022,(4:0.0000
06,5:0.000006):0.000016):0.000035,((2:0.000002,3:0.000002):0.000
047,(9:0.000027,((10:0.000000,11:0.000000):0.000015,(15:0.000003
,(12:0.000003,13:0.000003):0.000000):0.000012):0.000011):0.00002
2):0.000009):0.000164,(8:0.000087,14:0.000087):0.000134):0.00126
1,(6:0.000307,7:0.000307):0.001175):0.000186);
[176]((16:0.000015,17:0.000015):0.001653,(7:0.001482,(6:0.000643
,(((1:0.000022,(4:0.000006,5:0.000006):0.000016):0.000035,((2:0.
000002,3:0.000002):0.000047,(9:0.000027,((10:0.000000,11:0.00000
0):0.000015,(15:0.000003,(12:0.000003,13:0.000003):0.000000):0.0
00012):0.000011):0.000022):0.000009):0.000164,(8:0.000087,14:0.0
00087):0.000134):0.000422):0.000838):0.000186);
[100]((16:0.000015,17:0.000015):0.001653,(7:0.001482,(6:0.000643
,(((1:0.000022,(4:0.000006,5:0.000006):0.000016):0.000035,((2:0.
000002,3:0.000002):0.000047,(9:0.000027,((10:0.000000,11:0.00000
0):0.000015,(15:0.000003,(12:0.000003,13:0.000003):0.000000):0.0
00012):0.000011):0.000022):0.000009):0.000164,(8:0.000087,14:0.0
00087):0.000134):0.000422):0.000838):0.000186);
```

Command Line 3: G-PhoCS model using the inferred demographic parameters with a divergence time equal to 240 years (80 generations) between the Portuguese Wolf and the Spanish Wolf

```
./ms 17 15000 -r 920 1000 -I 9 1 2 2 2 2 2 2 2 -n 1 0.0002920 -
n 2 0.0000770 -n 3 0.0000350 -n 4 0.0008340 -n 5 0.0002000 -n 6
0.0000720 -n 7 0.0002450 -n 8 0.0000990 -n 9 0.0006970 -m 4 2 0.0
-m 4 1 0.0 -m 7 1 6060.61 -m 7 2 606.06 -m 7 3 303.03 -m 3 4 303.03
-m 6 8 0.0 -m 5 9 0.0 -ej 0.0000008 7 8 -em 0.0000008 7 1 0.0 -em
0.0000008 7 2 0.0 -em 0.0000008 7 3 0.0 -em 0.0000008 6 8 0.0 -en
0.0000008 8 0.0000160 -ej 0.000009 6 8 -en 0.000009 8 0.0001070 -
ej 0.00001 5 8 -em 0.00001 5 9 0.0 -en 0.00001 8 0.0006530 -ej
0.000032 4 8 -en 0.000032 8 0.0000630 -ej 0.000019 2 1 -em 0.000019
4 2 0.0 -em 0.000019 4 1 0.0 -en 0.000019 1 0.0000020 -ej 0.000019
3 1 -em 0.000019 3 4 0.0 -en 0.000019 1 0.0001790 -ej 0.000033 8
1 -em 0.000033 9 1 1132.78 -en 0.000033 1 0.0017500 -ej 0.001366
9 1 -em 0.001366 9 1 0.0 -en 0.001366 1 0.0005580 -T -seeds 1 2 3
-p 10 | tail +4 | grep -v // >treefile2
```

Command Line 4: G-PhoCS model using the inferred demographic parameters assuming that the Portuguese Wolf and the Spanish Wolf are part of a panmictic population

```
./ms 17 15000 -r 920 1000 -I 9 1 2 2 2 2 2 4 0 2 -n 1 0.0002920 -
n 2 0.0000770 -n 3 0.0000350 -n 4 0.0008340 -n 5 0.0002000 -n 6
0.0000720 -n 7 0.0000160 -n 8 0.0000990 -n 9 0.0006970 -m 4 2 0.0
-m 4 1 0.0 -m 7 1 6060.61 -m 7 2 606.06 -m 7 3 303.03 -m 3 4 303.03
-m 6 8 0.0 -m 5 9 0.0 -ej 0.0 7 8 -em 0.0 7 1 0.0 -em 0.0 7 2 0.0
-em 0.0 7 3 0.0 -em 0.0 6 8 0.0 -en 0.0 8 0.0000160 -ej 0.000009
6 8 -en 0.000009 8 0.0001070 -ej 0.00001 5 8 -em 0.00001 5 9 0.0
-en 0.00001 8 0.0006530 -ej 0.000032 4 8 -en 0.000032 8 0.0000630
-ej 0.000019 2 1 -em 0.000019 4 2 0.0 -em 0.000019 4 1 0.0 -en
0.000019 1 0.0000020 -ej 0.000019 3 1 -em 0.000019 3 4 0.0 -en
0.000019 1 0.0001790 -ej 0.000033 8 1 -em 0.000033 9 1 1132.78 -
en 0.000033 1 0.0017500 -ej 0.001366 9 1 -em 0.001366 9 1 0.0 -en
0.001366 1 0.0005580 -T -seeds 1 2 7 -p 10 | tail +4 | grep -v //
>treefile3
```

G-PhoCS parameter estimates

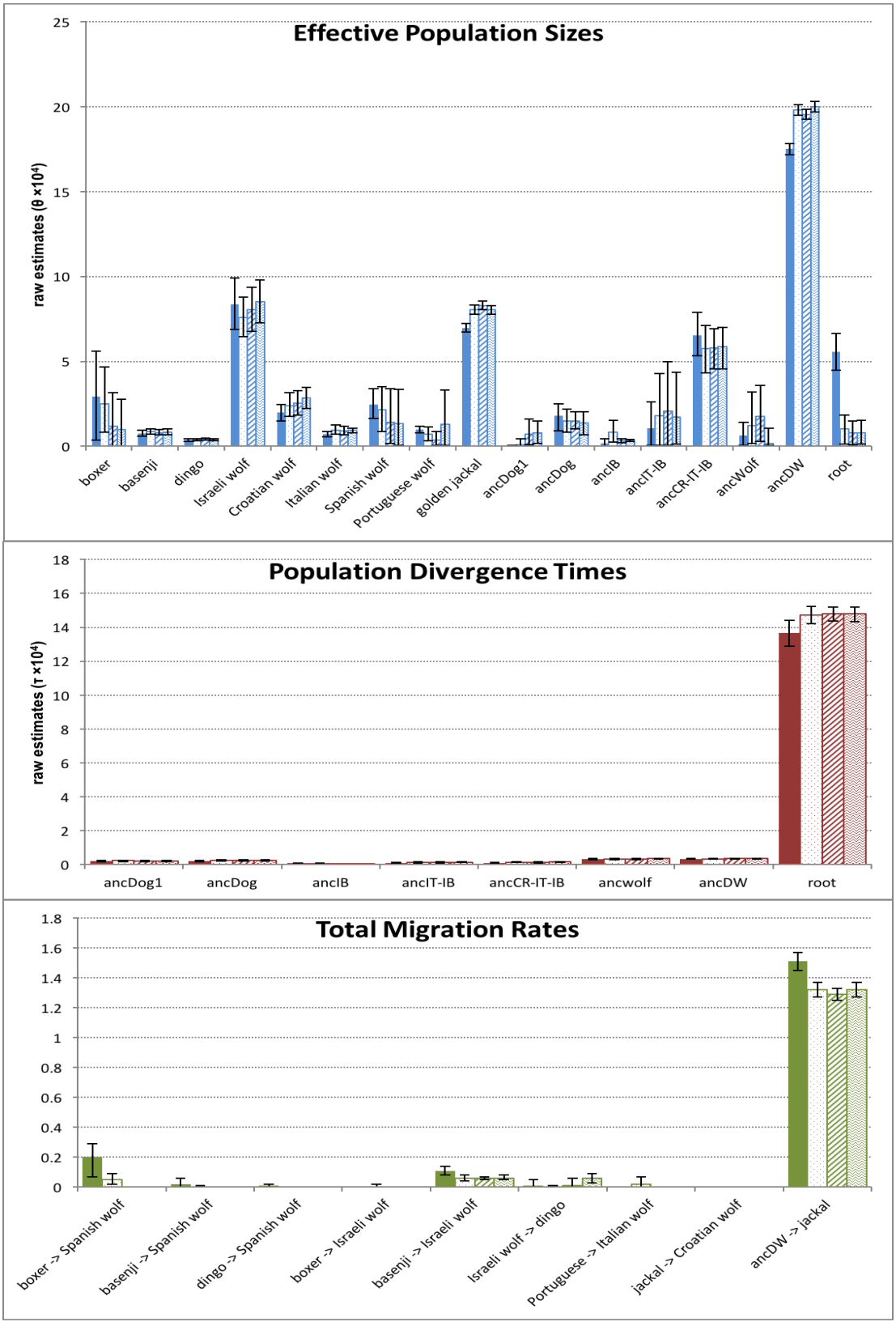
Table S1: Demographic parameter estimates by G-PhoCS. Calibrated estimates are provided assuming a mutation rate of $0.4x10^{-4}$ /bp/gen or $1x10^{-4}$ /bp/gen, and an average generation time of 3 years.

parameter	raw estimate $x10^4$	calibrated estimate ($0.4x10^{-4}$)	calibrated estimate ($1x10^{-4}$)
N_{BOX}	2.92 (0.36-5.59)	18279 (2,263-34,938)	7,312 (905-13,975)
N_{BAS}	0.77 (0.61-0.94)	4,791 (3,814-5,904)	1,916 (1,526-2,362)
N_{DIN}	0.35 (0.29-0.44)	2,183 (1,785-2,724)	873 (714-1,090)
N_{ISW}	8.34 (6.88-9.91)	52,121 (42,981-61,949)	20,849 (17,192-24,780)
N_{CRW}	2 (1.51-2.47)	12,474 (9,439-15,443)	4,990 (3,776-6,177)
N_{ITW}	0.72 (0.57-0.87)	4,514 (3,583-5,449)	1,806 (1,433-2,180)
N_{SPW}	2.45 (1.64-3.4)	15,329 (10,250-21,270)	6,132 (4,100-8,508)
N_{PTW}	0.99 (0.81-1.19)	6,191 (5,054-7,422)	2,477 (2,022-2,969)
N_{JAC}	6.97 (6.71-7.23)	43,574 (41,963-45,177)	17,430 (16,785-18,071)
$N_{ancDOG1}$	0.02 (0-0.04)	95 (27-233)	38 (11-93)
N_{ancDOG}	1.79 (0.92-2.49)	11,214 (5,769-15,539)	4,486 (2,308-6,216)
N_{ancIB}	0.16 (0.01-0.44)	998 (79-2,726)	399 (32-1,090)
$N_{ancIT-IB}$	1.07 (0.04-2.61)	6,704 (281-16,332)	2,682 (112-6,533)
$N_{ancCR-IT-IB}$	6.53 (5.33-7.89)	40,841 (33,299-49,304)	16,336 (13,320-19,722)
$N_{ancWOLF}$	0.63 (0.05-1.41)	3,940 (281-8,833)	1,576 (113-3,533)
N_{ancDW}	17.5 (17.16-17.84)	109,365 (107,273-111,483)	43,746 (42,909-44,593)
N_{root}	5.58 (4.49-6.65)	34,888 (28,070-41,576)	13,955 (11,228-16,631)
$T_{ancDOG1}$	0.19 (0.16-0.24)	14,468 (11,685-17,843)	5,787 (4,674-7,137)
T_{ancDOG}	0.19 (0.16-0.24)	14,520 (11,708-17,955)	5,808 (4,683-7,182)
T_{ancIB}	0.08 (0.07-0.09)	5,978 (5,198-7,110)	2,391 (2,079-2,844)
$T_{ancIT-IB}$	0.09 (0.08-0.11)	6,945 (5,625-8,288)	2,778 (2,250-3,315)
$T_{ancCR-IT-IB}$	0.1 (0.08-0.12)	7,418 (5,873-8,685)	2,967 (2,349-3,474)
$T_{ancWOLF}$	0.32 (0.29-0.35)	24,180 (22,065-26,438)	9,672 (8,826-10,575)
T_{ancDW}	0.33 (0.3-0.36)	24,923 (22,680-27,128)	9,969 (9,072-10,851)
T_{root}	13.66 (12.87-14.42)	1,024,508 (965,220-1,081,530)	409,803 (386,088-432,612)
$m_{BOX->SPW}$	0.2 (0.07-0.29)		
$m_{BAS->SPW}$	0.02 (0-0.06)		
$m_{DIN->SPW}$	0.01 (0-0.02)		
$m_{BOX->ISW}$	0 (0-0)		
$m_{BAS->Isw}$	0.11 (0.08-0.14)		
$m_{ISW->DIN}$	0.01 (0-0.05)		
$m_{PTW->ITW}$	0 (0-0)		
$m_{JAC->CRW}$	0 (0-0)		
$m_{ancDW->JAC}$	1.51 (1.45-1.57)		

Table S2: Comparison of demographic parameter estimates by G-PhoCS from simulated data.

parameter	control simulation	smaller divergence time PTW-SPW	PTW-SPW panmixia
N_{BOX}	2.54 (0.87-4.58)	1.15 (0-3.31)	1.08 (0-3.05)
N_{BAS}	1.08 (0.98-1.18)	0.81 (0.66-0.96)	0.68 (0.53-0.82)
N_{DIN}	0.51 (0.47-0.55)	0.37 (0.31-0.44)	0.33 (0.27-0.4)
N_{ISW}	6.39 (5.16-7.62)	6.41 (4.33-8.23)	7.15 (5.94-8.33)
N_{CRW}	2.77 (2.16-3.4)	2.02 (1.47-2.48)	1.76 (1.17-2.35)
N_{ITW}	0.89 (0.72-1.04)	0.76 (0.58-0.91)	0.67 (0.42-0.89)
N_{SPW}	1.69 (0.96-2.48)	1.79 (0.05-4.06)	1.5 (0.02-3.47)
N_{PTW}	0.7 (0.39-1.03)	1.81 (0.08-4.06)	1.38 (0.03-3.24)
N_{JAC}	7.91 (7.65-8.17)	8.2 (7.93-8.47)	7.96 (7.71-8.24)
$N_{ancDOG1}$	1.13 (0.01-3.11)	1.99 (0.1-4.6)	1.91 (0.03-4.49)
N_{ancDOG}	0.01 (0-0.03)	1.79 (1.13-2.4)	1.89 (1.4-2.43)
N_{ancIB}	1.04 (0.45-1.62)	0.18 (0.12-0.23)	0.14 (0.06-0.19)
$N_{ancIT-IB}$	1.6 (0.24-3.92)	1.54 (0.03-3.76)	1.89 (0.17-5.03)
$N_{ancCR-IT-IB}$	3.28 (1.83-4.82)	4.16 (1.46-5.97)	5.92 (4.2-7.43)
$N_{ancWOLF}$	2.52 (0.29-5.36)	7.05 (1.21-12.45)	1.6 (0.1-3.34)
N_{ancDW}	20.53 (20.22-20.84)	19.88 (19.56-20.22)	19.93 (19.63-20.24)
N_{root}	1.15 (0.3-1.74)	0.7 (0.14-1.46)	0.66 (0.02-1.34)
$T_{ancDOG1}$	0.27 (0.25-0.29)	0.2 (0.16-0.23)	0.17 (0.13-0.2)
T_{ancDOG}	0.27 (0.25-0.29)	0.2 (0.16-0.23)	0.18 (0.14-0.21)
T_{ancIB}	0.06 (0.03-0.09)	0.01 (0-0.02)	0.01 (0-0.02)
$T_{ancIT-IB}$	0.12 (0.1-0.14)	0.1 (0.08-0.12)	0.08 (0.05-0.11)
$T_{ancCR-IT-IB}$	0.14 (0.11-0.17)	0.11 (0.08-0.12)	0.1 (0.07-0.13)
$T_{ancWOLF}$	0.24 (0.21-0.27)	0.24 (0.16-0.31)	0.3 (0.26-0.32)
T_{ancDW}	0.27 (0.25-0.3)	0.34 (0.31-0.37)	0.32 (0.3-0.34)
T_{root}	14.63 (14.24-15.08)	14.66 (14.2-15.03)	14.97 (14.54-15.42)
$m_{BOX->SPW}$	0.04 (0.03-0.07)	0 (0-0)	0 (0-0)
$m_{BAS->SPW}$	0 (0-0)	0 (0-0)	0 (0-0)
$m_{DIN->SPW}$	0 (0-0)	0 (0-0)	0 (0-0)
$m_{BOX->ISW}$	0 (0-0)	0 (0-0)	0 (0-0)
$m_{BAS->Isw}$	0 (0-0)	0 (0-0)	0 (0-0)
$m_{ISW->DIN}$	0 (0-0)	0 (0-0.01)	0 (0-0)
$m_{PTW->ITW}$	0 (0-0.03)	0 (0-0)	0 (0-0.02)
$m_{JAC->CRW}$	0 (0-0)	0 (0-0)	0 (0-0)
$m_{ancDW->JAC}$	1.27 (1.23-1.32)	1.31 (1.26-1.35)	1.35 (1.31-1.39)

Figure S1: Comparison between demographic parameter estimates by G-PhoCS for different simulated scenarios regarding the Portuguese-Spanish wolf split time. Solid bars: main analysis; dotted bars: control simulation (data simulated under scenario inferred in main analysis); dashed bars: Portuguese-Spanish divergence 10x lower than main analysis; waved bars: Portuguese and Spanish wolves belong to panmictic population.



Supplementary Material for Paper III - Genome Sequencing Highlights the Dynamic Early of Dogs

(formatted as published)

S1 Samples

Rena M. Schweizer[1], Adam H. Freedman[1], Holly Beale[2], Elaine Ostrander[2], Robert K. Wayne[1], John Novembre[1]

[1]University of California, Los Angeles

Department of Ecology and Evolutionary Biology Los Angeles, California, United States of America

[2]National Institutes of Health

Cancer Genetics Branch National Human Genome Research Institute Bethesda, Maryland, United States of America.

S1.1 Samples for High Coverage Sequencing

We chose specific samples for each lineage based upon our ability to obtain high molecular weight genomic DNA containing low levels of protein and RNA. Following phenol-chloroform extraction for a panel of samples for each lineage, we estimated DNA concentrations with a Nano-Drop 2000 spectrophotomer, selecting samples with >200 ng/uL concentration and a 260/280 ratio within an ideal range that indicates minimal protein contamination. We also estimated DNA quality and quantity with a Quant-iT PicoGreen dsDNA assay (Life Technologies). Gel electrophoresis was used to run out all selected samples on 1% agarose gels, after which we selected a sample containing a high molecular weight band (>1500bp, indicating intact genomic DNA), or the sample with the least amount of smearing. In order to eliminate DNA fragments <1000 bp in length, each sample was subjected to Ampure Bead cleanup following manufacturer protocol (Beckman Coulter Genomics). Following this step, we concentrated the DNA via ethanol precipitation, and repeated the Nano-Drop, PicoGreen, and electrophoresis protocols.

The final samples selected for our study (see Table S1.1) were then genotyped with a species-diagnostic SNP panel in order to rule out the possibility that any were interspecies hybrids. We have developed a panel of 26 diagnostic SNP genetic markers that are able to distinguish between the gray wolf (*Canis lupus*), the dog (*Canis lupus familiaris*), the Coyote (*Canis latrans*), and their first and second generation hybrids [1]. These 26 SNPs (17 resolving wolf vs. dog, 2 resolving dog vs. coyote, and 7 resolving coyote vs. wolf) were identified and validated in a panel of 832 dogs, 180 gray wolves and 53 coyotes analyzed on the Affymetrix Canine SNP v2. microarray of 127,000 SNP

markers [2,3]. The 26 SNP markers were assayed for the final sample set of DNA samples using high resolution melting (HRM) SNP detection analysis [4] on a Roche 480 LightCycler quantitative PCR instrument, along with sets of pure gray wolf, domestic dog, and coyote control standards that were homozygotes for the specific SNP diagnostic alleles. The HRM results for each of our final samples were compared to the control standards for each SNP. The basenji, dingo, Israeli wolf, Croatian wolf, and Chinese wolf did not have any alleles that suggested they were interspecies hybrids (results not shown). The golden jackal was further tested by HRM genotyping of 7 SNPs that distinguish golden jackals and coyotes from wolves (as above), and sequencing of cytb [5] (Koepfli, in preparation). No wolf SNP alleles were observed in the golden jackal (results not shown).

Table S1.1. Sample origins, and sequencing effort by sample, platform, and library.

Sample	Sample ID	Sample Origin	Sex	SOLiD LMP[a]	SOLiD fragment	HiSeq[b]
Basenji	RKW 13764	Bethesda, MD, USA	M	1	—	1
Dingo	RKW13760	Bargo Dingo Sanctuary, Australia	M	1	2[c]	1
Israeli wolf	RKW13759	Neve Ativ, Golan Heights, Israel	F	1	1[d]	1
Chinese wolf	RKW13451	San Diego Zoo, CA, USA	F	—	—	3
Croatian wolf	RKW 3919	Perković, Croatia	F	1	1[d]	1
Golden jackal	RKW 1332	Tel Aviv, Israel	F	2	1.75[d]	1

[a] Number of slides, long mate pair, 1.5kb insert, 50bp per end

[b] Number of lanes, paired end 400bp insert, 100bp per end

[c] Number of slides, 75bp

[d] 50bp

The golden jackal cytb sequence was compared to the Indian golden jackal sequence in Genbank (accession no. AY291433) and against sequences for pure gray wolves, dogs, coyotes, Ethiopian wolves, side-striped and blacked jackals and African wild dogs. It grouped closely with the Indian golden jackal. The sex of each sample whose sex was previously unknown or only suggested from field observation was tested in a PCR reaction using the DBX6 and DBY7 markers from [6]. The basenji, whose sex was known, was included as a control.

S1.2 12 Dog Breeds for Moderate Coverage Illumina Sequencing

We utilized data from an ongoing companion study (Beale et al., unpublished data) in the form of moderate coverage sequencing of 12 additional dog breeds, selected to

represent the phenotypic and phylogenetic ranges of contemporary domestic dog breeds. Sequenced breeds were: Beagle, Bulldog, Chihuahua, Chow Chow, Flatcoated Retriever, Great Dane, Mastiff, Pekingese, Saluki, Scottish Terrier, Siberian Husky, and Toy Poodle. Three of these breeds—Chow Chow, the Siberian Husky, and Saluki—were previously found to be basal in phylogenetic studies [3,7]. Blood samples were obtained with the consent of dog owners at American Kennel Club (AKC)-sanctioned dog shows, specialty events, breed clubs, and veterinary clinics using a protocol approved by the NIH Institutional Animal Care and Use Committee (IACUC).

Briefly, DNA was extracted as previously described [8], and libraries were prepared and sequenced on an Illumina Hi-Seq 2000 according to manufacturer protocols, producing 72-108 million reads per sample. Reads were aligned to Canfam3.1 with BWA version #0.5.9-r16 [9] and alignments were refined according to GATK best practices, using GATK version 1.5-11 (Best Practice Variant Detection with the GATK v3, 2012) [10]. Randomly selected reads were removed from alignments in all samples to normalize coverage, which reduced the original depths (ranging from 5.2x to 8.3x per dog) to an average of 5.3x (5.1x- 5.5x per dog).

References

1. VonHoldt BM, Pollinger JP, Earl DA, Parker HG, Ostrander EA, et al. (2013) Identification of recent hybridization between gray wolves and domesticated dogs by SNP genotyping. Mamm Genome 24: 80-88.

2. vonHoldt BM, Pollinger JP, Earl DA, Knowles JC, Boyko AR, et al. (2011) A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. Genome Res 21: 1294-1305.

3. vonHoldt BM, Pollinger JP, Lohmueller KE, Han EJ, Parker HG, et al. (2010) Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. Nature 464: 898-902.

4. Wittwer CT, Reed GH, Gundry CN, Vandersteen JG, Pryor RJ (2003) High-resolution genotyping by amplicon melting analysis using LCGreen. Clin Chem 49: 853-860.

5. Irwin DM, Kocher TD, Wilson AC (1991) Evolution of the Cytochrome-B Gene of Mammals. J Mol Evol 32: 128-144.

6. Seddon JM (2005) Canid-specific primers for molecular sexing using tissue or non-

invasive samples. Conserv Genet 6: 147-149.

7. Parker HG, Kim LV, Sutter NB, Carlson S, Lorentzen TD, et al. (2004) Genetic structure of the purebred domestic dog. Science 304: 1160-1164.

8. Parker HG, Kukekova AV, Akey DT, Goldstein O, Kirkness EF, et al. (2007) Breed relationships facilitate fine-mapping studies: A 7.8-kb deletion cosegregates with Collie eye anomaly across multiple dog breeds. Genome Research 17: 1562-1571.

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

10. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297-1303.

S2 Sequencing

Adam H. Freedman[1], Kevin M. Squire[2], Vasisht Tadigotla[3], Clarence Lee[3], Timothy Harkins[3], Stanley F. Nelson[2], Robert K. Wayne[1], John Novembre[1]

[1]University of California, Los Angeles

Department of Ecology and Evolutionary Biology Los Angeles, California, United States of America

[2]University of California, Los Angeles

Department of Human Genetics Los Angeles, California, United States of America

[3]Life Technologies

Foster City, California, United States of America

S2.1 General Strategy

The goal of our sequencing strategy was to obtain ≥ 20x coverage across the mappable regions of the genome. Such coverage has been seen to give good resolution for genotyping heterozygous sites within single individual samples (e.g., 1000 Genomes Project Consortium 2010), and this proved to be true in our sample, as we found by validating our sequence-based genotype calls with array-based genotypes (Text S5). For our first 5 samples (Basenji, Dingo, Israeli wolf, Croatian wolf, Golden jackal), we generated short read data on primarily the SOLiD platform, and added at least one lane per sample of Illumina HiSeq. Subsequent to this sequencing, we took advantage of available high coverage data generated for the Chinese wolf (but only using the HiSeq platform), because of the benefits of adding a sample representing of an additional hypothesized domestication center.

S2.2 Library Construction and Sequencing

S.2.2.1 SOLiD 4 Library Preparation (non-ECC)

In all cases, the manufacturers protocols were closely followed, either using the following protocol guides, or earlier versions. For SOLiD protocols, we refer the reader to the Applied Biosystems SOLiD 4 System Library Preparation Guide (4445673 Rev. A, March 2010), and Applied Biosystems SOLiD 4 System Templated Bead Preparation Guide (4448378 Rev. B, March 2010). For EZ Bead-based template bead preparation protocols, see the EZ Bead user guides for the Applied Biosystems SOLiD EZ Bead

Emulsifier (4441486 Rev. E, October 2011), Applied Biosystems SOLiD EZ Bead Amplifier (4443494 Rev. E, October 2011), and the Applied Biosystems SOLiD EZ Bead Enricher (4443496 Rev. E, October 2011).

SOLiD LMP library preparation—For the long mate-paired library, we followed the 2x50 mate- paired library protocol from the Applied Biosystems SOLiD 4 System Library Preparation Guide (March 2010). The same protocol was used for all LMP libraries. For each sample, we started with 20 ug of genomic DNA. The DNA was sheared with a HydroShear Standard Shearing Assembly at speed code 5 (SC5) for 20 cycles, for a target fragment size of 1-2kb. The sample was purified using PureLink columns from a SOLiD Library Column Purification Kit. The purified DNA was end-repaired with End Polishing Enzymes 1 and 2 to convert the ends to 5'- phosphorylated blunt-ended DNA. LMP CAP adapters from the SOLiD Mate-Paired Library Oligos Kit were ligated the end-repaired DNA, and the DNA was column purified again. The LMP CAP adapter is missing a 5' phosphate from one of its oligonucleotides, which causes a nick on each DNA strand when the DNA is circularized. To remove unbound CAP adaptors, the DNA was run on a 1% agarose gel. The gel was cut to select 1.5Kb DNA fragments and purified using a SOLiD Library Quick Gel Extraction Kit. The DNA fragments were then circularized with a biotinylated internal adaptor from the SOLiD Mate-Paired Library Oligos Kit., and column purified. Plasmid-Safe ATP-Dependent DNase was used to eliminate uncircularized DNA, and resulting DNA was column purified again. All libraries consisted of more than 100 ng of circularized product at this point. The circularized DNA was treated with E. coli DNA polymerase I to translate the nick into the genomic DNA region and column purified. After nick-translation, the DNA was digested with T7 exonuclease (to create a single strand gap around the nick) and S1 nuclease (to cleave most of the library molecule from the circularized template), and column purified. The DNA was again treated with End Polishing Enzyme 1 and 2 for end-repair and phosphorylation of the 5' ends for subsequent ligation. To purify the library, the DNA molecules were bound to Dynabeads MyOne Streptavidin C1 beads, which binds specifically to the biotin-labeled internal adapter, and then separated with a Dynal magnet. P1 and P2 adapters from the SOLiD Mate-Paired Library Oligos Kit were ligated to the end-repaired DNA, and the molecules bound to streptavidin beads were washed and purified from ligation side-product. Again, the DNA was nick-translated with DNA polymerase I. The library was then trial amplified using Library PCR Primers 1 and 2 with the Platinum PCR Amplification Mix, and run on a 2% E-Gel EX Gel to determine the required number of PCR amplifications. All libraries required between 12 and 16 amplifications. The resulting library was size selected to between 250 and 350 bp using the SOLiD Library Size Selection gel, and was extracted and desalted using columns. The libraries were then

quantitated using the SOLiD Library TaqMan Quantitation Kit, using a qPCR run on the MJ Research DNA Engine Opticon 2 Real-Time Cycler.

SOLiD Fragment Library Preparation—Fragment library preparation is much simpler than LMP library preparation. We started with 5ug of DNA, which was sheared to a range of 150-180 bp (before adapter ligation) using a Covaris S2 System. As above, DNA ends were repaired using End Polishing Enzymes 1 and 2, and column purified. P1 and P2 adapters were ligated, followed by another column purification. The DNA was run on a SOLiD Library Size Selection gel, and a section corresponding to the post-ligation size of 200-230bp was cut. The DNA was nick- translated with DNA Polymerase I, amplified for 2 PCR cycles, and then column purified. qPCR quantification showed a yield of 50-150ng of DNA.

SOLiD Templated Bead Preparation—Most next-generation technologies have a DNA template immobilization strategy, followed by template amplification. The strategy employed by SOLiD sequencing is bead-based template immobilization, followed by emulsion PCR. Here we describe bead preparation. In early 2011, Life technologies introduced their EZ Bead system, which greatly simplified this step of sequencing preparation. However, we were not able to take advantage of this for our early LMP libraries, so we describe both bead preparation steps below.

Manual Template Bead Preparation—Bead preparation was conducted according to the macro (4 ePCR Reaction) protocol described in the Applied Biosystems SOLiD4 System Templated Bead Preparation Guide (March 2010). Bead preparation consists of emulsion PCR (ePCR), followed by a wash, template enrichment, and 3' end modification. For the ePCR step, the oil phase, aqueous phase (DNA template), and beads (SOLiD P1 DNA Beads) for the emulsion were prepared separately according to the full-scale ePCR reaction protocol for 2x50 or fragment library, and the emulsion was performed using a ULTRA-TURRAX Tube Drive from IKA and transferred to a 96 well plate for ePCR thermal cycling. After thermal cycling, the emulsion was broken and the beads were washed with Bead Wash Buffer and TEX Buffer according to protocol. Beads were then quantitated using NanoDrop. Depending on the library, between 800 million and 1 billion beads were produced. Next, beads with full length templates were isolated via oligo hybridization using the P2 primer. The P2-enriched beads were extended with a Bead Linker and Terminal Transferase enzyme according to library protocol. These were quantitated with a Nanodrop ND-1000 and the bead concentration was calculated using a Work Flow Analysis (WFA).Bead yield was around 700-900 million templated beads per library.

SOLiD EZ Bead-based Preparation—The EZ Bead system automates the emulsion, amplification, and enrichment steps above with three instruments (the EZ Bead Emulsifier, EZ Bead Amplifier, and EZ Bead Enricher). We used this system for SOLiD fragment library construction, and found the preparation much easier and the bead yield similar.

S2.2.2 SOLiD Sequencing (non-Exact Call Chemistry)

The library templated beads were loaded onto a 1-well flow-cell at a density of roughly 750,000 beads/ul, and run on a SOLiD 4 Sequencer with SOLiD 4 sequencing reagents using 50+50 LMP sequencing, 75+50 ECC LMP sequencing, or 50+35 Fragment sequencing protocols.

S2.2.3 SOLiD 4 Library Preparation (ECC)

Fragment and long mate pair libraries were prepared for both Basenji and Dingo DNA following prescribed protocols for the SOLiD 4 system library preparation guide, which will be briefly described here.

All DNA was quantified using the Qubit dsDNA HS assay. Fragment libraries were constructed using 1 microgram of genomic DNA for each species. The DNA is sheared sonically into small fragments using a Covaris S2 system, with a mean fragment size of ~150 to 180 bp. The sheared DNA is end-repaired and purified, prior to ligation of SOLiD P1 and P2 library adaptors. The ligated DNA with 200-230 bp length is subsequently isolated using an Invitrogen E-Gel SizeSelect agarose gel. The selected DNA is nick translated and PCR amplified using appropriate primers for 9 cycles. The resulting fragment libraries were purified and quantified using an Agilent Bioanalyzer and the corresponding high sensitivity DNA kit.

Long mate pair libraries were generated from 25-30 micrograms of genomic DNA. Genomic DNA was sheared using a Digilab Hydroshear with a standard shearing assembly. The DNA is end-repaired and purified prior to ligation of LMP CAP adaptors, which are missing a 5’ phosphate. The adapted DNA is purified and size-selected for a 1.5 kb insert size by running the DNA in a 1% agarose gel and excising the appropriate length as determined by using a 1 kb DNA ladder. The size-selected DNA is extracted from the agarose cut and circularized by ligating a biotinylated internal adaptor. Plasmid-Safe ATP-Dependent DNase is used to isolate the desired circularized DNA product. Because of the missing 5’ phosphates in the CAP adaptors, circularization of the DNA results in a nick on each strand. A nick translation is performed followed by digestion

with T7 exonuclease and S1 nuclease, resulting in library molecules with the desired mate pair tags being cleaved from the circularized DNA template. DNA molecules were again end-repaired, and the desired DNA molecules with internal adaptors were isolated using Dynabeads MyOne Streptavidin C1 beads. SOLiD P1 and P2 adaptors were ligated onto the end-repaired DNA. As with the fragment libraries, the ligated DNA underwent nick translation and PCR amplified using appropriate primers, and quantified.

The clonal amplification of libraries through emulsion PCR was performed using the SOLiD EZBead system as prescribed.

S2.2.4 SOLiD ECC Sequencing

Sequencing was done using the SOLiD 4 system with Exact Call Chemistry (ECC), generating 75 bp and 2x50 bp reads for the fragment and mate par libraries respectively. The principles of ECC are based on standard techniques used in communication and data storage system to minimize measurement error through redundancy and employing different encoding schemes. Due to its ligation-based approach, SOLiD sequencing can leverage the use an additional probe set that complements the standard two-base encoding scheme.

S2.2.5 Illumina Paired-End Sample Preparation and Sequencing

For Illumina library preparation protocols, see the Illumina Paired-End Sample Preparation Guide (1005063 Rev. E, February 2011). For all samples except the Chinese wolf, genomic paired-end sequencing libraries with average insert size of 300-500 bp were constructed according the manufacturer's recommended protocol. Briefly, ~5 µg of purified genomic DNA was fragmented by sonication using the Covaris Adaptive Focused Acoustics (AFA) System. 3’ and 5’ overhangs of the recovered genomic DNA fragment were converted into blunt ends using T4 DNA polymerase and Klenow enzyme (New England Biolabs). After end repair, an ‘A’ base was added to the blunt phosphorlated DNA fragments using Klenow 3’->5’ Exo- (New England Biolabs). The standard paired-end adaptors were ligated to the ‘A’ tailed DNA fragments using a Quick DNA ligation kit (New England Biolabs). The ligated products were separated on a 2% argrose gel and the desired DNA fragments were recovered from the gel by the QIAquick Gel Extration kit (Qiagen). After the initial denaturation at 98°C for 30 seconds, the PCR reaction was carried out for 8 cycles of 98°C for 10 sec, 65°C for 30 sec, and 72°C for 30 sec using Phusion DNA polymerase (Finnzymes). The final extension was for 5 min at 72°C. Libraries were sequenced on an Illumina HiSeq2000 following the manufacturer’s standard cluster generation with a V2 Paired End Cluster generation kit,

and sequencing protocol with TruSeq SBS sequencing reagents. Base calling was performed with the on-instrument computer using RTA version 1.7. For the Chinese wolf sample (carried out at BGI), the same protocols were carried out except for an additional library QC step using the Agilent 2100 Bioanlayzer and ABI StepOnePlus Real-Time PCR system.

S2.2.6 Data Generation

Sequencing runs were distributed across instruments, sequencing centers and dates, such that the chances for cross-contamination would be minimized (Table S2.2.1). This is important because one of our objectives was to assess admixture between wild and domestic lineages, and such contamination has the potential to create spurious signals of gene flow.

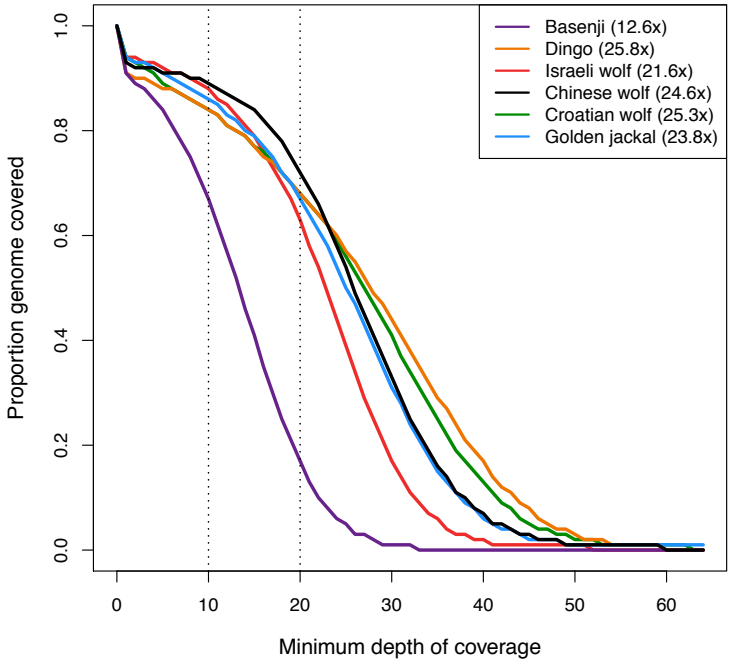
Combining SOLiD and Illumina sequencing reads, we generated between 969 and 3366 million reads per sample (Table S1). We generated the most reads for the golden jackal, in order to achieve >20x coverage despite a high PCR duplication rate indicative of library simplification (probably due to partial degradation of the tissue sample). The smallest data set and lowest coverage were generated for the basenji, as downstream analyses belatedly revealed issues with a SOLiD single-end fragment library generated for that sample undetected by standard quality control metrics, forcing us to exclude an additional >10x of poor quality sequencing data. Overall, we were able to align 69 - 94% of reads to the boxer reference, leading 20 ~ 28 Gb (Basenji) to 71 Gb (Dingo) of uniquely aligned bases (Table S1). With the exception of the basenji, all samples were genotyped to > 20x coverage (Table S1), such that (with the exception of the basenji), > 80% and 60% of the genome was covered by at least 10 or 20 reads, respectively (Figure S2.2.1).

Table S2.2.1. Dates and institutions where sequencing was carried out.

Sample	Library	Library Location	Sequencing Dates	Sequencing Location
Basenji	LMP[a]	LT	02.17.11	LT
Basenji	HiSeq	UCLA	12.10.10	UCLA
Dingo	FRAG[a]	LT	01.14.11	LT
Dingo	LMP[a]	LT	02.15.11	LT
Dingo	HiSeq	UCLA	12.10.10	UCLA
Israeli wolf	FRAG	UCLA	08.27.10	UCLA
Israeli wolf	LMP	UCLA	06.25.10	UCLA
Israeli wolf	HiSeq	UCLA	09.08.10	Stanford
Croatian wolf	FRAG	UCLA	11.19.10	UCLA
Croatian wolf	LMP	UCLA	12.02.10	UCLA
Croatian wolf	HiSeq	UCLA	08.15.10	UCLA
Chinese wolf	HiSeq	BGI	03.02.12	BGI
Golden jackal	FRAG	UCLA	09.13.10; 01.20.11	UCLA
Golden jackal	LMP-1	UCLA	07.28.10	UCLA
Golden jackal	LMP-2	UCLA	11.23.10	UCLA
Golden jackal	HiSeq	UCLA	12.10.10	UCLA

[a] Samples sequenced using Exact Call Chemistry (see Text S2 for details).

Figure S2.2.1. Proportion of the genome covered as a function of raw minimum depth of coverage. The vertical dashed lines at 10x and 20x are provided to aid interpretation.



S3 Genotyping Pipeline

Adam H. Freedman[1], Vasisht Tadigotla[2], Robert K. Wayne[1], John Novembre[1]

[1]University of California, Los Angeles

Department of Ecology and Evolutionary Biology Los Angeles, California, United States of America

[2]Life Technologies

Foster City, California, United States of America

S3.1 Pipeline Design

We implemented a sequencing alignment and genotyping pipeline customized for combining SOLiD and Illumina HiSeq short read data (Figure S3.1), using aligners tailored to the specific platforms, then post-processing alignments using the Picard (http://picard.sourceforge.net) and Genome Analysis Toolkit (GATK) toolsets [1]. This pipeline converted short read raw data to .bam format alignment files [2], and from bam files to genotype files in .vcf format (http://www.1000genomes.org/node/101).

S3.1.1 Sequence Alignment

All short read data were aligned to the most current assembly of the dog genome (CanFam 3.0), generated from a boxer breed individual. CanFam 3.0 represents an early release of the update from CanFam 2.0, that was made publicly available by the Broad Institute but not added to NCBI or the UCSC Genome Browser as available downloads. CanFam 3.0 differs from the currently available CanFam 3.1 only in the length of N buffers at the beginning of each chromosome (3MB for autosomes in CamFam 3.0), and similarly, the length of those buffers between scaffolds assembled into chromsomes. To maximize the probability of proper alignment of short reads, data generated by the SOLiD and Illumina HiSeq platforms were each aligned using different alignment algorithm.

For SOLiD ECC reads, the ECC decoding pipeline was run offline using the image files (.spch) generated by the SOLiD instrument to generate corrected csfasta and qual files. The algorithm for generating corrected color calls is described at http://www3.appliedbiosystems.com/cms/groups/global_marketing_group/documents/generaldocuments/cms_091372.pdf

Both SOLiD ECC and non-ECC reads were aligned to Canfam 3.0 using the mapping and pairing modules in the BioScope1.2 pipeline. For the former, since this protocol used

an early version of the ECC decoding the quality values (QV) were not properly calibrated in the mapped BAM files. QVs that were greater than 40 were reduced to 40. All Illumina reads were aligned to CanFam 3.0 using novoalign (version 2.07.11) (www.novocraft.com), with soft-clipping turned on.

Aligned reads from both sequencing platforms were merged and stored in bam format [2]. Reads corresponding to PCR duplicates were marked (and later removed) with Picard MarkDuplicates (picard.sf.net). Additional processing steps described below were then applied to the merged .bam files.

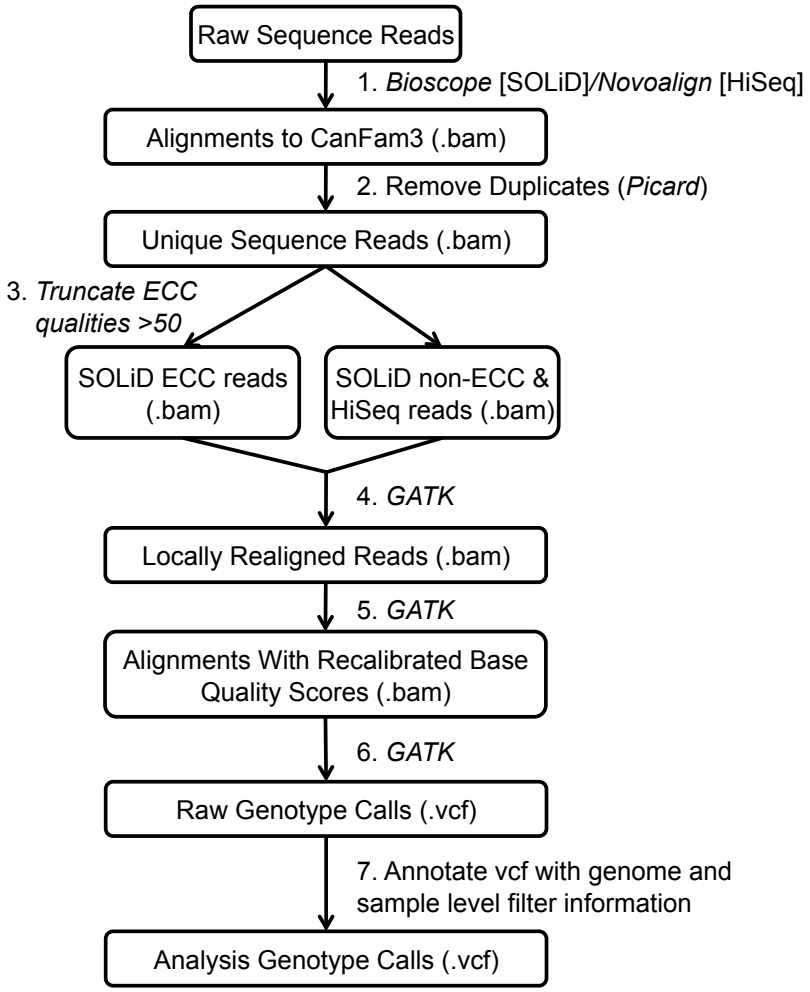


Figure S3.1. Schematic of sequence alignment and genotyping pipeline carried out separately for each lineage.

S3.1.2 Local Realignment

Short read alignment algorithms operate on each read independently, with the result that false SNVs can be detected in regions where repeated alignment errors occur across

overlapping reads. A large proportion of such regions contain indels, with misalignment occurring most frequently for reads overlapping the indel near the read start or end. We used the GATK IndelRealigner [1] to perform local multiple alignment leading to a consensus indel call, and reducing the occurrence of false positive SNV sites. This three-step process entails first identifying suspicious intervals that may require realignment, followed by local realignment within these intervals, then 'rescuing' the mate pairing lost during the local realignment process, using the program Picard. Specific generic command lines are:

1) Interval detection

```
java -Xmx4g -Djava.io.tmpdir=GATKtemp -jar
Path/To/GenomeAnalysisTK.jar -T RealignerTargetCreator
-rf BadCigar -I Path/To/Infile/Infile.bam -L
ChromosomeName -o Path/To/IntervalsOutfile/intervals -
R Canfam3.fa
```

2) Local realignment

```
java -Xmx4g -Djava.io.tmpdir=GATKtemp -jar
Path/To/GenomeAnalysisTK.jar -T IndelRealigner
-I Path/To/File/To/Realign/file.bam -o
RealignedFile.bam -R Canfam3.fa -
targetIntervals
Path/To/Suspicious/Intervals/File/IntervalsFile
Name
```

3) Fix mate pair information

```
java -Xmx4g -Djava.io.tmpdir=GATKtemp -jar
FixMateInformation.jar INPUT=Path/To/Infile/infile.bam
OUTPUT=Path/To/Outfile/Realigned_infile.bam
SO=coordinate VALIDATION_STRINGENCY=SILENT
```

S3.1.3 Base Quality Recalibration

Quality scores assigned to individual base calls are intended to reflect confidence in the specified nucleotide, but these scores may be weakly correlated with the actual probabilities of erroneous base calls. Important with respect to our study, the range of possible quality scores and the nature of quality score assignment differ substantially between SOLiD and Illumina sequencing platforms. To standardize quality scores across sequencing runs, libraries, and technologies, we performed empirical quality score recalibration using GATK. Recalibration involves three steps: 1) liberally defining a set of SNV-containing sites that are excluded from subsequent steps, 2) for all other sites,

tabulating the frequency of base calls that are correct (i.e. consistent with homozygous-reference genotype) vs. incorrect as a function of covariates reflecting features of the underlying sequence context stratified by library/sequencing run, and 3) replacing the instrument-assigned quality scores with the genome-wide empirical error rates conditional on each unique covariate set. Step 1 was undertaken by genotyping in the same manner as below (see S3.1.4), but only requiring that genotypes containing an SNV had a genotype quality score ≥10.We used the three default covariates: read group (i.e. library), dinucleotide context, and position within the read. For SOLiD reads, reference bias introduced due to reference correction was removed by using the --solid_recal_mode SET_Q_ZERO and -- solid_nocall_strategy PURGE_READ options to the walker. Specific generic command lines are:

1) Create recalibration table:

```
java -Xmx4g -Djava.io.tmpdir=GATKtemp -jar
PathTo/GenomeAnalysisTK.jar -l INFO -T CountCovariates
-cov ReadGroupCovariate -cov CycleCovariate -cov
DinucCovariate --default_platform solid -I
PathToInfile/infile.bam - B:mask,VCF
PathToVariatnSitesToExclude/rod_file.vcf -R Canfam3.fa
-recalFile RecalibrationTable.csv --solid_recal_mode
SET_Q_ZERO --solid_nocall_strategy
LEAVE_READ_UNRECALIBRATED
```

2) Generate recalibrated .bam file:

```
java -Xmx4g -Djava.io.tmpdir=GATKtemp -jar
Path/To/GenomeAnalysisTK.jar -l INFO -T
TableRecalibration --default_platform solid -I
PathToInfile/infile.bam --out outfile.bam -R Canfam3.fa
-recalFile RecalibrationTable.csv --
doNotWriteOriginalQuals --solid_recal_mode SET_Q_ZERO -
-solid_nocall_strategy PURGE_READ
```

S3.1.4 Base and Indel Genotyping with GATK

To call genotypes for our five novel canid genomes, we used the GATK Unified Genotyper (UG). UG employs a Bayesian genotype likelihood model that takes as input the base calls and associated quality scores for a locus, and emits the most likely genotype, the posterior probabilities that the locus is segregating and for the three possible genotypes. Only three genotype calls are possible because UG makes the simplifying assumption that a site is bi-allelic [1]. Although UG has multi-sample genotyping capabilities that enable estimation of population allele frequency across a set

of samples, our focus on comparative genomics amongst evolutionarily distinct lineages (rather than, for example, variant discovery within a population of interest) led us to genotype each lineage separately. In addition, separate genotyping runs allowed us to specify separate priors on heterozygosity for each lineage, in keeping with known differences among wild and domestic canids. Specifically, priors were set based upon the heterozygosity estimates obtained by [3]. Because only one golden jackal was sampled in that study, for our golden jackal we used the value provided for wolves. To evaluate the sensitivity of our genotype calling to the assumed priors, we calculated the proportion of heterozygous genotype calls at these values, as well as for separate genotyping runs with values ± 50%. Priors had little effect on the frequency of heterozygous calls regardless of any hard threshold for minimum genotype quality (Figure S3.2). As a result, for each lineage we took as our final priors the average across the three runs, at genotype quality=20, the value we used as a sample-level hard filter (see Text S4). An example generic command line is as follows:

```
java -Xmx3g -jar Path/To/GenomeAnalysisTK.jar -T
UnifiedGenotyper -l INFO -- genotyping_mode DISCOVERY -
-output_mode EMIT_ALL_CONFIDENT_SITES -I $file -L
<chromosome_name> --min_base_quality_score 20 --
standard_min_confidence_threshold_for_emitting 0.0 --
standard_min_confidence_threshold_for_calling 0.0 --
heterozygosity <lineage- specific prior> -A GCContent -
o outfile.vcf -metrics outfile.vcf_metrics.txt -R
Path/To/Canfam3.fa -dt NONE
```

Because we implemented our own conservative set of filters post-genotyping, we set both standard minimum confidence thresholds to zero.

Accurate calling and discovery of indel variants from next-generation sequencing is still subject to considerable uncertainty, and little prior information is available concerning the distribution of indels in the dog genome, let alone for other wild canid species. Furthermore, we had no way to validate indel calls at the genome-wide scale in a manner comparable to that available for SNV calls (see Text S4). Thus, we called indels only for use in the filtering out of SNVs proximate to them that might be false positives (see Text S4), accepting that the indel calls are only approximations. Our generic command line for indel calling using UG employed default settings for indel_heterozygosity, min_indel_count_for_genotyping, and other indel-calling specific settings, as follows:

```
java -Xmx4g -jar Path/To/GenomeAnalysisTK.jar -T
UnifiedGenotyper -l INFO - glm INDEL --
indel_heterozygosity 0.000125 --
min_indel_count_for_genotyping 5 --genotyping_mode
```

```
DISCOVERY --output_mode EMIT_ALL_CONFIDENT_SITES --
min_base_quality_score 20 --
standard_min_confidence_threshold_for_emitting 0.0 --
standard_min_confidence_threshold_for_calling 0.0 --
heterozygosity <lineage-specific prior> -I infile.bam -
L <chromosome name> -o indel_outfile.vcf -metrics
outfile.vcf_indel_metrics.txt -R Path/To/Canfam3.fa
```

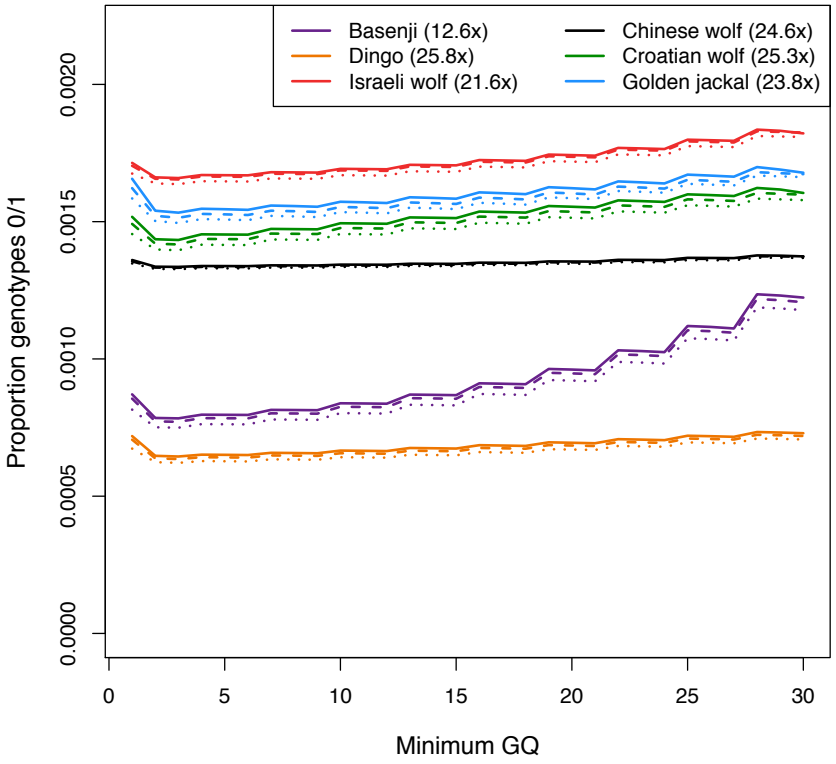


Figure S3.2. Proportion of genotypes typed as heterozygous using three different heterozygosity priors with UG, plotted against minimum genotype quality. Dashed, dotted and solid lines represent priors set at the mean, mean -50%, and mean +50% of nucleotide diversity estimates from Gray et al. [3].

References

1. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491-498.

2. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.

3. Gray MM, Granka JM, Bustamante CD, Sutter NB, Boyko AR, et al. (2009) Linkage Disequilibrium and Demographic History of Wild and Domestic Canids. Genetics 181:

1493-1505.

S4 Quality Filtering

Adam H. Freedman[1], Pedro Silva[2], Marco Galaverni[3], Robert K. Wayne[1], John Novembre[1]

[1]University of California, Los Angeles

Department of Ecology and Evolutionary Biology Los Angeles, California, United States of America

[2]University of Porto *CIBIO-UP - Research Center in Biodiversity and Genetic Resources Porto, Portugal*

[3] Istituto Superiore per la Protezione e la Ricerca Ambientale *Laboratorio di Genetica Ozzano dell'Emilia, Italy*

S4.1.1 Filtering Conventions

In line with previous studies utilizing next-generation sequencing data, we developed a series of conservative data quality filters, implemented post-genotyping. Filters served two purposes. First, we sought to minimize the effects of sequencing and alignment errors that might bias downstream analyses [1,2]. Second, we sought to exclude regions of the genome that, irrespective of such errors, might show accelerated rates of evolution for reasons other than positive selection on the dog lineage, and might falsely appear as outliers in our selection scans; such regions might also be prone to misalignment of short reads. We established sets of criteria with which to filter at both the level of genomic position and individual lineages. *Genome feature filters* were applied to genomic positions based upon intrinsic features of the reference (Canfam3) and polymorphism across samples (i.e. tri-allelic and CpG sites), while *sample feature filters* were applied to individual lineage genotypes based upon features of the data underlying the corresponding genotype call. We annotated our VCF files according to whether genomic positions and samples passed the respective filtering criteria.

S4.1.2 Genome feature filters

Genomic positions in a VCF file were flagged as not passing the genome feature filter according to the following criteria.

1. *Repeat Regions.* We identified all genomic positions falling within repeat regions of the reference genome identified with RepeatMasker [3] and Tandem Repeat Finder (TRF) [4]. We annotated our VCF file according to the class of repeat detected,

collapsing the output repeat classes into a reduced set of 14 classes: SINE, LINE, LTR, DNA, RNA, rRNA, scRNA, snRNA, srpRNA, tRNA, Satellite, Simple_repeat, Low complexity sequence, and Unknown. Because ancient repeats can make up a substantial portion of genomes, and because these regions will have diverged enough to allow accurate read mapping with short read alignment algorithms, we sought to retain these, and only mask out younger repeats prone to sequence misalignment. We considered that erroneous mapping of short reads to these regions should lead to increased frequency of heterozygous genotype calls, and plotted the frequency of heterozygote genotype calls against divergence from the repeat libraries employed by RepeatMasker (Figure S4.1.1). We conservatively chose 25% divergence as our minimum repeat divergence threshold, as repeats in this interval show no increase in heterozygosity with decreasing repeat age.

2. *CpGs*. Mutation rates at CpG sites are substantially higher than non-CpG sites [5], so that regions enriched for CpGs may display elevated diversity and/or divergence leading to outliers in window-based analyses, independent from any demographic or selective forces germane to our investigation of domestication. If in any of our six lineages, a nucleotide that otherwise passed filter fell within a CpG dinucleotide, because at least some proportion of our data fell into that hyper-mutable site category, we flagged the genomic position.

3. *Copy Number Variants (CNVs).* When true CNVs are not included in a reference genome assembly, or when samples mapped to the reference contain novel CNVs, misalignment of paralogous reads is more probably, and can lead to false positive SNVs that can bias estimated levels of polymorphism and divergence. To minimize the effects of such misalignment, we constructed a set of CNV regions to exclude from downstream analyses, by combining a set of

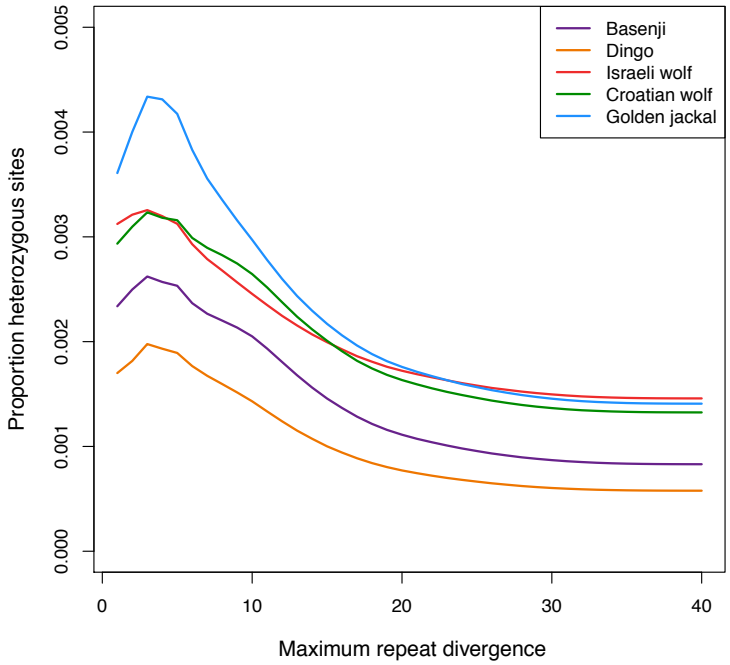


Figure S4.1.1. Proportion of heterozygous sites genotyped in repeat regions, as a function of the maximum divergence (of all repeats intersecting the genomic position of interest) between the observed repeat and the matching repeat motif used by RepeatMasker and Tandem Repeat Finder.

previously discovered CNVs reported in a diverse panel of dog breeds [6], and those we discovered directly from the short read data generated for our six canid lineages. See Text S5 regarding CNV detection methods.

4. *Triallelic sites.* Preliminary comparisons of genotypes from sequencing with those from the Illumina CanineHD BeadChip (see S.4.1 below), indicated triallelic sites were more prone to genotyping errors, and so these sites, while making up a relatively small fraction of the genome, were excluded.

We created genome feature filters at two levels: more stringent, using filters from all fours of the above categories, and less stringent, using only RM/TRF, CNV, and triallelic site filters. We used the more stringent filter for window-based analyses. We implemented the less stringent filtering for analyses of coding positions, as filtering out CpGs would a priori exclude a fraction of amino acids containing the CpG dinucleotide. We also reasoned that, because coding sequence is likely under evolutionary constraints, those constraints should reduce the disparity between mutation rates at CpG vs. non-CpG sites.

S4.1.3 Sample Feature Filters

1. *Proximity to Indel.* Short reads generated by next-generation sequencing platforms are prone to misalignment near indels, and attempts at local realignment around indels may not fully rectify this problem. As a result, these indel-proximate misaligned regions may be enriched for false positive SNVs. To account for this potential source of bias, for each sample we excluded any genotype containing an alternative allele relative to Canfam3 that was within 5bp (either up or downstream) of another SNV containing genotype within the same sample.

2. *Genotype Quality.* Genotype quality (GQ) metrics output by the GATK Unified Genotyper (UG) represent phred-scaled probabilities that the called genotype does not match the true underlying genotype, i.e $-10*\log_{10}(P[error])$. We chose a hard minimum GQ threshold of 20 (P[error]=0.01) based upon two considerations. First, we sought to minimize genotyping errors as measured by discordance with an independent, high quality genotype data set from the Illumina SNP chip (see S4.1). Second, we sought to balance the competing goals of retaining maximum genomic coverage while being able to correctly identify specific mutations of functional significance, particularly those fixed between dogs and wild canid species. Hard genotype quality thresholds may lead to undercalling of heterozygotes in samples with low or moderate coverage, but works well with those at >20x coverage [2]. All but one of our canid lineages were sequenced at >20x. Two additional lines of evidence support our use of a hard GQ threshold. First, the majority of all emitted genotypes have GQ >20 (Basenji 83.1%, Dingo 93.5%, Israeli wolf 95.6%, Croatian wolf 93.2%, Chinese wolf 98.9%, golden jackal 93.7%). Second, for our lowest coverage sample, the basenji, filtering on GQ appears to exclude more low quality homozygous genotypes, as the proportion of heterozygous calls shows an increasing trend with GQ above GQ=20 (Figure S3.2).

3. *Excess Depth of Coverage.* Extremely high depth of coverage relative to the genome-wide average likely indicates misalignment of reads generated from paralogous positions in the genome, particularly those containing CNVs. Indeed, excess depth of coverage is a typical metric used to define CNV regions, but CNV filtering alone will fail to detect finer-resolution CNV signatures. Thus, we conservatively filtered all sites where depth of coverage exceeded twice the mean depth of coverage recorded for each lineage. GATK UG filters out reads that fail to meet certain criteria (see above). As a result, post-GATK filtering, depth of coverage may fall below our 2x threshold, even when the GATK filtering of hundreds of reads would indicate a region that may intrinsically be prone to read

misalignment. Thus, our filtering on depth of coverage is based upon the number of reads overlapping a genomic position prior to imposition of the UG's internal filters.

4. *Clustered SNVs.* Within any sample, we excluded all SNV-containing genotypes falling within 5 bp of another SNV-containing genotype. In identifying clustered SNVs, to be conservative we required that proximate SNVs only have a minimum genotype quality of 10, rather than the 20 employed in our downstream evolutionary analyses.

Sample-level filters were employed as hard filters. For analyses involving estimation of genome- wide patterns of diversity, we used combinations of filters designated GF2 and SF (Table S5.1.1). For quantifying the number of dog and wolf specific variants, and for analysis of functional regions where the potential for elevated mutation rates at CpG sites should be constrained by functional consequences, we included CpG sites, equivalent to filters GF3 and SF (Table S5.1.1).

References

1. Jordan G, Goldman N (2012) The Effects of Alignment Error and Alignment Filtering on the Sitewise Detection of Positive Selection. Mol Biol Evol 29: 1125-1139.

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

3. Smit AFA, Hubley R, Green P (1996-2010) RepeatMasker Open-3.0 4. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic

Acids Res 27: 573-580. 5. Hodgkinson A, Eyre-Walker A (2011) Variation in the mutation rate across mammalian

genomes. Nat Rev Genet 12: 756-766. 6. Nicholas TJ, Baker C, Eichler EE, Akey JM (2011) A high-resolution integrated map of

copy number polymorphisms within and between breeds of the modern domesticated dog. BMC Genomics 12:414.

S5 Validation of Genotype Calls

Adam H. Freedman[1], Rena M. Schweizer[1], Pedro Silva[2], Marco Galaverni[3], Robert K. Wayne[1], John Novembre[1]

[1]University of California, Los Angeles *Department of Ecology and Evolutionary Biology Los Angeles, California, United States of America*

[2]University of Porto *CIBIO-UP - Research Center in Biodiversity and Genetic Resources Porto, Portugal*

[3] Istituto Superiore per la Protezione e la Ricerca Ambientale *Laboratorio di Genetica Ozzano dell'Emilia, Italy*

S5.1 Effective Genomic Coverage and Number Variants Discovered

Despite the fact that coverage for the basenji was substantially less than for the other canid samples, a large proportion of the genome was still genotyped with genotype qualities (GQ) ≥ 20. 77% of the genome was covered for the basenji at GQ ≥ 20, while coverage for the other five genomes ranged from 88% to 97% (Figure S5.1.1). The precipitous decline of coverage with increasing GQ for the basenji presumably reflects the difficulty for genotype callers of making high confidence calls with fewer reads.

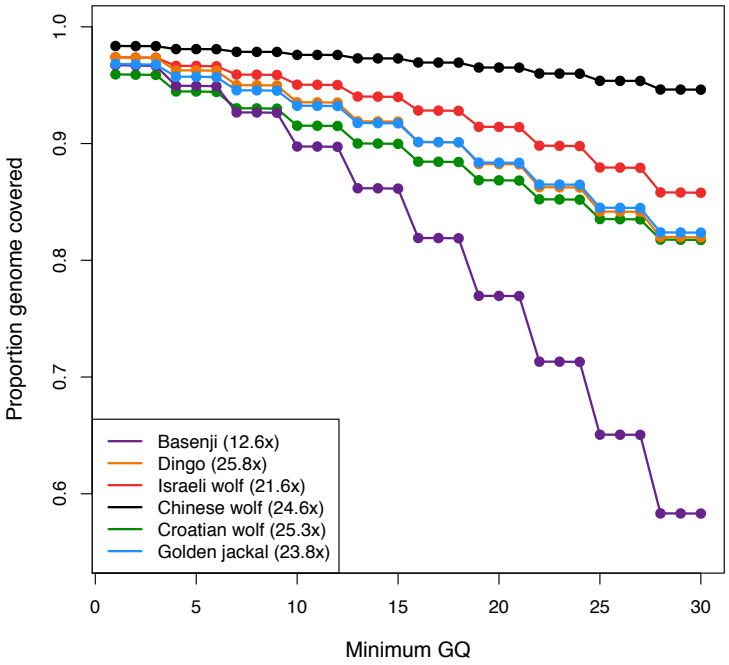


Figure S5.1.1. Genomic coverage per sample as a function of genotype quality, before imposing additional genome and sample level filters.

Imposing genome-level filters led to a large reduction in the amount of the genome available for genotyping. In order to access more sites in the boxer reference, instead of imposing a hard filter on repeat-masked sites, we chose to only filtered out young repeats to which it would be challenging to unambiguously align reads. This allowed us to gain an additional ~10% of the genome (GF1 vs. GF2, Table S2). Nevertheless, without filtering CpGs, genome level filtering (GF3) enabled us to genotype 68-69% of the genome for all six samples (Table S2). Using our combination of genome and sample level filters (GF3+SF), genomic coverage was 55% for the basenji and 63 - 68% for the other canids. Our filtering scheme reduced the discrepancy in coverage between the Basenji and the other samples observed from filtering on genotype quality alone (Figure S.5.1.1). This is not surprising, as difficult to align repeat regions that we filtered out would be harder to genotype effectively with lower coverage, as mapping qualities for individual reads in these regions would, on average, be lower than for more unique regions of the genome.

At present, high coverage genome sequences are not available for any wolf lineage. Looking across out novel wolf genomes, without any genome feature filtering we cover 94% of the boxer reference with at least one wolf genotype call, and even with such filters in place are able to provide genotypes for >60% of the non-N positions in the boxer reference for at least two different wolves (Figure S5.1.2).

After implementing genome and sample level filters (Text S4), we detected a large number of SNV sites relative to the boxer reference, ranging from 2.1-2.4 million for the 2 novel dog samples, 3.3-3.6 million for the 3 wolf samples, and 5.2 million for the golden jackal sample. (Table S2). Unsurprisingly, the golden jackal contained more unique SNV sites relative to the boxer reference than either dogs or wolves, and while wolves contained more unique SNV sites than dogs, a large portion of SNV sites were shared between dogs and wolves (Figure S5.1.3). Sites with no missing data in any lineage allowed us to classify variants as private to dogs, wolves, wild canids, or individual lineages, including 428,339 variants found in dogs but not in wolves, 867,656 variants in wolves that were absent in dogs, 1,524,761 variants shared between dogs and wolves, and 16,604 variants that were fixed between dogs and wild canids (Table S3).

S5.2 Comparison with Illumina Chip-Based Calls

The Illumina CanineHD BeadChip consists of >170,000 markers evenly spaced throughout the dog genome, ascertained from a diverse panel of dog breeds, and

selected primarily from the Dog Genome Project 2.5 million SNP set. With call rates >99% and error rates less that 0.1%, the BeadChip data provide a high-quality benchmark against which to compare sequencing- based genotypes generated with our pipeline. Despite the high quality of the chip genotypes, we imposed an additional set of filters aimed at further reducing error rates. Specifically, based upon prior genotyping of a panel of 96 dogs, we only included positions with minor allele frequency >0, call rate ≥ 90%, and SNP heterozygosity < 120% of the expectation under Hardy-Weinberg equilibrium. A small number of positions were also excluded because they either a) produced a homozygous non-reference genotype for the original boxer used to build the reference genome (Canfam2), or b) produced genotypes inconsistent with those generated with the Affymetrix canine SNP array (v2). Because genomic positions on the chip are based upon the earlier Canfam2 reference, prior to comparison with our genotypes, we converted all chip coordinates to

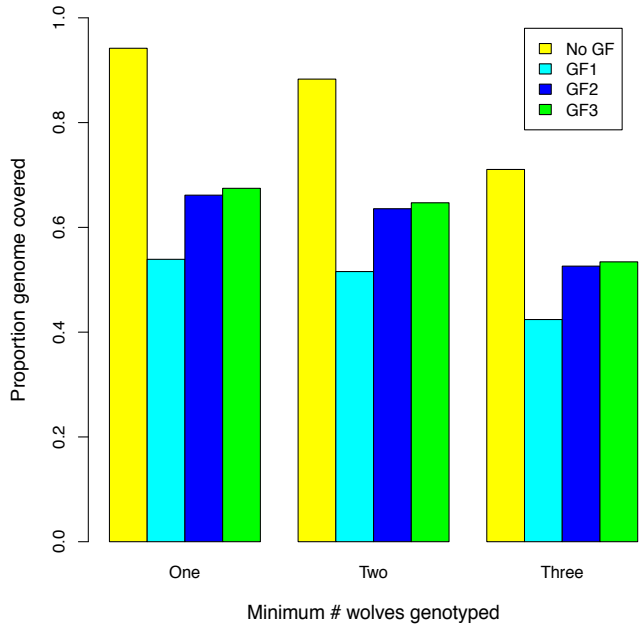


Figure S.5.1.2. Proportion of non-N bases in the boxer reference covered by ≥1 wolf as a function of different genome filters, and with the sample level filters in effect.

homologous coordinates on Canfam3.0 using the *liftover* utility available from the UCSC genome browser website (http://hgdownload.cse.ucsc.edu/admin/exe/). Because all downstream analyses filter out, at a bare minimum, genotypes that do not pass genome (GF3) and sample (SF) level filters, we only compared filter-passing genotypes with those from the BeadChip. This comparison revealed a high degree of concordance (Table S4). An important benchmark with respect to genotype quality is the error rate at known heterozygous positions, as heterozygous genotype calls are more difficult to call

correctly, and typically require higher coverage to do so, relative to either homozygous reference or homozygous non- reference positions. Heterozygote discordance rates were low and typical of other sequencing studies of similar coverage (Dingo, 0.010%, Israeli wolf 0.015%, Croatian wolf 0.018%, Chinese wolf 0.013%, Basenji 0.034%, Table S4). Although the heterozygous site error rate was higher for the golden jackal (0.058%), the frequency of chip heterozygous positions in the jackal was nearly half that of the lowest frequency observed in the other samples (jackal 7.0%; dogs and wolves, 12.0–26.8%), reducing the effect of such errors on downstream evolutionary analyses. The observed discordances showed evidence of bias towards the reference genome (Table S5), with errors towards the reference being anywhere from three to seven times higher than errors away from the reference. However, given the heterozygote discordance rates indicate the overall error rates are very low, and that the reference bias is similar across samples (Tables S4-S5), the effects on analyses should be small.

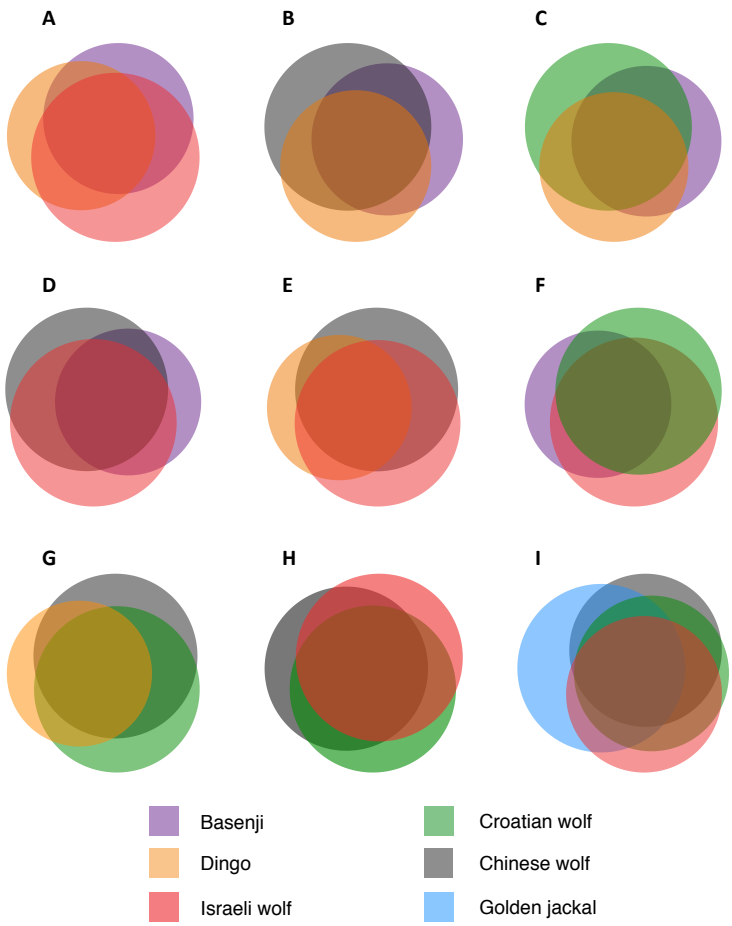


Figure S5.1.3. Euler diagrams showing relative frequency and proportional overlap of autosomal genomic positions containing an SNV. Size of circles is proportional to the number of genomic positions containing a variant allele relative to the boxer reference. As expected, wolves harbor more SNV sites than dogs, with much of the dog polymorphism

overlapping with that in wolves, and jackals harbor a disproportionately large number of private alleles.

S5.3 Ti:Tv

The Ti/Tv ratio has been shown to be an important metric for evaluating the specificity of SNV calls, and the 1000 Genomes and other large-scale sequencing studies show a consistent ratio for whole genome data of ~ 2 [1-3]. Ti:tv for all six canid samples was close to the expected value of ~2 with a mean of 2.16 (range: 2.12 – 2.20; Table S5.3.1)

Table S5.3.1. Ti:Tv ratios for autosomes of 6 canid genomes.

Sample	Transitions	Transversions	Ti:Tv
Basenji	1934897	910072	2.126
Dingo	1995944	941029	2.121
Israeli wolf	2861812	1307555	2.189
Croatian wolf	2685632	1235497	2.174
Chinese wolf	2806363	1288970	2.177
Golden jackal	4422539	2015001	2.195

S5.4 Comparison of NGS Samples to CanMap Reference Samples

S5.4.1 SNP Selection

To assist in validating that our 6 NGS samples were accurate representatives of their respective taxon, we compared the genotypes of our sequenced samples to those from previous SNP genotyping studies, as follows. We chose a set of 10 Boxers, 13 Basenjis, 6 Dingoes, 22 Middle Eastern wolves, 9 European wolves, 10 Chinese wolves, and two golden jackals that were previously genotyped on the Affymetrix Canine version 2 genome-wide SNP mapping array ("CanMap" genotypes; see [4-6] for methods details). We chose a subset of 14,655 SNP positions that overlapped between the two array platforms and were validated using genotype data from our Croatian wolf sample, since it was genotyped on both platforms. This subset of SNPs was further selected so that all SNP positions passed our genome filter GF3, and SNP positions had to be non-missing in the basenji, since that was our lowest coverage sample. This final set of 11,763 SNPs (hereto referred to as the 12K data set) was extracted from our 6 NGS samples and coded as missing if the position did not pass filters (GF3, SF). Thus, for our 6 NGS samples, we constructed a set of genotypes extracted from sequences ("seq-based"), and a set of genotypes from the Illumina array ("Illumina array-based"). For some analyses, we used Illumina array-based genotype data from a second golden jackal individual. All genotypes were converted to PLINK format for subsequent analysis [7].

S5.4.2 Heterozygosity

Individual heterozygosity was calculated using PLINK [7] for the 12K SNP data set of CanMap samples and both the seq-based and Illumina array-based genotypes for the 6 focal samples of this study. The Illumina array-based genotypes for the second Israeli golden jackal were included. The heterozygosity of sequence-based and array-based genotypes of our 6 focal taxa agree very closely (Figure S5.4.1). The heterozygosity levels of the focal taxa are concordant with that of the CanMap samples, though Canmap samples of Basenjis and Middle Eastern wolves appear less heterozygous than the focal samples for this study.

S5.4.3 Principal Components Analysis

Principal components analysis was performed on the 12K data set to visualize the predominant patterns of genetic differentiation, and to verify that our 6 NGS samples clustered within the appropriate species or breed group, in accordance with findings from previous analyses [4,5]. Given that we wanted each individual represented only once within the PCA, we chose the sequencing-based genotypes for the 6 NGS samples, the Illumina array-based genotype data for the second golden jackal, and the CanMap data set. The first 15 PCs were analyzed using the SMARTPCA package within the EIGENSTRAT software [8].

Principal components analysis of individual SNPs separates Boxers, Basenjis, and Dingoes from the wild canids on the first two axes, which together explain over 30% of the variation (Figure S5.4.2). These groupings, and those on subsequent PCs (Figure S5.4.2) confirm previous findings [4,5]. In general, our 6 NGS samples are placed correctly within their respective taxon grouping for the first 15 PC axes. One exception occurs in PC9, which separates our golden jackal from the two CanMap golden jackals. This is most likely due to population- level differentiation within the golden jackals, since the CanMap samples are from Kenya and our sample is from Israel. We also found evidence for differentiation within the CanMap Chinese wolves on PC6, with the NGS samples clustering with one subgroup of the CanMap Chinese samples.

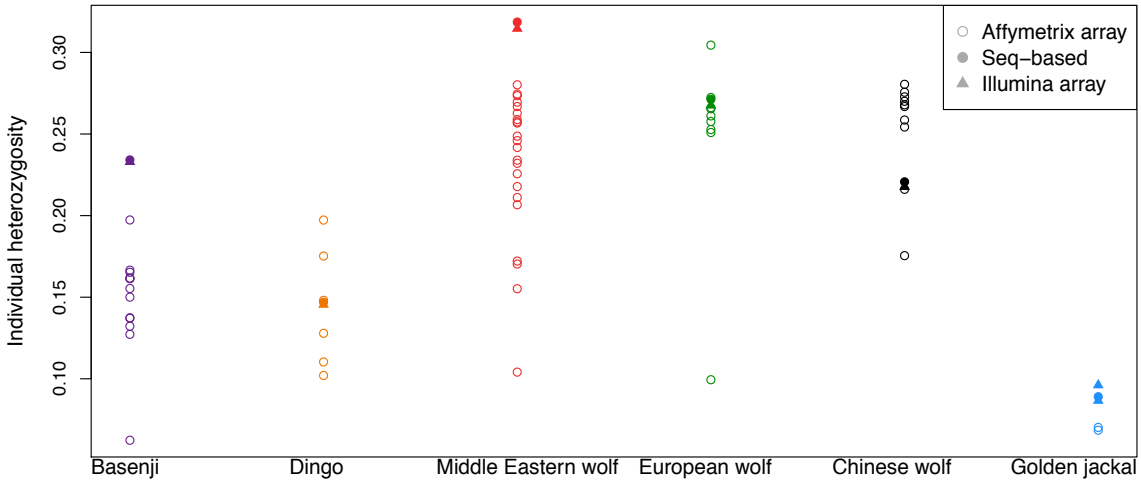


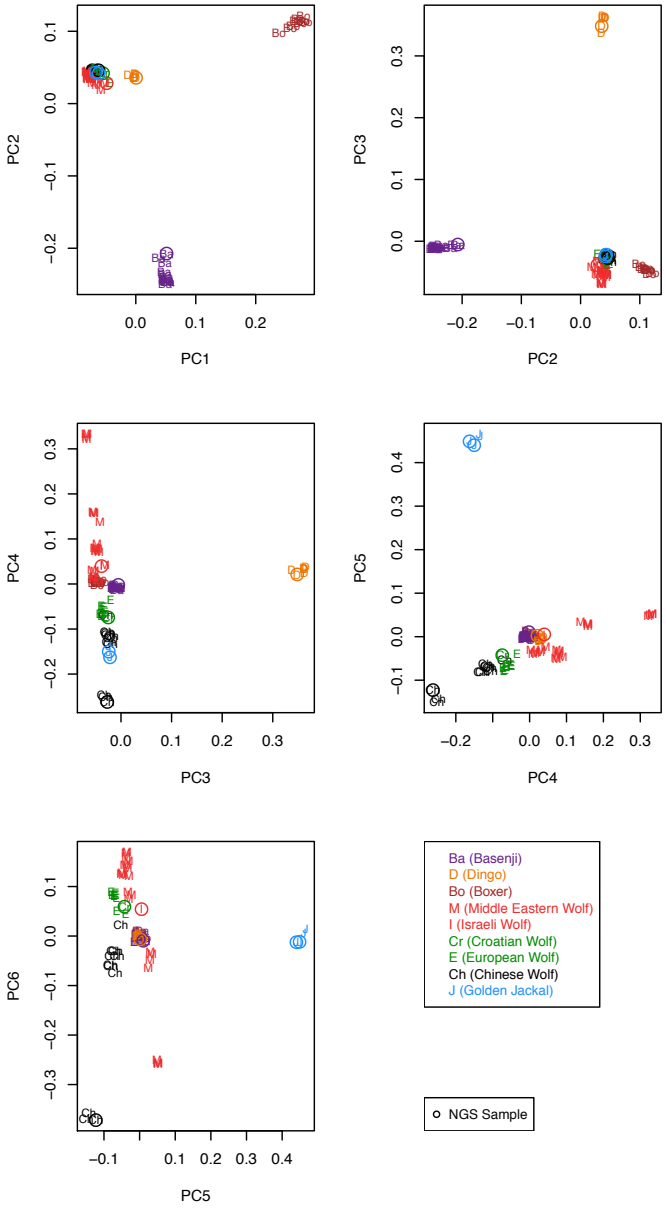
Figure S5.4.1. Heterozygosity estimates by individual for 12K SNP positions overlapping between sequencing- based genotype calls, and those from the Affymetrix and Illumina BeadChip SNP arrays. Sequencing-based genotypes show no apparent bias in heterozygosity. In all 6 taxonomic groups, the same individual was genotyped both via sequencing and the Illumina BeadChip platform, and the corresponding heterozygosity estimates are nearly identical in every instance.

S5.5 Comparison of Pairwise Distances to Previously Published Values

As a test of our data quality we compared pairwise genetic distances computed from our sample to those of a previous study of 12 exons in an overlapping set of lineages [9]. Lindblad-Toh *et al.* [9] defined a set of 12 exons that were used to reconstruct a phylogeny of 31 canid species. Those exons were selected based on the high percentage of bases that were informative in a phylogeny that included humans, dog, mouse and rat; that the mammal phylogeny reconstructed using those 4 mammalian species was consistent with their known phylogeny and that the exons could be amplified in the 31 canid lineages. Using those 12 exonic regions, they computed a matrix of pairwise distances, and reported pairwise distances for three species examined in our study. Those distances were: golden jackal - gray wolf (0.00062), golden jackal – domestic dog (0.00062) and gray wolf – domestic dog (0.00037).

From our sequencing data, we computed genetic distances among our six canid lineages for the same exonic regions (see Text S9.1 and Table S5.5.1). Mean pairwise distances computed from our samples were similar to the ones obtained by Lindblad-Toh *et al.* [9]: golden jackal - gray wolf (0.00077), golden jackal – domestic dog (0.00067) and gray wolf – domestic dog (0.00018). This concordance further supports the quality of our genotype calls.

A


PC2
PC1
PC3
PC2
PC4
PC3
PC5
PC4
PC6
PC5
Ba (Basenji)
D (Dingo)
Bo (Boxer)
M (Middle Eastern Wolf)
I (Israeli Wolf)
Cr (Croatian Wolf)
E (European Wolf)
Ch (Chinese Wolf)
J (Golden Jackal)
NGS Sample

B

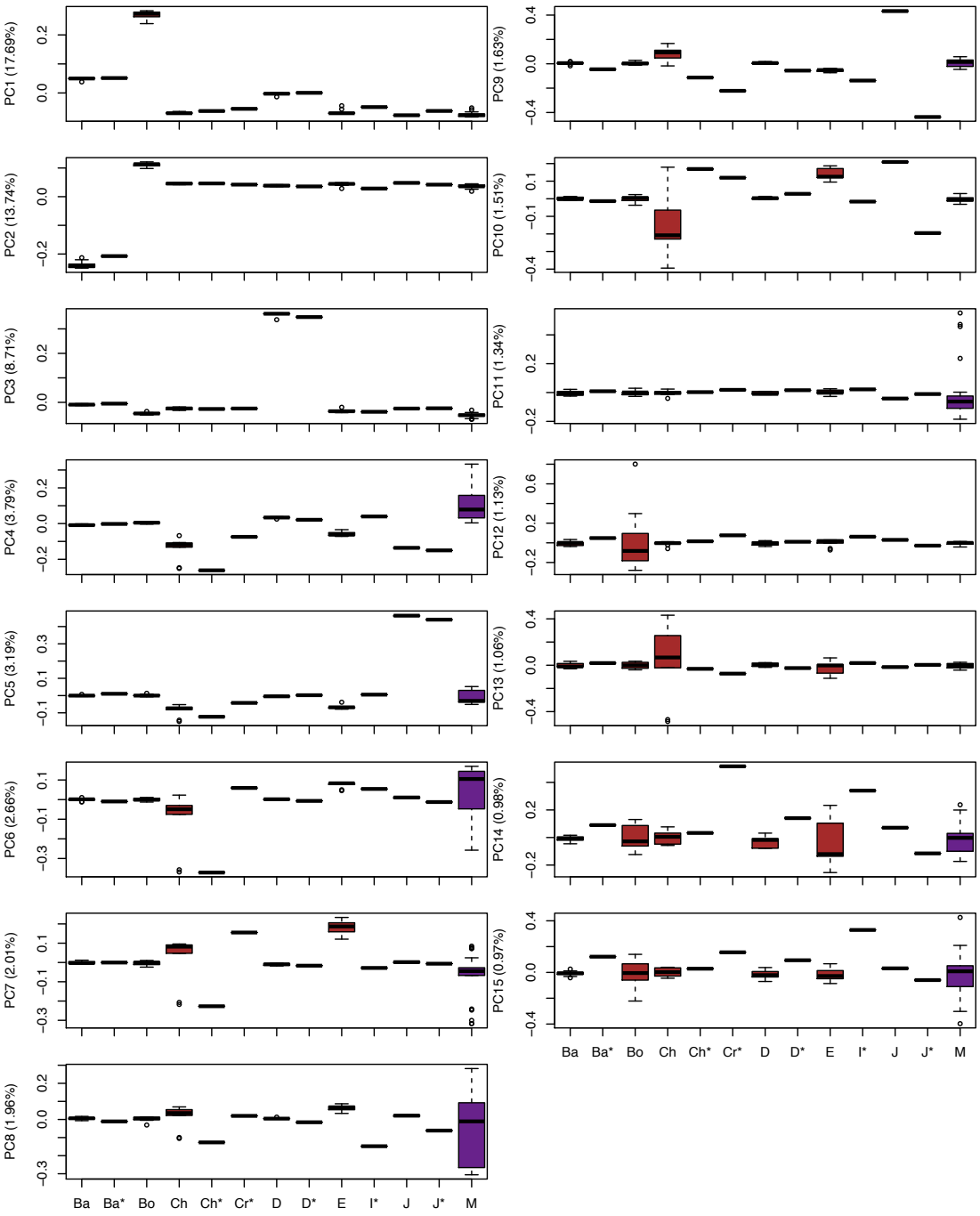

PC1 (17.69%)
PC2 (13.74%)
PC3 (8.71%)
PC4 (3.79%)
PC5 (3.19%)
PC6 (2.66%)
PC7 (2.01%)
PC8 (1.96%)
PC9 (1.63%)
PC10 (1.51%)
PC11 (1.34%)
PC12 (1.13%)
PC13 (1.08%)
PC14 (0.98%)
PC15 (0.97%)
Ba Ba* Bo Ch Ch* Cr* D D* E I* J J* M

Figure S5.4.2. (A) PCA plot of ~ 12,000 overlapping SNPs between CanMap samples genotyped on the Affymetrix canid array, representing taxonomic groupings within which our NGS samples fall, compared to the same genotypes obtained from our sequencing data. (B) Individual boxplots of PCs 1-15, with asterisks indicating genotypes from next-generation sequencing.

Table S5.5.1. Pairwise sequence divergence for exonic regions sequenced by Lindblad-Toh et al. (2005).

	Boxer	Basenji	Dingo	Israeli wolf	Croatian wolf	Chinese wolf	Golden jackal
Boxer							
Basenji	0.00000						
Dingo	0.00011	0.00011					
Israeli wolf	0.00021	0.00021	0.00032				
Croatian wolf	0.00000	0.00000	0.00011	0.00021			
Chinese wolf	0.00021	0.00021	0.00032	0.00042	0.00021		
Golden jackal	0.00063	0.00063	0.00074	0.00084	0.00063	0.00084	

S5.6. Genome-wide Heterozygosity

To quantify genome-wide patterns of heterozygosity, for each genome we calculated the frequency of heterozygous genotype calls across all positions that passed the GF3 and SF filters (see Text S4). As expected, autosomal heterozygosity in dogs was lower in dogs than in wild canids, with levels in the former approximately half of those observed in the latter (Table S6.). Within dogs, the lower heterozygosity in the Dingo likely reflects the historical isolation and strong bottleneck experienced by that lineage. Conversely, the higher heterozygosity in the Basenji is due, in part, to admixture in wild canids (see main text and Text S8.4, S9). The low mean heterozygosity observed in the Chinese wolf relative to other wild canids was produced, in part, by large apparent runs of homozygosity revealed by analyses of 5MB-wide sliding windows (Figure S5.5.1). We formally identify runs of homozygosity (Figure S5.5.1) with the program PLINK v1.07 [8] using the following command line:

```
 plink --tfile <input tfile> --homozyg --homozyg-snp 200 \
--homozyg-kb 10000  --homozyg-window-missing 30 \
--homozyg-window-het 10 --dog
```

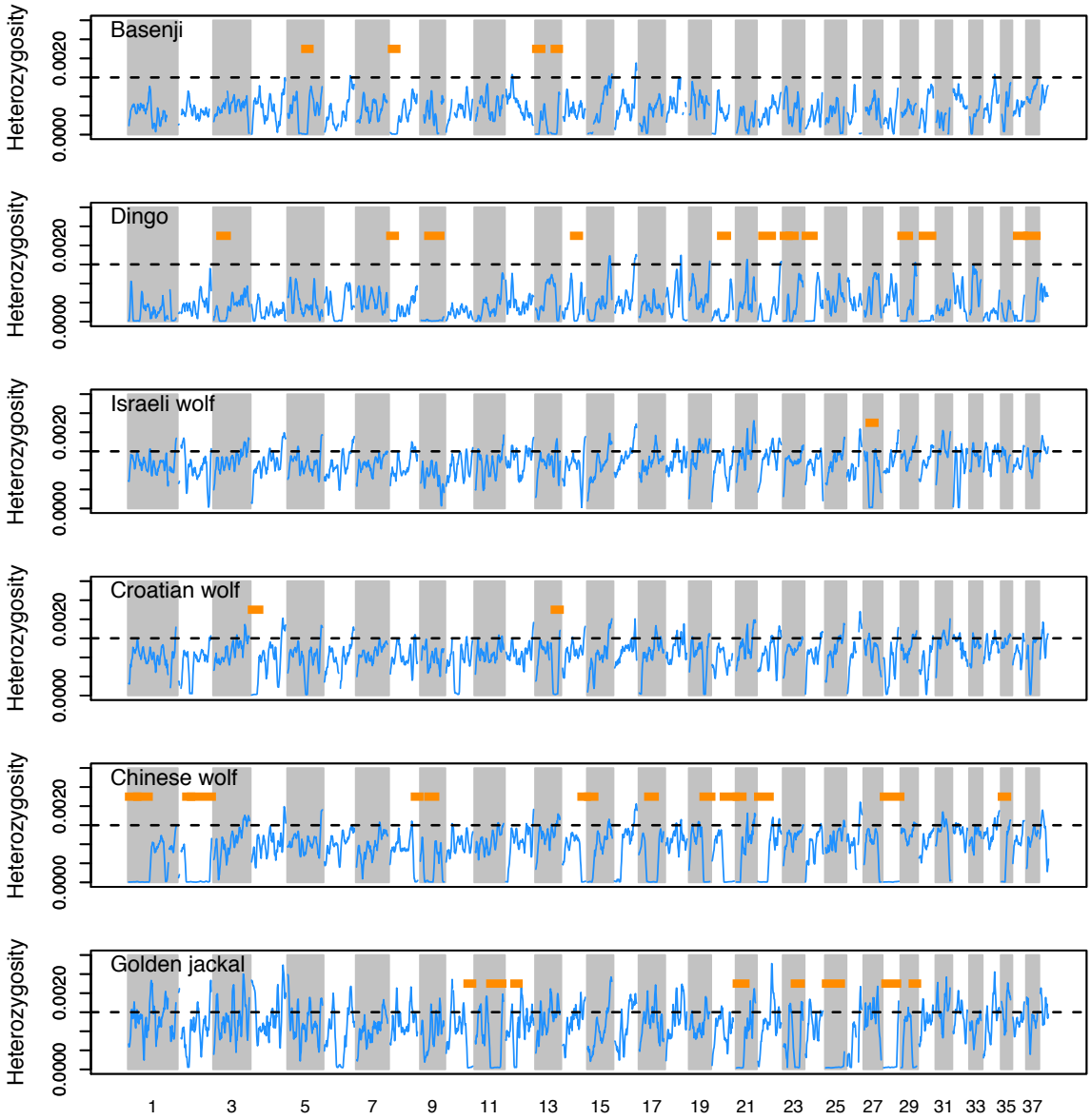


Figure S5.6.1. Genome-wide patterns of mean heterozygosity per sample in 5MB windows, sliding 1MB per step. A minimum of 2MB of genotypes passing GF3 and SF filters were required in order for a window to be displayed. Orange bars indicate runs of homozygosity detected using PLINK.

References

1. Ebersberger I, Metzler D, Schwarz C, Paabo S (2002) Genomewide comparison of DNA sequences between humans and chimpanzees. Am J Hum Genet 70: 1490-1497.

2. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491-498.

3. Altshuler D, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061- 1073.

4. vonHoldt BM, Pollinger JP, Lohmueller KE, Han EJ, Parker HG, et al. (2010) Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. Nature 464: 898-902.

5. vonHoldt BM, Pollinger JP, Earl DA, Knowles JC, Boyko AR, et al. (2011) A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. Genome Res 21: 1294-1305.

6. Boyko AR, Quignon P, Li L, Schoenebeck JJ, Degenhardt JD, et al. (2010) A Simple Genetic Architecture Underlies Morphological Variation in Dogs. Plos Biology 8: e1000451.

7. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575.

8. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904-909.

9. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438: 803-819.

S6 Structural Variant Calling

Belen Lorente-Galdos[1], Farhad Hormozdiari[2], Can Alkan[3], Tomas Marques-Bonet[1,4]

[1]Institut de Biologia Evolutiva (UPF-CSIC)

Departament de Ciències Experimental i de la Salut Parc de Recerca Biomèdica de Barcelona Dr. Aiguader 88, 08003 Barcelona

[2]University of California, Los Angeles

Department of Computer Science 4732 Boelter Hall Los Angeles, CA 90095

[3]Bilkent University

Department of Computer Engineering Ankara, 06800 Turkey

[4]ICREA. Institut Catala de Recerca i Estudis Avançats.

Catalonia, Spain

S6.1 Genome-wide Structural Variant Detection

We used whole-genome shotgun paired-end sequence data generated with both Illumina and Applied Biosystems SOLiD platforms from the genomes of six canid samples (including a additional Basenji only sequenced to low coverage on the Illumina platform, but excluding the Chinese wolf), to estimate the fraction of the genome with segmental duplications. Our goal was to determine potentially duplicated regions to filter out for the final SNP call set.

We identified the segmental duplication (SD) content in these genomes using the Whole-genome Shotgun Sequence Detection (WSSD) approach [1]. This strategy is based on determining regions with a significant excess of depth of coverage. Briefly, WGS reads are allowed to map to multiple locations to a reference genome, and therefore we expect that paralogous copies map into all locations. Highly identical duplicated genomic regions would be detected with an excess of depth of coverage. In our case, we used the dog assembly (canFam2) downloaded from the UCSC Genome Browser. Repeats detected by RepeatMasker and simple tandem repeats with period smaller than 12 detected by the Tandem Repeat Finder were pre- masked. We aligned the Illumina reads allowing 94% of sequence identity using mrFAST v2.0.0.5 [2] and SOLID reads with

drFAST v0.0.0.3 [3].

We calculated the absolute copy numbers of non-overlapping windows of 1 kb of unmasked sequence using mrCaNaVaR version 0.31 (http://mrcanavar.sourceforge.net/). We identified SDs as regions with at least 5 consecutive windows with a copy number higher than 2.5. We detected between 1,379 and 1,413 SD segments larger than 10 kb in the five genomes we analyzed. These regions comprise 52.77 to 55.01 Mb in total that correspond to 2.09% to 2.17% of the reference assembly (Table S6.1.1).

The results are highly similar to the results we obtained using the SOLiD data with the same analysis protocols described above (Table S6.1.2). However, these results are likely to be conservative compared to a previous study [4] where 4.11% of the dog

genome was reported as being duplicated with the same method. In that study, Nicholas et al. [4] used Sanger capillary reads from the same dog that was also used to build the canFam2 reference genome. This difference is likely due to different treatment of repeat sequences in Sanger vs. next-generation sequencing datasets.

To help reduce potential false negatives of the conservative approach outlined above, we applied an alternative strategy for SD calling that is highly similar to the one used for Sanger reads. We identified SDs as regions having a higher read depth than the mean coverage plus 4 standard deviation in at least 6 out of 7 overlapping windows of 5 kb of unmasked and non- gapped sequence. We predicted between 7,456 and 8,202 regions as SDs longer than 10 kb representing between 4.93% to 5.63% of the reference assembly (Table S6.3). We also added the WSSD regions from the reference genome to this dataset [4] and the final list was used to exclude paralogous regions in the SNP calling (Figure S6.1.1). We note that the conservative strategy may have higher rate of false negatives, while this alternative method potentially has a higher false positive rate.

Table S6.1.1. Segmental duplications detected as regions with at least 5 consecutive non-overlapping windows with a copy number higher than 2.5 from Illumina reads.

		>10kb		>20kb	
	Sample ID	**# Intervals**	**# bps**	**# Intervals**	**# bps**
Basenji 1[a]	RKW 13764	1,402	53,445,975	784	44,666,307
Basenji 2	1756	1,409	53,142,107	761	43,921,743
Croatian wolf	RKW 3919	1,413	54,845,287	817	46,428,462
Dingo	RKW13760	1,386	53,042,846	776	44,439,961
Golden Jackal	RKW 1332	1,379	52,769,259	774	44,155,006
Israeli wolf	RKW13759	1,368	55,014,821	796	46,875,363

[a] Basenji used for all other analyses throughout the paper.

Table S6.1.2. Segmental duplications detected as regions with at least 5 consecutive non-overlapping windows with a copy number higher than 2.5 from SOLiD reads. The overlapping bps with the predicted SDs from the Illumina dataset are also shown.

	>10kb			>20kb		
	# Intervals	# bps	Intersection with Illumina data	# Intervals	# bps	Intersection with Illumina data
Basenji	1453	54,580,095	45,605,293[a]	810	45,406,265	38,384,115[a]
Croatian wolf	1276	51,410,183	49,612,023	753	43,912,580	42,457,271
Dingo	1422	54,604,031	48,165,059	808	41,228,862	41,228,862
Golden Jackal	1234	49,330,533	47,072,480	740	42,241,860	40,316,493
Israeli wolf	1289	52,032,514	49,980,206	743	44,192,703	42,796,663

[a] For Basenji, we took the intersection of the two Illumina lanes from the two individuals.

Table S6.1.3. Segmental duplications detected with at least 6 out of 7 5kb overlapping windows showing a read depth higher than the 4 standard deviations above the average, detected using Illumina reads.

		>10kb		>20kb	
Sample	**Sample ID**	**# Intervals**	**# bps**	**# Intervals**	**# bps**
Basenji 1	RKW 13764	7,597	126,674,600	5,506	99,205,302
Basenji 2	1756	8,202	142,461,157	5,966	112,954,332
Croatian wolf	RKW 3919	7,632	128,344,367	5,551	100,981,951
Dingo	RKW13760	8,043	140,574,072	5,798	110,892,809
Golden Jackal	RKW 1332	7,456	124,758,677	5,397	97,652,991
Israeli wolf	RKW13759	7,724	129,350,771	5,545	100,780,739

[a] Basenji used for all other analyses throughout the paper.

Figure S6.1.1. SD distribution on the dog genome (CanFam2). Each horizontal line refers to a chromosome in the dog assembly. Tasha refers to the duplications detected in the reference [4].

S6.2 Copy Number Variation at the Amylase (*AMY2B*) Locus The amylase activity, which cleaves the starch into maltose, has been affected by gene duplication events in the recent history of both humans and dogs [5,6]. In dogs the *AMY2B* gene, that encodes the alpha-2B-amylase enzyme, is present with a high variety of copy number states while in wolves has always being found as single copy. To date, this gene model in dogs has been predicted by Ensembl (http://www.ensembl.org/) and is localized in three regions on the Unknown chromosome of CanFam2. Moreover, there is a partial or unresolved copy of this gene on chromosome 6 detected using BLAT (see Table S6.2.1, Figure S6.2.1).

Table S6.2.1. Location of *AMY2B* in CanFam2.

Chromosome	Start	End	Length	Strand	Gene ID*
ChrUn	4,462,782	74,468,964	6,183	+	ENSCAFG00000032684
ChrUn	62,479,904	62,496,097	16,194	+	ENSCAFG00000030588
ChrUn	6,712,667	46,719,782	7,116	+	ENSCAFG00000031239
Chr6	50,008,123	50,014,414	6,292	-	

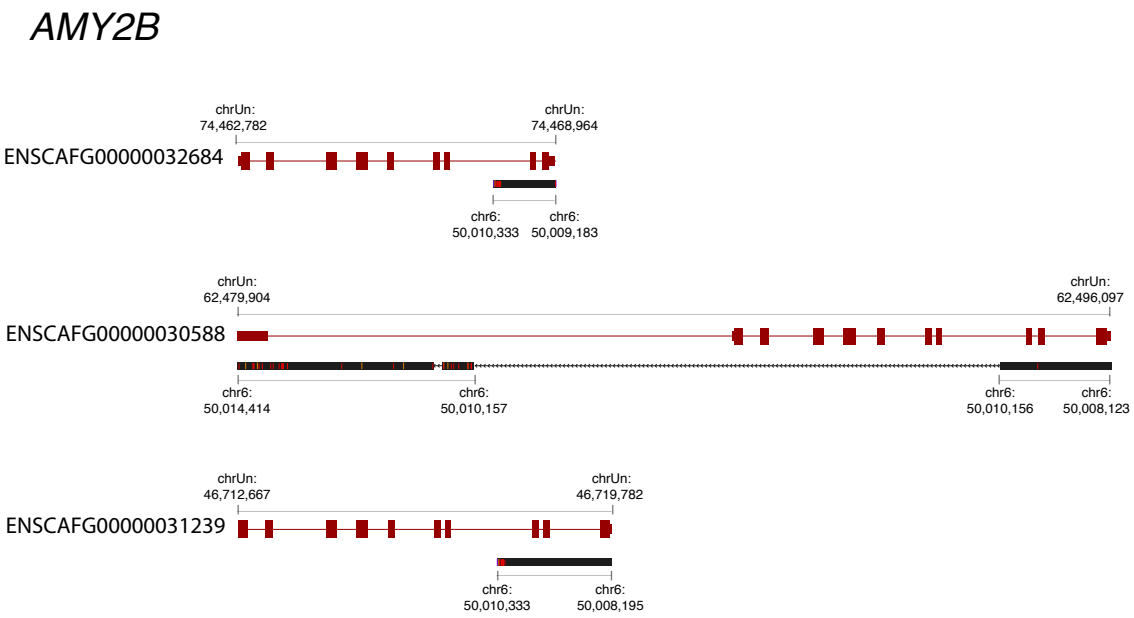


Figure S6.2.1 Region of chromosome 6 mapped to *AMY2B* genes in CanFam2. A duplication with a deleted fragment or a bad representation of this region in the assembly might explain the data.

To determine the duplication status of the amylase gene in our samples we calculated the read depth on contiguous 1kb windows of non-repetitive sequence. The number of copies of any window is estimated by dividing between the average coverage in the genome. In such a way, the expected value for diploid single-copy regions would be around 2. For *AMY2B* this is the case in all our samples except for Basenji, for which the

copy number is higher than 2. The average of the windows containing the gene ENSCAFG00000032684 and ENSCAFG00000031239) in Basenji is around 9 copies. In the case of the predicted gene ENSCAFG00000030588, the first exon, which is non-coding, might be single-copy, as can be inferred from its copy number. This fragment is also shown in the region of chromosome 6 where the gene is partially represented (Figures S6.2.2 and S6.2.3, Table S6.2.2).

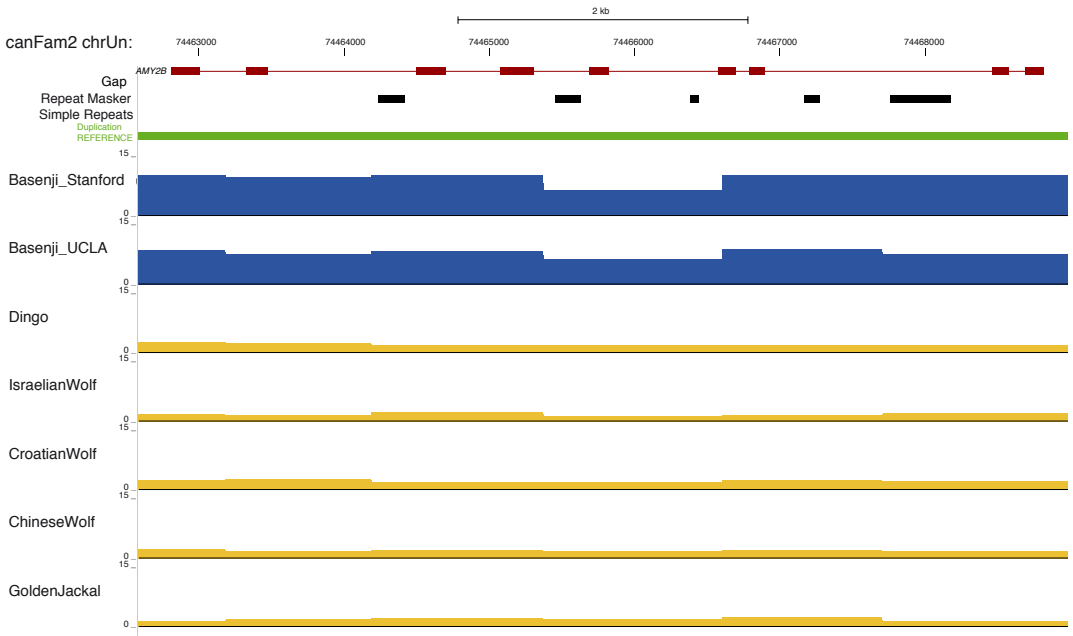


Figure S6.2.2 Average copy number on 1kb windows in the region of ENSCAFG00000032684 (in red) in chrUn. In yellow, expected copy number of single copy regions. In blue, higher copy numbers. In green we represented a duplicated region according to the dog reference genome assembly (Tasha) [4]. Repeats and gaps of the region are also shown.

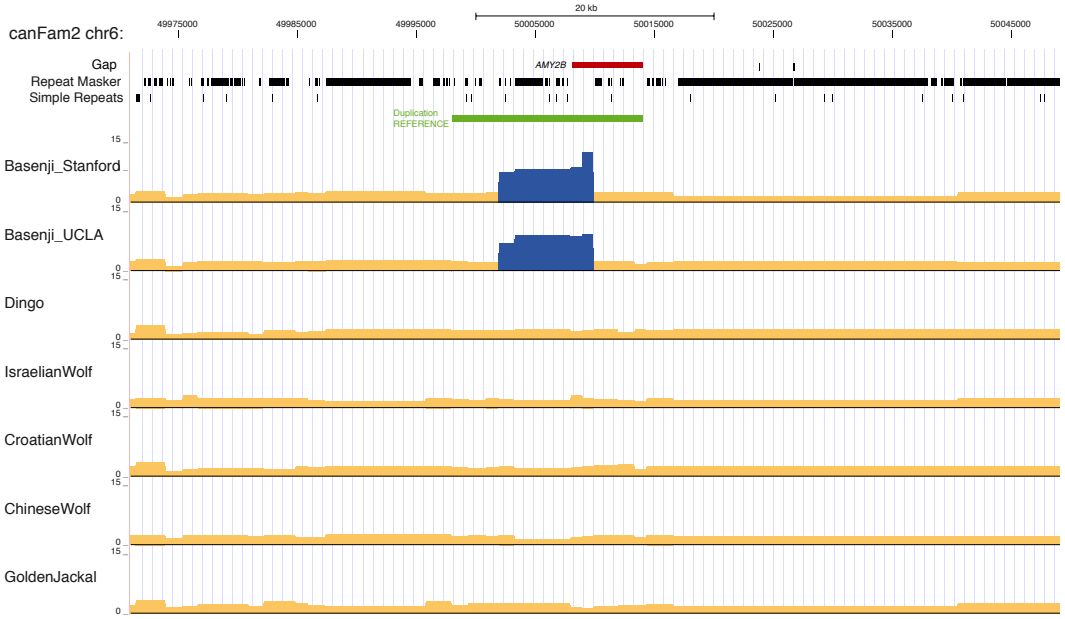


Figure S6.2.3. Average copy number on 1kb windows in the region of chromosome 6 where *AMY2B* is partially represented. The putative location of the gene was determined by aligning into this region the sequence of the predicted genes, and the maximum region (that corresponds to ENSCAFG00000030588) is shown here. In yellow, expected copy number of single copy regions. In blue, regions with higher copy number.

Table S6.2.2 Copy number on windows containing *AMY2B* for Basenji samples. Windows showing evidence for gene duplications are highlighted in green.

	Windows			Copy Number	
	Chr	Start	End	Basenji_Stanford	Basenji_UCLA
ENSCAFG00000032684 (ChrUn:74,462,782-74,468,964)	chrUn	74,462,186	74,463,186	10.17	8.51
	chrUn	74,463,186	74,464,186	9.48	7.63
	chrUn	74,464,186	74,465,370	9.81	8.29
	chrUn	74,465,370	74,466,604	6.41	6.19
	chrUn	74,466,604	74,467,705	10.02	8.55
	chrUn	74,467,705	74,469,120	9.90	7.50
ENSCAFG00000030588 (ChrUn:62,479,904-62,496,097)	chrUn	62,473,212	62,480,658	6.42	4.64
	chrUn	62,480,658	62,481,658	2.52	2.21
	chrUn	62,481,658	62,482,870	1.86	1.71
	chrUn	62,482,870	62,485,453	1.92	2.02
	chrUn	62,485,453	62,488,822	6.54	5.58
	chrUn	62,488,822	62,489,822	9.70	8.51
	chrUn	62,489,822	62,490,822	9.44	6.89
	chrUn	62,490,822	62,491,999	9.25	8.06
	chrUn	62,491,999	62,493,056	7.59	6.68
	chrUn	62,493,056	62,494,572	9.45	8.72
	chrUn	62,494,572	62,495,572	12.42	9.26
	chrUn	62,495,572	62,507,548	6.37	6.47
ENSCAFG00000031239 (ChrUn:46,712,667-46,719,782)	chrUn	46,710,448	46,712,776	9.52	7.74
	chrUn	46,712,776	46,713,776	9.94	8.50
	chrUn	46,713,776	46,714,961	9.60	7.55
	chrUn	46,714,961	46,716,138	7.23	6.93
	chrUn	46,716,138	46,717,296	8.88	7.64
	chrUn	46,717,296	46,718,711	10.85	8.81
	chrUn	46,718,711	46,719,711	10.81	8.86
Chr6:50008123-50014,414	chr6	50,007,907	50,008,907	8.78	8.39
	chr6	50,008,907	50,009,907	12.51	9.13
	chr6	50,009,907	50,011,943	2.42	2.10
	chr6	50,011,943	50,013,300	2.36	2.25
	chr6	50,013,300	50,014,300	2.51	1.65
	chr6	50,014,300	50,016,584	2.48	1.95

S6.3 Validation of copy number of *AMY2B* by real-time quantitative PCR (qPCR) We explore the variation in *AMY2B* copies using qPCR across additional breed dogs (n=52), dingoes (n=6) and a globally distributed panel of wolves (n=40) (Table S13). This new data improve specially the variability presented in wolves, with the

analysis of samples from 9 wolf populations, 5 of them not previously explored. Also this data allow us to validate the copy number of *AMY2B* estimated based on whole genome sequencing (Table S6.3.1). Estimation of copy number was performed using the Multiplex TaqMan assays previously described by Axelsson et al. [5].The duplex reaction contained a reference assay designed to amplify *C7orf28B* that is known to exist in two copies in a canid genome (900 nM of forward and reverse primers, 250 nM VIC and TAMRA labeled probe, Applied Biosystems), and the *AMY2B* as a target gene (300 nM of forward and reverse primers, 250 nM FAM labeled MGB probe, Applied Biosystems) in genomic DNA. For each sample we performed three replicates.

Table S6.3.2. Amylase copy number in 10 dogs estimated by qPCR and genome sequencing.

Sample	Copy Number from qPCR	Copy Number from Genome Sequencing
Beagle	6	7
Bulldog	14	15
Chihuahua	10	10
Flat-coated retriever	12	10
Great dane	16	16
Mastiff	8	12
Pekingese	14	14
Saluki	23	29
Scottish terrier	8	9
Siberian husky	3	3

References

1. Bailey JA, Gu ZP, Clark RA, Reinert K, Samonte RV, et al. (2002) Recent segmental duplications in the human genome. Science 297: 1003-1007.

2. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, et al. (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet 41: 1061-1067.

3. Hormozdiari F, Hach F, Sahinalp SC, Eichler EE, Alkan C (2011) Sensitive and fast mapping of di-base encoded reads. Bioinformatics 27: 1915-1921.

4. Nicholas TJ, Cheng Z, Ventura M, Mealey K, Eichler EE, et al. (2009) The genomic architecture of segmental duplications and associated copy number variants in dogs. Genome Res 19: 491-499.

5. Axelsson E, Ratnakumar A, Arendt M-J, Maqbool K, Webster MT, et al. (2013) The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature 495:360-364.

6. Meisler MH, Ting CN (1993) The Remarkable Evolutionary History of the Human Amylase Genes. Crit Rev Oral Biol Med 4: 503-509.

S7 Gene Annotations

Pedro Silva[1], Marco Galaverni[2], Rena M. Schweizer[3], Adam H. Freedman[3], Robert K. Wayne[3], John Novembre[3]

[1] University of Porto

CIBIO-UP - Research Center in Biodiversity and Genetic Resources Porto, Portugal

[2] Istituto Superiore per la Protezione e la Ricerca Ambientale

Laboratorio di Genetica Ozzano dell'Emilia, Italy

[3] University of California, Los Angeles

Department of Ecology and Evolutionary Biology Los Angeles, California, United States of America

In order to construct a set of neutral regions for use in demographic analyses, we generated a set of annotations for regions likely evolving in a non-neutral fashion. These consisted of genic and other regions showing a high degree of conservation or that might otherwise play a functional role, and were identified as described below.

S7.1 Identification of Genes

In order to build a comprehensive set of annotated genes in the domestic dog, we compiled the available information from three different sources: the refGene file from the UCSC genome browser database [1] (downloaded from ftp://hgdownload.cse.ucsc.edu/goldenPath/canFam2/database/ on Aug 15, 2011); Ensembl [2] (all protein coding genes from Ensembl Release 63 downloaded Jul 26, 2011 via the BioMart MartView tool: http://www.ensembl.org/biomart/martview), and SeqGene files from the NCBI database (downloaded Mar 02, 2011 from ftp://ftp.ncbi.nih.gov/genomes/Canis_familiaris/mapview/). All downloaded information pertains to the May 2005 assembly of the dog genome (canFam2) [3]. The retrieved information included annotated gene names and symbols, genomic coordinates of coding exons and untranslated regions (UTRs) (including annotations for alternatively spliced transcripts) of both confirmed and predicted genes.

UCSC refGene contained 1,168 entries (transcripts) corresponding to 1,131 unique gene symbols, data from Ensembl yielded 30,914 transcripts from 24,660 different genes, and NCBI's seq_gene contained 33,636 transcripts from 19,758 genes. As currently available

gene annotations are all provided with reference to genomic coordinates in canFam2, as with the Canine BeadChip genotypes (see Text S4.1), we used the command-line version of the UCSC *liftover* tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver) to convert annotations to canFam 3.0 (see SI Text S3.1.1 for details concerning the reference) coordinates. 36 transcripts (27 genes) from RefGene, 736 (600 genes) from Ensembl and 1,368 (977 genes) from NCBI failed to be converted.

Many of the entries in NCBI's SeqGene file had provisional LOC codes as the only available information, so an effort was made to obtain gene symbols and descriptions for those loci. For this, we used NCBI's BatchEntrez tool (http://www.ncbi.nlm.nih.gov/sites/batchentrez) to query the 'Gene' database. Many of the records were already discontinued, so those entries were eliminated. The final SeqGene annotation from NCBI comprised 32,200 transcripts from 18,601 genes.

From the initial concatenated set of annotations from the three sources, we created a reduced set in which we merged entries that appeared to be duplicates. In order for any two annotated genes to be considered duplicates, they had to have overlapping coordinates and be transcribed on the same strand. In addition, they had to either possess similar symbols, similar gene descriptions, a symbol of one matching the transcript IDs of the other, or share ≥90% of their exonic sequence. While the last of these criteria is somewhat arbitrary, we chose it based upon the observation that only a very small percentage of annotations known to be unique displayed such a high degree of exonic overlap. The majority of annotation merges depending upon this criterion were sparsely annotated, typically falling into the "unknown gene" category. As an understanding of the functional role for many predicted genes in the dog is incomplete, we chose to retain such annotations. The final gene annotation set consisted of 28,805 genic regions with 63,510 associated transcripts.

We distinguished what appeared to be functional transcripts ('CDS OK'), containing properly positioned start and stop codons and a transcript length that is a multiple of 3 bp, from those that were not. Approximately 19% of transcripts from UCSC RefGene, 23% from Ensembl, and 89% from NCBI SeqGene satisfied these conditions. From these, we retained the longest transcript from each unique gene annotation, and used these to build our final transcript annotation set; in those cases where a gene contained more than one transcript with the same length, one was chosen randomly. This final transcript set ('CDS-OK longest transcripts') consisted of 19,910 transcripts. Figure S7.1 shows the provenance of the transcripts that constitute this final set.

The transcripts that did not pass the CDS filters in the boxer genome, probably due for

the most part to improper annotation, were still retained and grouped in an additional annotation dataset ('CDS-fail transcripts' set of 23,079 transcripts).

S7.2 Identification of conserved non-coding regions

Recent research has indicated conserved non-coding elements can play an important role in modulating the regulation of gene expression [4,5]. In particular, such regions conserved across vertebrates, but showing acceleration on the human lineage (HARs), have been implicated in the rapid acquisition of traits unique to humans [4]. To identify conserved elements in dogs, we first identified conserved genomic regions in a set of mammals not including the dog, through examination of a multi-genome alignment of 11 species of the mammalian Euarchontoglires clade, using mouse as reference: mouse, rat, guinea pig, rabbit, human, chimpanzee, orangutan, rhesus macaque, marmoset, bushbaby and tree shrew. The Euarchontoglires (Supraprimates) represent a sister clade to the Laurasiatheria clade that includes carnivores and allows us to identify mammalian conserved regions of the genome without the influence of dog/canid specific changes.

We identified conserved non-coding elements (CNEs) using phastCons scores [6]

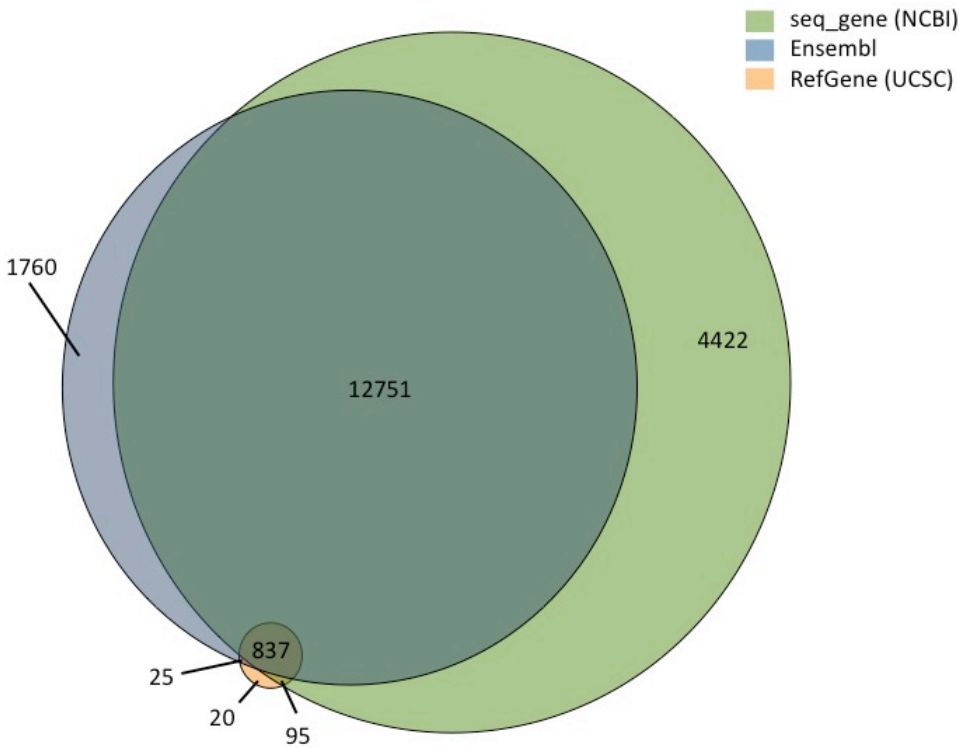


Figure S7.1. Origin of the transcripts in the final CDS OK transcript set. Numbers denote the amount of transcripts found

in each database; intersections represent merged gene entries.

provided for the Euarchontoglires clade available on UCSC for the mouse genome (http://hgdownload.cse.ucsc.edu/goldenPath/mm9/phastCons30way/euarchontoglires/) . Conserved regions of the mouse genome were defined as stretches of consecutive bases with phastCons scores > 0.7 longer than 50 bp. The 50 bp threshold was chosen because this approximates the lower size limit of miRNA genes (www.mirbase.org), and such genes have been previously discovered within HARs [4]. The genomic locations of these regions were then converted to the CanFam 3.0 assembly of the dog genome using the command-line version of the UCSC *liftover* tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver). CNEs were then defined by the intersection of these conserved regions with the non-coding portion of CanFam 3.0.

S7.3 Regulatory Region Variation (UTR, Promoters, Dog-wolf Differences in Dog-Conserved Binding Motifs) Given the possible effects of gene regulation on phenotypic traits that differentiate dogs from their wild ancestors, we classified regulatory regions flanking coding sequences including the 5' and 3' untranslated regions (UTRs) and promoter regions. These regions were defined based on our CDS OK annotation set. UTRs were defined as the regions between the annotated transcription start/end site and the first base of the initiation codon/last base of the stop codon. 5'UTR were defined for 8,427 (~42%) and 3'UTR for 11085 (~56%) of CDS OK transcripts. 6,581 (~33%) of the transcripts had both types of UTRs.

Promoter regions were considered as the 1Kb regions upstream of the transcription start site, considering strand orientation. Putative transcription factor binding sites (TFBS) were searched within the promoter regions using the profiles in the JASPAR PHYLOFACTS database (http://jaspar.cgb.ki.se/) since this database contains count matrices of conserved motifs in human, mouse, rat and dog, originally identified by [7].The motifs were converted to probability weight matrices and used with the motif finding program FIMO [8], part of the MEME package (http://meme.sdsc.edu) to find matching occurrences in the promoter regions of the dog genome. A total number of 866,242 putative binding sites were identified.

References

1. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, et al. (2011) The UCSC Genome Browser database: update 2011. Nucleic Acids Rese 39: D876-D882.

2. Birney E, Andrews TD, Bevan P, Caccamo M, Chen Y, et al. (2004) An overview of ensembl. Genome Res 14: 925-928.

3. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438: 803-819.

4. Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, et al. (2006) An RNA gene expressed during cortical development evolved rapidly in humans. Nature 443: 167-172. 5. Lowe CB, Kellis M, Siepel A, Raney BJ, Clamp M, et al. (2011) Three Periods of Regulatory

Innovation During Vertebrate Evolution. Science 333: 1019-1024. 6. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou MM, et al. (2005) Evolutionarily

conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15:

1034-1050. 7. Xie XH, Lu J, Kulbokas EJ, Golub TR, Mootha V, et al. (2005) Systematic discovery of

regulatory motifs in human promoters and 3 ' UTRs by comparison of several mammals.

Nature 434: 338-345. 8. Grant CE, Bailey TL, Noble WS (2011) FIMO: scanning for occurrences of a given motif.

Bioinformatics 27: 1017-1018.

S8 Demographic Analyses Using Sequence Divergence, ABBA/BABA Tests and PSMC

Diego Ortega Del Vecchyo[1], Zhenxin Fan[2], Adam H. Freedman[1], Rena M. Schweizer[1], Pedro Silva[3], Robert K. Wayne[1], John Novembre[1]

[1]University of California, Los Angeles

Department of Ecology and Evolutionary Biology Los Angeles, California, United States of America

[2]Sichuan University

Sichuan Key Laboratory of Conservation Biology on Endangered Wildlife, College of Life Sciences

Chengdu, People's Republic of China

[3]University of Porto

CIBIO-UP - Research Center in Biodiversity and Genetic Resources Porto, Portugal

S8.1 Distance Matrices and Phylogenetic Tree Reconstruction

S.8.1.1 Distance Metrics

We computed a matrix with the pairwise genetic distances between each of the 6 canid genomes and the reference Boxer sequence using the genetic distance metric from Gronau et al. [1]:

$$d(X,Y) = \frac{1}{L}\sum_{i=1}^{L}[1 - \frac{1}{2}\max\left(\delta_{a_i c_i} + \delta_{b_i d_i}, \delta_{a_i d_i} + \delta_{b_i c_i}\right)]$$

(E9.1)

where X and Y represent the two genomes being compared, L is the total number of sites utilized in the analysis, a_i and b_i are the two allele copies carried by individual X, c_i and d_i are the two allele copies carried by individual Y and δ_{jk} represents the Kronecker delta function (i.e. in this case equals one if allele j is identical to allele k and 0 otherwise). This measure represents a conservative estimate of the expected number of differences per site between individual chromosomes drawn (Gronau et al, 2011, S3.2).

We also computed the average number of nucleotide differences per site among a pair

of randomly drawn alleles from each individual, using the following equation:

$$d(X,Y) = \frac{1}{L}\sum_{i=1}^{L}[1 - \frac{1}{4}\max\left(\delta_{a_i c_i} + \delta_{b_i d_i} + \delta_{a_i d_i} + \delta_{b_i c_i}\right)]$$

(E9.2)

In order to be included in the analysis, sites had to pass the GF2 and SF filters and had no missing genotypes for all of the six samples.

S9.1.2 Results on Genome-wide Pairwise Distances

We took all of the sites across the genome that passed the quality filters defined above to compute a matrix of pairwise distances between all canid genomes using E8.1 and E8.2 (Tables S8.1.1 and S8.1.2, respectively). The distances of all taxa to the golden jackal are very similar (approximately 0.0021) while the distances between dogs and wolves were about a half of that (0.0011). We used the matrix of pairwise distances generated by E8.1 and E8.2 to generate phylogenetic trees using the neighbor joining method as implemented on the program *neighbor* of the phylogenetic package *PHYLIP* [2].

In the neighbor-joining tree generated by using E8.1 (Figure S8.1.1A), all dogs were clustered into a single clade. Wolves also comprised a single clade, separated from other species by a branch of relatively short length. The Dingo was recovered as the outgroup to a clade comprised of Basenji and Boxer. Similarly, the Chinese wolf was inferred as the outgroup to the clade formed by the Israeli and Croatian wolves. Thus, the phylogenetic tree supports the hypothesis that dogs and wolves are reciprocally monophyletic taxa.

The tree created using E8.2 (Figure S8.1.1B) differs from the previous tree in the position of the Chinese Wolf lineage. The Chinese Wolf appears as an outgroup to the clade comprised of the remaining dogs and wolves. However, the bootstrap support is low for both the branch that joins that lineage to the whole wolf-dog clade (54.2%) and the branch ancestral to the clade comprised of the Croatian and Israeli wolves 53.7%).

Table S8.1.1. Genome-wide pairwise sequence divergence, estimated using E8.1 using all the genomic sites that passed the genomic quality filters outlined in S.8.1.1.

	Boxer	Basenji	Dingo	Israeli Wolf	Croatian Wolf	ChineseWolf	Jackal
Boxer							

Basenji	0.00087					
Dingo	0.00094	0.00097				
Israeli wolf	0.00111	0.00105	0.00111			
Croatian wolf	0.00113	0.00110	0.00112	0.00101		
Chinese wolf	0.00114	0.00111	0.00111	0.00106	0.00105	
Golden jackal	0.00211	0.00211	0.00212	0.00209	0.00209	0.00210

Table S9.1.2. Genome-wide pairwise sequence divergence, estimated using E8.2 using all the genomic sites that passed the genomic quality filters outlined in S.8.1.1.

	Boxer	Basenji	Dingo	Israeli Wolf	Croatian Wolf	Chinese Wolf	Jackal
Boxer							
Basenji	0.00087						
Dingo	0.00094	0.00100					
Israeli wolf	0.00111	0.00112	0.00116				
Croatian wolf	0.00113	0.00117	0.00116	0.00115			
Chinese wolf	0.00114	0.00117	0.00115	0.00118	0.00115		
Golden jackal	0.00211	0.00214	0.00214	0.00214	0.00214	0.00214	

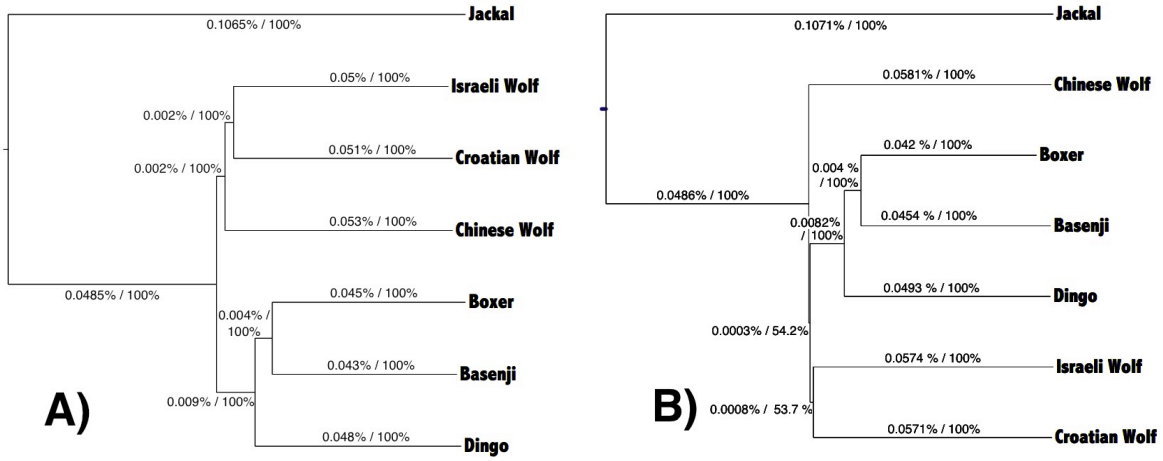


Figure S8.1.1. Neighbor-joining tree of canid samples plus the Boxer reference (CanFam3.0) for all positions passing the GF2 and SF filters and for which there was no missing data for any sample. The distance metrics used were E8.1 and E8.2 for panel A) and B), respectively. For each branch, we report the genetic distance (left side of the slash) and the bootstrap support (right side of the slash). Bootstrap replicates were generated by dividing the genome of each species into windows of 500 kb based on the genomic coordinates of the Boxer reference, and then resampling with replacement from those windows until the bootstrapped genomes for each species contain an equal or greater number of sites called as the true genomes.

S8.2 Population Size Change From Single Genome Sequences

S8.2.1 PSMC: General Approach

We used the methods developed by Li and Durbin [3] to infer the trajectory of population sizes across time for the six canid genome sequences. Briefly, the method uses the distribution of heterozygote sites across the genome and a pairwise sequentially Markovian coalescent (PSMC) model that defines a Hidden Markov Model, where the parameters are the mutation rate, recombination rate and the effective population sizes through time. The parameters are inferred through an Expectation-Maximization algorithm.

The genotypes for each diploid genome sample that passed the GF2 and SF filters were transformed into a sequence of '0', '1' and '.', with one character for each 100bp, and where a '1' was assigned if there were heterozygous sites in the window, 0 if there were none, and a '.' was given if more than 90 positions were missing in the 100 bp window. Passing this data into the PSMC software, we ran 20 iterations of the Expectation-Maximization algorithm [3]. The EM algorithm was run using an upper bound on the time to the most recent common ancestor equal to 10 in a $2N_0$ scale and an initial θ/ρ set to the default value of 5. Following [3], the N_e was inferred across 64 different intervals for each dog genome, where the interval boundaries were set equal to:

$$t_i = 0.1 \exp\left[\frac{1}{n} \log\left(i + 100\right)\right] - 0.1$$

on a $2N_0$ scale ,where *i* takes values from 0 to 64. In a preliminary run we found that the number of recombination events inferred in the most recent time intervals by PSMC falls below 10. In such situations, the authors of PSMC recommend refraining from inferring a population size during such time intervals. Thus, we merged the first 6 intervals such that only a single N_e is inferred across them while the next 58 intervals were allowed to have interval-specific N_e values (in the Chinese wolf, the number of recombination events was higher and thus we continued to use all 64 intervals).

To translate from time units of generations to calendar years, we assume a generational time of 3 years for the wolves and the golden jackal. For the Dingo and the basenji, we used a generational time of 2 years from the present until the N_e interval that reached 10,000 years ago and for all N_e intervals further into the past, we used a generational time of 3 years. We found this scaling improved the concordance of the trajectories during the ancestral period where we expect them to be identical across lineages and is motivated by the known shorter generation time in domestic dogs. Following Lindblad-Toh *et al.* [4], the mutation rate assumed was 1.0×10^{-8} per generation.

The full results including the golden jackal are shown here (Figure S8.2.1). The golden

jackal shows an apparent large increase in effective populations size around 80,000 years ago. We address interpretations of this signal in more detail in the results of our validation study (see below).

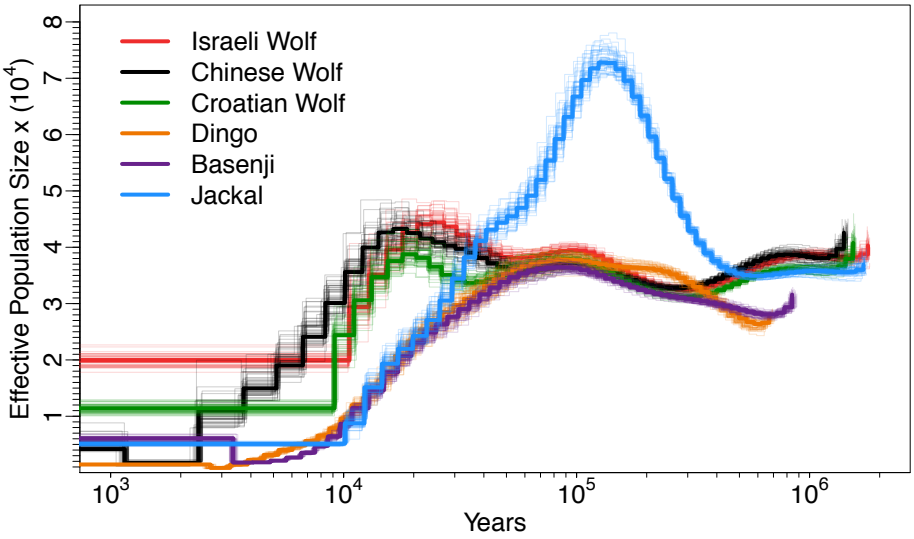


Figure S8.2.1. N_e trajectories of 6 canid lineages reconstructed using the PSMC method of Li and Durbin [3]. Dark and light lines indicate whole genome based estimates and bootstrap estimates, respectively.

S8.2.2 Validation

We assessed the confidence in our PSMC findings in three ways. First, to assess the certainty in the inferred N_e trajectories, we ran the PSMC method using the same settings for the initial estimations, assessing the variance in those estimates from 100 bootstrap replicates for each genome. To sample a bootstrap replicate, we divided the genome into segments of 5Mb, sampled with replacement from those segments until we obtained a sequence with approximately the same length as the original genome as defined by using the "-b" option in the PSMC software, and re-ran the EM-based N_e estimation procedure. This analysis revealed a low variability among the N_e traces, comparable to what has been recovered in the analysis of human genome sequences (Figure S8.2.1) [3].

Second, we tested the sensitivity of the methods to long runs of homozygosity (RoH), as the Chinese wolf sample evidenced several runs (see Text S5.6). To test if long runs of homozygosity could bias the inference of Ne trajectories, we identified runs of homozygosity with the program PLINK [5] (see Text S5.6, Figure S5.6.1). As can be seen in Figure S8.2.2, the estimated trajectories are not affected by the removal of the RoH regions. This implies that the degree of inbreeding in the Chinese wolf is not large enough to bias the inference of ancestral demographic events estimated by the PSMC method.

Third, to investigate the sensitivity of PSMC to our choice of minimum acceptable genotype quality (GQ≥20), we ran the PSMC analysis including the genotypes that passed the GF2 and SF filters, but relaxing the GQ component of SF such that we included sites with GQ≥10 (as a contrast, Figure S8.2.1 and Figure 3B use the genotypes that passed the GF2 and SF1 filters and had a GQ >= 20). Using this more liberal GQ threshold, values of N_e are lower by approximately 1,000 along the trajectory of all canids (Figure S8.2.3), however the N_e trajectories remain largely concordant. The effect is particularly strong in the golden jackal between 50,000 – 300,000 years ago, where using a lower GQ threshold reduces the estimates of N_e by 2,000. The difference between the dog and wolf N_e at earlier times (5,000-70,000 years) is more noticeable when using a higher GQ threshold. The reductions in N_e across the PSMC traces are consistent with expectations with respect to how confidence in genotype quality scales differently for homozygous versus heterozygous genotype calls. Homozygous sites can be called confidently with less data that is of lower quality. Conversely, heterozygous calls will require more and higher quality data, such that genotype qualities at those sites will be higher. As a result, lowering the GQ threshold leads to the inclusion of disproportionately more homozygous genotypes than low quality heterozygous ones, reducing the observed heterozygosity within defined intervals, and as a result, the inferred N_e. Overall, although changes in GQ filtering does influence the estimates of the N_e trajectories, the magnitude of the changes are not large, and more importantly, the major patterns in the inferred trajectories are preserved.

Fourth, we simulated genome sequences arising from the demographic history inferred from the model analyzed by *G-PhoCS* which assumes that wolves and dogs are reciprocally monophyletic taxa (see Table S9.2 and Figure S9.1) to determine if we could accurately reconstruct changes in N_e conditional on such a history. Specifically, for each species we simulated one hundred regions of 30Mb apiece using the program MaCS [6]. We conducted these simulations under three different scenarios, varying the levels of gene flow between lineages. We used parameter values from the main results obtained with *G-PhoCS* (see Table S9.2). The scenarios tested used:

1) The full model inferred from *G-PhoCS* (Command Line 1, see command-line parameter listings below). 2) Our model inferred with *G-PhoCS* but with no gene flow between any species at any time (Command Line 2).

3) The model inferred by *G-PhoCS* but with only one form of gene flow, from golden jackal to the ancestor of dogs and wolves (Command Line 3). 4) The model inferred by *G-PhoCS* but with only one form of gene flow, from the ancestor of dogs and wolves to the golden jackal (Command Line 4).

5) The model inferred by *G-PhoCS* but only with gene flow from the Israeli wolf to the golden jackal (Command Line 5).

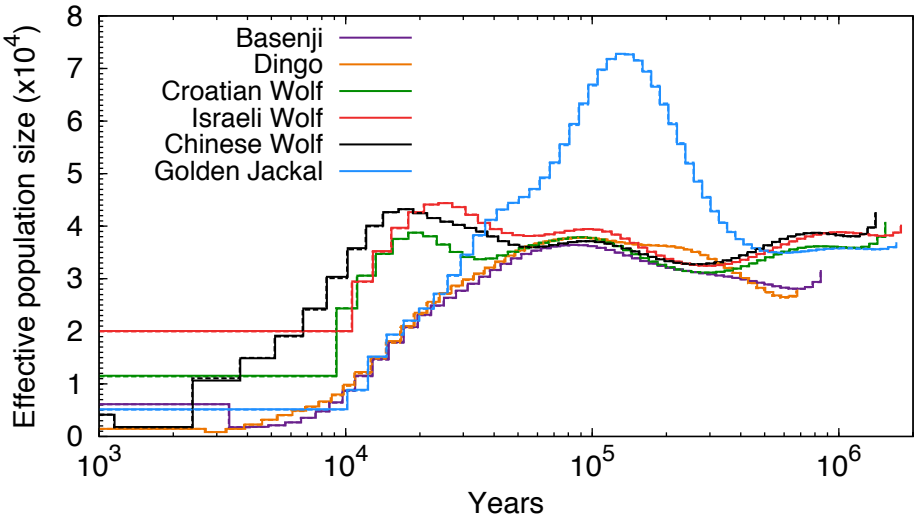


Figure S8.2.2. N_e trajectories of 6 canid lineages reconstructed using the PSMC method of Li and Durbin [3], using all the genomic information that passed our quality filters (dashed lines) and excluding 43 regions with runs of homozygosity (solid lines).

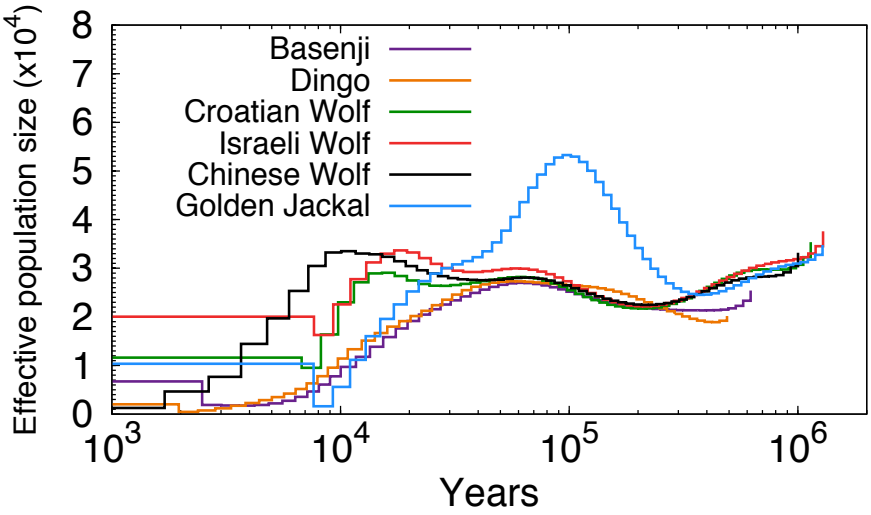


Figure S8.2.3. N_e trajectories of 6 canid lineages reconstructed using the PSMC method of Li and Durbin [3] using the sites that had a GQ >= 10 and passed the SF and GF2 filters.

There are 7 different genomes being simulated in the command lines for each scenario. They are a haploid genome of the Boxer and diploid genomes for the Basenji, Dingo, Israeli wolf, Croatian wolf, Chinese wolf and Golden Jackal, respectively. Only the diploid genomes were used in this analysis. The output of MaCS was processed using perl scripts, so that each of the 30Mb regions was transformed into a binary sequence of '1' and '0', where each character was determined by the presence or absence of a

heterozygote site in contiguous windows of 100bp. Then, for each lineage we used the 100 transformed binary sequences of 30Mb to run the PSMC method using the following command line:

```
./psmc -N20 -t10 -r5 -p "1*6+58*1" -o <Output
file> <Input
file>.
```

The recombination rate in all scenarios was assumed to be equal to 0.92 cM/Mb, a value that is equal to the mean recombination rate estimated in the dog genome in a linkage map generated using microsatellites [7]. In these simulations, we set the generational time to 3 years and mutation rate to 1×10^{-8} per bp per generation for all species.

We compared the N_e trajectories specified in the simulations with the estimations done by the PSMC method for each canid species. Scenarios 2 (Figure S8.2.4) and 3 (Figure S8.2.5) have remarkably similar and accurate trajectories inferred using the PSMC method for all species of canids. In scenarios 4 (Figure S8.2.6), 5 (Figure S8.2.7) and 1 (Figure S8.2.8), the N_e trajectories are also accurate for all species of canids but the golden jackal, where the estimate of N_e is inflated in the interval from 10,000 - 300,000 years ago, with a distinctive sharp peak between 100,000 and 300,000 years ago.

Admixture with wolves or the ancestor of dogs and wolves appears to generate the extreme upward bias in the inferred ancestral jackal N_e. In PSMC inferences from simulated jackal demographic histories the presence of jackal - dog/wolf ancestor and jackal - Israeli wolf migration bands (Figures S8.2.6 – S8.2.8) produced an artefactual spike in the jackal N_e trajectory. This sharp peak is similar to the one observed in the empirical data from the golden jackal, although in the N_e trajectory reconstructed from that data, the peak is slightly more recent. Overall, we conclude the peak in the N_e trajectory observed in the data is likely due to post- divergence gene flow between ancestors of contemporary golden jackals and Israeli wolves or the ancestor of dogs and wolves. Ongoing work has found evidence for multiple highly divergent jackal or jackal-like lineages in Africa and the Middle East (Koepfli et al., unpublished data).

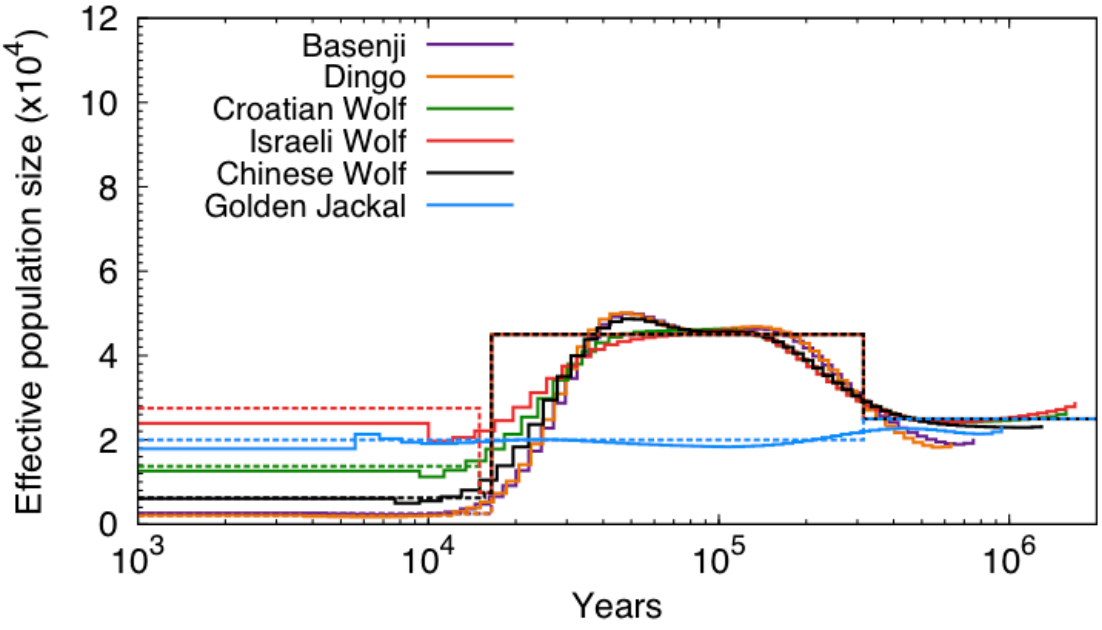


Figure S8.2.4. N_e trajectories of 6 canid lineages reconstructed using the PSMC method of Li and Durbin [3], for data simulated under the *G-PhoCS* inferred demographic history, excluding migration bands. The dotted lines show the actual N_e trajectories whereas the solid lines represent the inferred N_e trajectories.

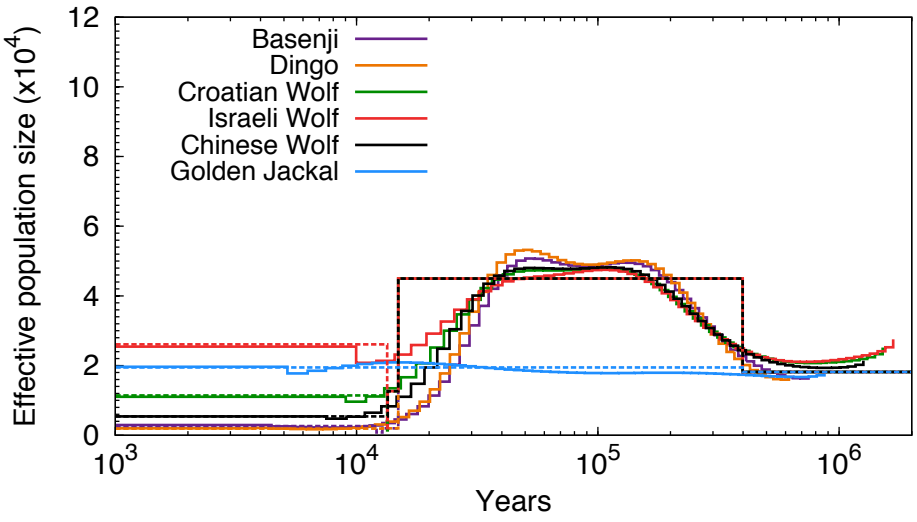


Figure S8.2.5. N_e trajectories of 6 canid lineages reconstructed using the PSMC method of Li and Durbin [3] for data simulated under the *G-PhoCS* inferred demographic history, only including gene flow from the golden jackal to the ancestor of dogs and wolves. Inferred N_e trajectories are shown with solid lines and the actual N_e trajectories are displayed with dotted lines.

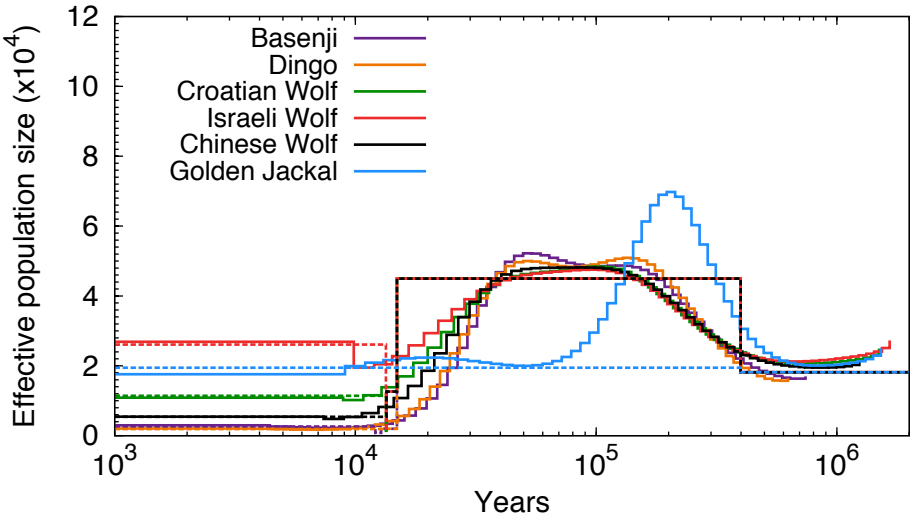


Figure S8.2.6. N_e trajectories of 6 canid lineages reconstructed using the PSMC method of Li and Durbin [3] for data simulated under the *G-PhoCS* inferred demographic history, only including gene flow from the ancestor of dogs and wolves to golden jackal. Inferred N_e trajectories are shown with solid lines and the actual N_e trajectories are displayed with dotted lines.

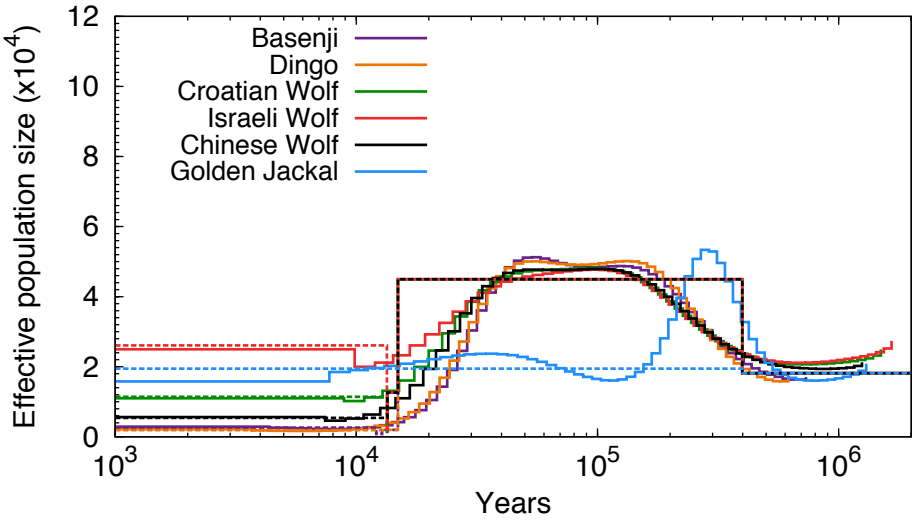


Figure S8.2.7. N_e trajectories of 6 canid lineages reconstructed using the PSMC method of Li and Durbin [3] for data simulated under the *G-PhoCS* inferred demographic history, only including gene flow from Israeli wolf to golden jackal. Inferred N_e trajectories are shown with solid lines and the actual N_e trajectories are displayed with dotted lines.

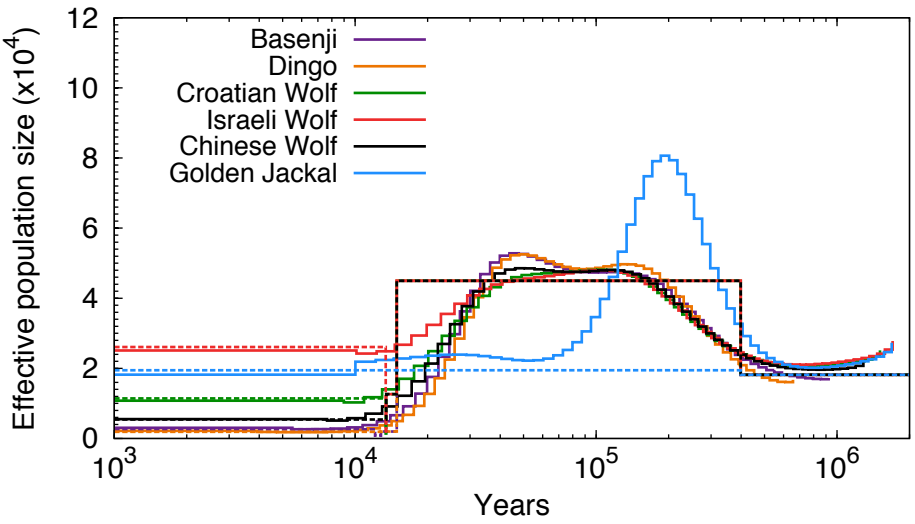


Figure S8.2.8. N_e trajectories of 6 canid lineages reconstructed using the PSMC method of Li and Durbin [3], for data simulated under the *G-PhoCS* inferred demographic history, including all detected gene flow. The actual N_e trajectories are shown as dotted lines whereas the inferred N_e trajectories are depicted by solid lines.

S8.3 Genealogies and Incomplete Lineage Sorting

S8.3.1 Definition of Neutral Loci

To assess patterns of incomplete lineage sorting, we focused on a set of neutral loci, 1kb in length, chosen so as to reduce potential confounding effects of natural selection, following guidelines set by several previous studies [1,8]. To create this set of loci, we scanned the boxer genome, examining sliding 1kb windows with a step size of 50bp. To be included in the neutral loci set, a region had to pass the following filters: 1) no coding DNA; 2) located at least 100kb away from the nearest gene (both "known" and predicted); 3) GC content within two standard deviations of the mean GC content of the boxer genome; 4) within 1kb, no 50bp window with a PhastCons score >0.5; 5) within 1 kb, no two consecutive 50bp windows with a mappability score >2, with mappability computed using the program TALLYMER [9]; 6) no RepeatMasked elements with divergence less than 25%; and 7) no N's in boxer reference genome. Loci were further selected to be located at least 50kb from one another, leading to a total of 5139 markers, 5073 of which were autosomal. Within each locus, CpG sites present within any of the genomes were masked from further analysis in all genomes.

S8.3.2 Neighbor-joining Trees

For the above 5073 neutral loci (see Text S9.3.1) we reconstructed putative genealogies using the neighbor-joining method as implemented in PHYLIP with the pairwise differences being calculated as in E9.1.

S8.3.3 Coalescent simulations

In order to compare the distribution of genealogies to those expected under the demographic history of dogs and wolves, we simulated genealogies of 5073 1-kb segments with the program *ms* [10] under the demographic history inferred by *G-PhoCS* (see Text S9), and then built a NJ tree from this simulated data. We repeated this procedure 1,000 times using the command line (Command Line 6).

From the 1,000 simulated genealogies, we counted the proportion of those in which dogs were monophyletic, the proportion of times we observed a particular outgroup to dogs (conditional on dog monophyly), and the frequency of different outgroups to the Israeli wolf. We report this last set of statistics because previous research on dog domestication found an excess of haplotype sharing between dogs and Israeli wolves [11], and because we detected substantial admixture between the Israeli wolf and basenji (see Text S8.4). Results from simulations are all reported as the mean values of the 1,000 runs.

S8.3.4. Results

In 385 of the 5073 genealogies recovered from our neutral loci, all branch lengths were equal to 0, and we excluded these from subsequent analyses. Within the remaining 4688 genealogies, 365 (7.79%, binomial 95%CI = 7.02% - 8.55%) contained a monophyletic dog clade. For the simulated genealogies, 212 (4.23%, 95%CI = 3.69% - 4.78%) contained a monophyletic dog clade. The neutral loci and simulated data contain different proportions of trees in which dogs are monophyletic. In both the empirical and simulated data, within the set of genealogies in which dogs were monophyletic, dogs did not have clear outgroup in most trees (neutral loci: 157 trees, 3.35%; simulated genealogies 158 trees, 3.16%; labelled 'NA in Figure S8.3.4.1A). These relatively high frequencies of neutral genealogies that are discordant with the genome-wide species tree point to a combination of a) a lack of resolution due to too few mutations within a 1-kb segment to resolve relationships, and b) incomplete lineage sorting, likely due to both the relatively recent timing of divergence, and recurrent admixture between wild and domestic canids.

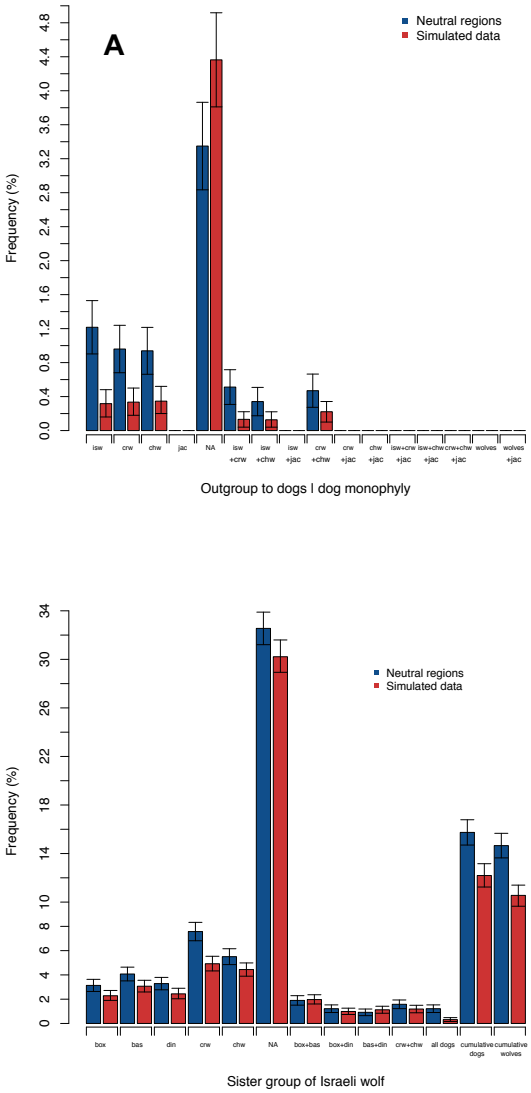


Figure S8.3.4.1. For both neutral loci extracted from sequencing data, and simulated histories inferred from *G- PhoCS*, (A) the frequency of different outgroups to dogs when dogs are recovered as a monophyletic clade in NJ trees, and (B) frequencies of outgroups to the Israeli wolf.

For the remainder of genealogies derived from empirical data, the Israeli wolf is the most common outgroup to dogs (57 trees, 1.22%), with the Croatian and Chinese wolves appearing at similar lower frequencies. In contrast, in the simulated data, the Chinese wolf is the most common outgroup to dogs (0.248%), although the proportions among the three wolves are very similar (isw: 0.237%; crw: 0.241%). Although the most frequent outgroup to dogs in neutral loci is Israeli wolf, the 95%CIs for the three wolves are overlapping (Figure 8.3.4.1A). In both neutral loci and simulated data, no trees were recovered in which a monophyletic wolf clade was sister to the dog clade. Inconsistent with the genome-wide species tree, in both the empirical and simulated data polytomies frequently preclude an assignment of an outgroup to the Israeli wolf. However, in those

cases where an outgroup can be assigned, both empirical and simulated data identify the Croatian wolf as the most common outgroup to the Israeli wolf, followed by the Chinese wolf (Figure 8.3.4.1B). Consistent with Israeli wolf - Basenji admixture, of the three dogs the Basenji was the most frequent outgroup to the Israeli wolf.

S8.4 Post-Divergence Gene Flow

To investigate the extent of gene flow between wolves and dogs subsequent to their divergence, we employed a method recently developed by Durand *et al.* [12]. This method tests for gene flow by testing for asymmetries in allele sharing between a source lineage (P3), and either of two receiving lineages (P1, P2). In this case, the ancestor of P1 and P2 is sister to the ancestor of P3. Given a site that is bi-allelic in (P1, P2, P3) where P3 is in state B and an outgroup (O) is in state A, there are two possible allelic configurations of P1-P2-P3-O that are informative with respect to gene flow between P3 and either P1 or P2: ABBA and BABA. In the absence of lineage- specific post-divergence gene flow and under selective neutrality, the genome-wide frequency of these configurations should be approximately equal. Thus, the null hypothesis is that there has not been gene flow between P3 and P1 or P2 after the divergence of P3 from P1 and P2. We defined an ABBA site as a site where P1 and the outgroup shared the same allele ‘A’ while P2 and P3 shared an alternative allele ‘B’. A site was defined as a BABA site when the outgroup and P2 shared the allele ‘A’ and the alternative allele ‘B’ was shared between P1 and P3. The rejection of the null hypothesis indicates that there has been gene flow between P3 and either P1 or P2. Deviations from the null expectation were quantified using the D-statistic:

$$D = \frac{\sum_{i=1}^{n} C_{ABBA}(i) - \sum_{i=1}^{n} C_{BABA}(i)}{\sum_{i=1}^{n} C_{ABBA}(i) + \sum_{i=1}^{n} C_{BABA}(i)} \quad \text{(Eq 8.3)}$$

where $C_{ABBA}(i)$ and $C_{BABA}(i)$ are indicator variables equal to 1 or 0 depending on the presence or absence of the ABBA and BABA sites at the i^{th} site. To calculate the D statistic, we specified the golden jackal as our outgroup, and divided the reference genome into 422 segments of 5 Mb each, excluding the chromosome ends where the remaining segment is < 5Mb. Within these segments, we used stringent filtering criteria, excluding genomic positions with missing data, and sites that failed either the GF2 or SF filters (see Text S4) For each species at each site, with the exception of the haploid boxer reference, we randomly sampled one allele from the called genotype. We then calculated the D statistic from a total of *n* sites that met our quality control filters.

To be consistent with the evolutionary history reflected in the recovered neighbor-joining tree (see S8.1 above), and to focus on gene flow most germane to evolutionary processes influencing wolf-dog divergence, we restricted testing to those cases where when one of the dog samples was P3, the other two (P1 and P2) were wolves, and vice versa (P3=wolf, P1 & P2 = dogs). Using these criteria, and including the boxer reference among the dogs, 18 tests were possible.

Following Durand *et al.* [12], the standard error of the statistic was calculated using a jackknife procedure [13]. A Z-score was then obtained by dividing the value of the D statistic by its standard error. Z-scores with an absolute value ≥3 were considered significant. Rejection of the null hypothesis indicates that there has been gene flow between P3 and either P1 or P2 [14]. Negative significant Z scores indicate gene flow between P1 and P3 while positive significant Z scores indicate gene flow between P2 and P3.

We found evidence for post-divergence gene flow between three pairs of samples: basenji/Israeli wolf, boxer/Israeli wolf, and dingo/Chinese wolf (Table S9.4.1). The mean absolute value of Z was highest in basenji/Israeli wolf ($\hat{Z}$ = 9.27; range = 5.64 -12.11), compared to Chinese wolf/Dingo $\hat{Z}$ = 6.58; range = 3.58 – 10.14), and Israeli wolf/Boxer ($\hat{Z}$ = 6.15; range = 5.33 - 6.71).

Because calculation of the D statistic does not account for the effects of gene flow between the outgroup and any of the three samples considered under a given test, it is possible that such gene flow could introduce bias. In particular, our analyses using *G-PhoCS* support gene flow between the jackal and the Israeli wolf and jackal and the ancestral wolf. Nevertheless, our ABBA/BABA results are not affected by this gene flow for the following reasons. First, only the gene flow with Israeli wolf could affect the calculation of the D statistic. Thus, this gene flow would not affect tests that did not include the Israeli wolf. Second, this gene flow would not affect tests with two dogs and one wolf (dog,dog,wolf,jackal = 1,2,3,4), as Israeli wolf –jackal gene flow would lead to an allelic configuration that is **AA or **BB and thus not evaluated in the test. It is possible that, in tests with two wolves (one of which is the Israeli wolf), jackal- Israeli wolf admixture could give appearance of gene flow between the dog in question and the other wolf in the test. For example, consider a test that includes Israeli wolf, Croatian wolf, Basenji, and Golden Jackal. If the 'B' allele resulted from a mutation that arose in the ancestor to dogs and wolves, the original configuration would be BBBA, but Israeli wolf –jackal admixture would convert it to ABBA, leading to an upwardly biased count of this configuration, which would contribute to a Croatian wolf-Basenji gene flow signal. Nevertheless, we found in all tests with two wolves and one dog that include the Israeli wolf, the significant gene flow that is detected is between the Israeli wolf and the dog in

question, the exact opposite of what would be expected if Israeli wolf –jackal gene flow were biasing the test statistic.

Table S8.4.1. Estimation of post-divergence gene flow using the D Statistic [12]. The outgroup in all comparisons is the golden jackal. Statistical significance is evaluated using a two-tailed Z test, with the additional requirement that that absolute value of the Z-score to be ≥3. Significant tests and sample pairs showing evidence for post-divergence gene flow are shown in bold.

P1	P2	P3	ABBA Sites	BABA Sites	D (%)	SE (%)	Z	p-value
Basenji	Dingo	Croatian wolf	164211	162364	0.57%	0.40%	1.42	0.16
Basenji	Dingo	**Israeli wolf**	158610	179656	-6.22%	0.51%	**-12.21**	**2.79×10^{-34}**
Boxer	Basenji	Croatian wolf	144942	146113	-0.40%	0.46%	-0.88	0.38
Boxer	**Basenji**	**Israeli wolf**	157007	147991	2.96%	0.52%	**5.64**	**1.67×10^{-8}**
Boxer	Dingo	Croatian wolf	177485	176031	0.41%	0.44%	0.94	0.35
Boxer	Dingo	**Israeli wolf**	176511	189294	-3.49%	0.52%	**-6.71**	**1.96×10^{-11}**
Croatian wolf	**Israeli wolf**	**Boxer**	226123	210897	3.48%	0.65%	**5.33**	**9.86×10^{-8}**
Croatian wolf	Israeli wolf	Dingo	213742	212876	0.20%	0.54%	0.38	0.71
Croatian wolf	**Israeli wolf**	**Basenji**	205695	182191	6.06%	0.62%	**9.74**	**1.99×10^{-22}**
Basenji	**Dingo**	**Chinese wolf**	173366	162030	3.38%	0.45%	**7.49**	**6.76×10^{-14}**
Boxer	Basenji	Chinese wolf	149172	147273	0.64%	0.41%	1.54	0.12
Boxer	**Dingo**	**Chinese wolf**	192400	175946	4.47%	0.44%	**10.14**	**3.77×10^{-24}**
Croatian wolf	Chinese wolf	Boxer	216145	219859	-0.85%	0.42%	-2.02	4.32×10^{-2}
Croatian wolf	**Chinese wolf**	**Dingo**	221737	212060	2.23%	0.44%	**5.10**	**3.48×10^{-7}**
Croatian wolf	Chinese wolf	Basenji	190706	191336	-0.16%	0.39%	-0.42	0.68
Chinese wolf	**Israeli wolf**	**Boxer**	242452	222327	4.33%	0.68%	**6.41**	**1.43×10^{-10}**
Chinese wolf	Israeli wolf	**Dingo**	223003	232071	-1.99%	0.56%	**-3.58**	**3.48×10^{-4}**
Chinese wolf	**Israeli wolf**	**Basenji**	216213	191475	6.07%	0.64%	**9.50**	**2.02×10^{-21}**

S8.5 Model fit using the ABBA/BABA/BBAA configurations statistics

We tested the fit of the three models analyzed with *G-PhoCS* using the proportion of sites that contain alleles that are shared between two lineages but not the other two when comparing four species. The ABBA and BABA sites are defined following the notation seen in Section S8.4. On the other hand, a BBAA site is defined as one where the lineages P_1 and P_2 share one allele while the two other lineages P_3 and O share a different allele. The proportion of those three types of sites is reflective of the genealogies contained in the data when comparing four lineages, where those genealogies are affected by gene flow and the divergence time between species. For a quartet of lineages

P_1, P_2, P_3 and O we estimated the frequency of a site being ABBA, BABA or BBAA given that there are two alleles, each present in two of the four species as:

$$f(ABBA \mid two\ alleles, each\ in\ two\ species) = \frac{N(ABBA)}{N(ABBA) + N(BABA) + N(BBAA)} \quad \text{(Eq 8.4)}$$

$$f(BABA \mid two\ alleles, each\ in\ two\ species) = \frac{N(BABA)}{N(ABBA) + N(BABA) + N(BBAA)} \quad \text{(Eq 8.5)}$$

$$f(BBAA \mid two\ alleles, each\ in\ two\ species) = \frac{N(BBAA)}{N(ABBA) + N(BABA) + N(BBAA)} \quad \text{(Eq 8.6)}$$

We refer to these estimates as relative frequencies of ABBA, BABA and BBAA sites, respectively. In the equations, N(ABBA), N(BABA) and N(BBAA) are the number of ABBA, BABA and BBAA sites.

The counts of ABBA, BABA and BBAA sites in the data were calculated using the 18 quartet configurations that are shown in Table S8.4.1 with two additional quartet configurations that contain either three dogs or three wolves. Those two additional configurations were added because they are informative about the actual phylogenetic relationships inside dogs and inside wolves, respectively. A demographic model would be more likely to be correct if it captures similar values for Eq.8.4-8.6 as those seen in data. The estimates of the number of ABBA/BABA/BBAA sites in the data are shown in Table S8.5.1, along with the estimates of the relative frequency of those sites.

To mimic the empirical analysis (see above) we initially simulated 422 regions of 5Mb using the three models analyzed by *G-PhoCS.* However, because this produced an excess of ABBA/BABA/BBAA sites, to match the counts of these site classes seen in the data, we reduced our region size, instead simulating 422 regions of 2Mb. The simulations were performed using the following command lines:

1) Model where the dogs and wolves are each a separate clade (Command Line 7). This command line is identical to Command Line 1, with the only difference being the number of bases simulated.

2) Regional domestication model (Command Line 8)

3) Origin of dogs from the Israeli wolf (Command Line 9)

As a measure of the fit of each model to the data, we calculated the total difference between each model and the data in the relative frequencies of the ABBA/BABA/BBAA sites using the following equation:

$$
\begin{aligned}
&Absolute\ Error \\
&= \sum_{i=1}^{combinations} | f(ABBA \mid two\ alleles, each\ in\ two\ species)_{model} \\
&- f(ABBA \mid two\ alleles, each\ in\ two\ species)_{data}| \\
&+ | f(BABA \mid two\ alleles, each\ in\ two\ species)_{model} \\
&- f(BABA \mid two\ alleles, each\ in\ two\ species)_{data}| \\
&+ | f(BBAA \mid two\ alleles, each\ in\ two\ species)_{model} \\
&- f(BBAA \mid two\ alleles, each\ in\ two\ species)_{data}|
\end{aligned}
\qquad \text{(Eq 8.7)}
$$

Overall, we found that the model which provided a better fit to the data, in terms of a smaller absolute error as estimated by Eq 8.7, was the model which assumes that the dogs and wolves are each a separate clade whereas the model which provided the worst fit was the one which assumes a regional domestication model (Table S8.5.2).

Using a threshold of 1.5% to look for important absolute differences between the data and the model in terms of relative frequencies, we found larger differences in the relative frequencies of BBAA sites in the data and the model that provided a better fit to the data in comparisons that included the Dingo, Chinese Wolf and another species of dog. We also found that the model which provided a better fit to the data incorrectly estimated the relative frequencies of ABBA sites in comparisons including the Chinese Wolf as P_1, Israeli wolf as P_2 and the Boxer or Basenji as P_3. Additionally, the number of BBAA sites in the quartet Boxer (P_1), Dingo (P_2) and Croatian Wolf (P_3) deviated substantially from those observed in the empirical data.

The regional domestication model overestimated the relative frequency of shared sites between Basenji and Dingo and underestimated the relative frequency of sites shared between (Dingo, Boxer) and (Boxer, Basenji) in comparisons that included the three dogs and the golden jackal. This shows that the phylogenetic relationships between dogs are more severely distorted under this model. This is also exemplified by the poor fit to the data in terms of the relative frequencies of ABBA/BABA/BBAA sites in the comparisons that include the Dingo, Boxer and another species of wolf. As in the model from Fig. 5A, the number of BBAA sites was also underestimated in the quartet Basenji (P_1), Dingo (P_2) and Chinese Wolf (P_3).

As with the best model, the model that posits the origin of dogs from the Israeli Wolf had poor fit to the data with respect to the relative frequency of BBAA sites in the comparisons of Boxer (P_1), Dingo (P_2) and Chinese Wolf (P_3). The latter model also had problems fitting the relative frequencies of the three types of sites we were inspecting in comparisons that included the Israeli Wolf, Croatian Wolf and a dog. The relative

frequency of BBAA sites in the comparison of Boxer, Dingo and Croatian Wolf was underestimated under this model.

Table S8.5.1. Estimates of the number of ABBA/BABA/BBAA sites in the six canid genomes. For each cell and each quartet comparison we report the number of ABBA/BABA/BBAA sites followed by the frequency of those three types of sites given that the site is bi-allelic with the two alleles found in two species each. The golden jackal was used as an outgroup in all comparisons.

			Data		
P1	**P2**	**P3**	**ABBA Sites**	**BABA Sites**	**BBAA Sites**
Basenji	Dingo	Croatian wolf	164211; 28.43%	162364; 28.11%	250958; 43.45%
Basenji	Dingo	Israeli wolf	158610; 27.18%	179656; 30.78%	245329; 42.04%
Boxer	Basenji	Croatian wolf	144942; 24.82%	146113; 25.02%	292896; 50.16%
Boxer	Basenji	Israeli wolf	157007; 26.71%	147991; 25.17%	282873; 48.12%
Boxer	Dingo	Croatian wolf	177485; 27.15%	176031; 26.93%	300095; 45.91%
Boxer	Dingo	Israeli wolf	176511; 26.50%	189294; 28.42%	300201; 45.07%
Croatian wolf	Israeli wolf	Boxer	226123; 34.16%	210897; 31.86%	224971; 33.98%
Croatian wolf	Israeli wolf	Dingo	213742; 32.78%	212876; 32.65%	225351; 34.56%
Croatian wolf	Israeli wolf	Basenji	205695; 35.29%	182191; 31.26%	194909; 33.44%
Basenji	Dingo	Chinese wolf	173366; 29.45%	162030; 27.52%	253270; 43.02%
Boxer	Basenji	Chinese wolf	149172; 24.91%	147273; 24.59%	302448; 50.50%
Boxer	Dingo	Chinese wolf	192400; 28.40%	175946; 25.97%	309223; 45.64%
Croatian wolf	Chinese wolf	Boxer	216145; 32.52%	219859; 33.08%	228675; 34.40%
Croatian wolf	Chinese wolf	Dingo	221737; 33.97%	212060; 32.49%	218959; 33.54%
Croatian wolf	Chinese wolf	Basenji	190706; 32.79%	191336; 32.90%	199502; 34.31%
Chinese wolf	Israeli wolf	Boxer	242452; 35.42%	222327; 32.48%	219803; 32.11%
Chinese wolf	Israeli wolf	Dingo	223003; 33.37%	232071; 34.73%	213209; 31.90%
Chinese wolf	Israeli wolf	Basenji	216213; 36.43%	191475; 32.26%	185855; 31.31%
Basenji	Dingo	Boxer	179362; 32.42%	216634; 39.16%	157265; 28.43%
Chinese Wolf	Croatian Wolf	Israeli Wolf	230181; 34.70%	208597; 31.44%	224601; 33.86%

Table S8.5.2. Estimates of the number of ABBA/BABA/BBAA sites in the three *G-PhoCS* models analyzed. For each cell and each quartet comparison we report: 1) The number of ABBA/BABA/BBAA sites; 2) The frequency of those three types of sites given that the site is bi-allelic with the two alleles found in two species each and 3) the difference of that frequency in the simulations minus what is estimated in the data (when this difference is bigger than 1.5%, we highlight the cell in bold). The lower row of the table indicates the fit of the model to the data as estimated by equation 8.7. The golden jackal was used as an outgroup in all comparisons.

			Fig. 5A model (Model where the dogs and wolves are each a separate clade)			**Fig. 5B model (Regional domestication model)**			**Fig. 5C model (Origin of dogs from the Israeli wolf)**		
P1	**P2**	**P3**	**ABBA Sites**	**BABA Sites**	**BBAA Sites**	**ABBA Sites**	**BABA Sites**	**BBAA Sites**	**ABBA Sites**	**BABA Sites**	**BBAA Sites**
Basenji	Dingo	Croatian wolf	177596; 28.53%; 0.10%	180202; 28.95%; 0.84%	264624; 42.52%; -0.94%	178773; 28.94%; 0.50%	177186; 28.68%; 0.57%	261870; 42.39%; -1.07%	178434; 28.88%; 0.45%	177152; 28.67%; 0.56%	262289; 42.45%; -1.00%
Basenji	Dingo	Israeli wolf	173506; 27.87%; 0.69%	191296; 30.72%; -0.06%	257817; 41.41%; -0.63%	173256; 27.83%; 0.65%	192556; 30.93%; 0.15%	256705; 41.24%; -0.80%	173222; 27.82%; 0.64%	188792; 30.32%; -0.46%	260580; 41.85%; -0.18%
Boxer	Basenji	Croatian wolf	157926; 25.24%; 0.42%	158158; 25.28%; 0.26%	309616; 49.48%; -0.67%	155013; 24.78%; -0.04%	156346; 24.99%; -0.03%	314275; 50.23%; 0.08%	158543; 25.42%; 0.60%	158872; 25.47%; 0.45%	306268; 49.11%; -1.05%
Boxer	Basenji	Israeli wolf	168735; 26.93%; 0.23%	155221; 24.78%; -0.40%	302524; 48.29%; 0.17%	165943; 26.52%; -0.19%	155130; 24.79%; -0.38%	304670; 48.69%; 0.57%	167349; 26.80%; 0.09%	155402; 24.89%; -0.29%	301725; 48.32%; 0.20%
Boxer	Dingo	Croatian wolf	172541; 27.69%; 0.53%	175379; 28.14%; 1.21%	**275228; 44.17%; -1.75%**	**148908; 23.69%; -3.47%**	**148654; 23.64%; -3.29%**	**331136; 52.67%; 6.76%**	172536; 27.92%; 0.76%	171583; 27.76%; 0.83%	**273917; 44.32%; -1.59%**
Boxer	Dingo	Israeli wolf	173388; 27.77%; 1.27%	177664; 28.45%; 0.03%	273358; 43.78%; -1.30%	**147173; 23.27%; -3.24%**	**155660; 24.61%; -3.82%**	**329753; 52.13%; 7.05%**	171562; 27.53%; 1.03%	175185; 28.11%; -0.31%	276446; 44.36%; -0.72%
Croatian wolf	Israeli wolf	Boxer	205879; 33.27%; -0.89%	201724; 32.60%; 0.74%	211157; 34.13%; 0.14%	208604; 33.71%; -0.44%	200215; 32.36%; 0.50%	209921; 33.93%; -0.06%	208423; 33.80%; -0.36%	**207350; 33.62%; 1.76%**	200941; 32.58%; -1.40%
Croatian wolf	Israeli wolf	Dingo	203877; 32.96%; 0.18%	201160; 32.53%; -0.13%	213431; 34.51%; -0.06%	202216; 32.78%; 0.00%	202568; 32.84%; 0.19%	212020; 34.37%; -0.19%	205800; 33.35%; 0.57%	209303; 33.92%; 1.27%	**201941; 32.73%; -1.84%**
Croatian wolf	Israeli wolf	Basenji	215597; 34.74%; -0.56%	197696; 31.85%; 0.59%	207361; 33.41%; -0.03%	216547; 34.95%; -0.35%	196012; 31.63%; 0.37%	207051; 33.42%; -0.03%	216467; 35.11%; -0.19%	**203118; 32.94%; 1.68%**	197038; 31.95%; -1.49%
Basenji	Dingo	Chinese wolf	188009; 30.16%; 0.71%	177552; 28.49%; 0.96%	**257728; 41.35%; -1.67%**	188470; 30.47%; 1.02%	174988; 28.29%; 0.77%	**254996; 41.23%; -1.79%**	185253; 29.98%; 0.53%	173424; 28.06%; 0.54%	259312; 41.96%; -1.06%
Boxer	Basenji	Chinese wolf	160801; 25.64%; 0.74%	158007; 25.20%; 0.61%	308245; 49.16%; -1.34%	156840; 25.13%; 0.22%	155804; 24.97%; 0.37%	311426; 49.90%; -0.60%	157369; 25.29%; 0.38%	159053; 25.56%; 0.97%	305845; 49.15%; -1.35%
Boxer	Dingo	Chinese wolf	184167; 29.48%; 1.09%	170916; 27.36%; 1.40%	**269545; 43.15%; -2.48%**	**159174; 25.32%; -3.08%**	**144656; 23.01%; -2.96%**	**324831; 51.67%; 6.03%**	178856; 28.94%; 0.55%	168711; 27.30%; 1.33%	**270441; 43.76%; -1.88%**

Croatian wolf	Chinese wolf	Boxer	203311; 32.95%; 0.43%	202091; 32.76%; -0.32%	211562; 34.29%; -0.11%	200348; 32.66%; 0.14%	198041; 32.28%; -0.80%	215078; 35.06%; 0.66%	204468; 33.45%; 0.93%	203864; 33.35%; 0.27%	202947; 33.20%; -1.20%
Croatian wolf	Chinese wolf	Dingo	213747; 34.53%; 0.57%	196438; 31.74%; -0.75%	208747; 33.73%; 0.18%	209895; 34.22%; 0.25%	193324; 31.52%; -0.97%	210107; 34.26%; 0.71%	210931; 34.50%; 0.53%	201135; 32.90%; 0.41%	199265; 32.60%; -0.95%
Croatian wolf	Chinese wolf	Basenji	205710; 33.27%; 0.48%	201464; 32.58%; -0.32%	211167; 34.15%; -0.15%	201556; 32.84%; 0.05%	196880; 32.08%; -0.82%	215250; 35.07%; 0.77%	203801; 33.28%; 0.49%	204552; 33.41%; 0.50%	203964; 33.31%; -1.00%
Chinese wolf	Israeli wolf	Boxer	**208018; 33.51%; -1.91%**	205083; 33.04%; 0.56%	207667; 33.45%; 1.35%	210840; 34.03%; -1.38%	204758; 33.05%; 0.58%	203911; 32.91%; 0.81%	210065; 34.00%; -1.42%	209596; 33.92%; 1.45%	198217; 32.08%; -0.03%
Chinese wolf	Israeli wolf	Dingo	200720; 32.34%; -1.03%	215312; 34.69%; -0.04%	204645; 32.97%; 1.07%	200301; 32.34%; -1.03%	217224; 35.07%; 0.34%	201859; 32.59%; 0.69%	204194; 33.06%; -0.31%	217493; 35.21%; 0.49%	195969; 31.73%; -0.18%
Chinese wolf	Israeli wolf	Basenji	**216436; 34.81%; -1.62%**	202781; 32.61%; 0.35%	202571; 32.58%; 1.27%	217724; 35.14%; -1.29%	201865; 32.58%; 0.32%	199982; 32.28%; 0.96%	218547; 35.38%; -1.05%	204447; 33.10%; 0.84%	194752; 31.53%; 0.21%
Basenji	Dingo	Boxer	190695; 31.36%; -1.06%	242304; 39.85%; 0.69%	175036; 28.79%; 0.36%	**244189; 40.10%; 7.69%**	**219636; 36.07%; -3.08%**	**145058; 23.82%; -4.60%**	192265; 31.81%; -0.60%	237327; 39.27%; 0.12%	174739; 28.91%; 0.49%
Chinese Wolf	Croatian Wolf	Israeli Wolf	208874; 33.63%; -1.06%	203245; 32.73%; 1.28%	208912; 33.64%; -0.22%	206703; 33.38%; -1.32%	198457; 32.05%; 0.61%	214034; 34.57%; 0.71%	204824; 33.27%; -1.43%	200458; 32.56%; 1.12%	210316; 34.16%; 0.31%
		Absolute Error	0.4298			0.8219			0.4668		

Simulation Command Lines Command Line 1. *G-PhoCS* model with the full set of migration bands inferred:

```
./macs 13 30000000 -t 1 -r 0.92 -I 7 1 2 2 2 2 2 2 -n
1 0.000010 -n 2 0.000106 -n 3 0.000077 -n 4 0.001044 -
n 5 0.000457 -n 6 0.000217 -n 7 0.000778 -m 2 4 4505.0
-m 4 2 1840.0 -m 3 6 573.0 -m 6 3 942.0 -m 4 7 58.0 -m
7 4 1162.0 -ej 0.0000403 2 1 -en 0.0000403 1 0.000032
-em 0.0000403 1 4 0.0 -em 0.0000403 4 1 0.0 -em
0.0000403 2 4 0.0 -em 0.0000403 4 2 0.0 -ej 0.0000427
3 1 -en 0.0000427 1 0.000080 -em 0.0000427 1 6 0.0 -em
0.0000427 6 1 0.0 -em 0.0000427 3 6 0.0 -em 0.0000427
6 3 0.0 -ej 0.0000446 5 4 -en 0.0000446 4 0.000056 -em
0.0000446 1 4 0.0 -em 0.0000446 4 1 0.0 -em 0.0000446
4 7 0.0 -em 0.0000446 7 4 0.0 -ej 0.0000449 6 4 -en
0.0000449 4 0.000505 -em 0.0000449 1 4 0.0 -em
0.0000449 4 1 0.0 -em 0.0000449 4 7 0.0 -em 0.0000449
7 4 0.0 -ej 0.0000496 4 1 -en 0.0000496 1 0.001800 -em
0.0000496 1 4 0.0 -em 0.0000496 4 1 0.0 -em 0.0000496 4 7
0.0 -em 0.0000496 7 4 0.0 -em 0.0000496 1 7 17.0 -
em 0.0000496 7 1 746.0 -ej 0.0013275 7 1 -en 0.0013275
1 0.000727 -em 0.0013275 1 7 0.0 -em 0.0013275 7 1 0.0
```

Command Line 2. The model inferred from *G-PhoCS* but with no gene flow between any species at any time:

```
./macs 13 30000000 -t 1 -r 0.92 -I 7 1 2 2 2 2 2 -n
1 0.000010 -n 2 0.000106 -n 3 0.000077 -n 4 0.001044 -
n 5 0.000457 -n 6 0.000217 -n 7 0.000778 -m 2 4 0.0 -m
4 2 0.0 -m 3 6 0.0 -m 6 3 0.0 -m 4 7 0.0 -m 7 4 0.0 -
ej 0.0000403 2 1 -en 0.0000403 1 0.000032 -em
0.0000403 1 4 0.0 -em 0.0000403 4 1 0.0 -em 0.0000403
2 4 0.0 -em 0.0000403 4 2 0.0 -ej 0.0000427 3 1 -en
0.0000427 1 0.000080 -em 0.0000427 1 6 0.0 -em
0.0000427 6 1 0.0 -em 0.0000427 3 6 0.0 -em 0.0000427
6 3 0.0 -ej 0.0000446 5 4 -en 0.0000446 4 0.000056 -em
0.0000446 1 4 0.0 -em 0.0000446 4 1 0.0 -em 0.0000446
4 7 0.0 -em 0.0000446 7 4 0.0 -ej 0.0000449 6 4 -en
0.0000449 4 0.000505 -em 0.0000449 1 4 0.0 -em
0.0000449 4 1 0.0 -em 0.0000449 4 7 0.0 -em 0.0000449
7 4 0.0 -ej 0.0000496 4 1 -en 0.0000496 1 0.001800 -em
0.0000496 1 4 0.0 -em 0.0000496 4 1 0.0 -em 0.0000496
4 7 0.0 -em 0.0000496 7 4 0.0 -em 0.0000496 1 7 0.0 -
em 0.0000496 7 1 0.0 -ej 0.0013275 7 1 -en 0.0013275 1
0.000727 -em 0.0013275 1 7 0.0 -em 0.0013275 7 1 0.0
```

Command Line 3. The model inferred from *G-PhoCS* but with only one event of gene flow, from the golden jackal to the ancestor of dogs and wolves:

```
./macs 13 30000000 -t 1 -r 0.92 -I 7 1 2 2 2 2 2 -n
1 0.000010 -n 2 0.000106 -n 3 0.000077 -n 4 0.001044 -
n 5 0.000457 -n 6 0.000217 -n 7 0.000778 -m 2 4 0.0 -m
4 2 0.0 -m 3 6 0.0 -m 6 3 0.0 -m 4 7 0.0 -m 7 4 0.0 -
ej 0.0000403 2 1 -en 0.0000403 1 0.000032 -em
0.0000403 1 4 0.0 -em 0.0000403 4 1 0.0 -em 0.0000403
2 4 0.0 -em 0.0000403 4 2 0.0 -ej 0.0000427 3 1 -en
0.0000427 1 0.000080 -em 0.0000427 1 6 0.0 -em
0.0000427 6 1 0.0 -em 0.0000427 3 6 0.0 -em 0.0000427
6 3 0.0 -ej 0.0000446 5 4 -en 0.0000446 4 0.000056 -em
0.0000446 1 4 0.0 -em 0.0000446 4 1 0.0 -em 0.0000446
4 7 0.0 -em 0.0000446 7 4 0.0 -ej 0.0000449 6 4 -en
0.0000449 4 0.000505 -em 0.0000449 1 4 0.0 -em
0.0000449 4 1 0.0 -em 0.0000449 4 7 0.0 -em 0.0000449
7 4 0.0 -ej 0.0000496 4 1 -en 0.0000496 1 0.001800 -em
0.0000496 1 4 0.0 -em 0.0000496 4 1 0.0 -em 0.0000496
4 7 0.0 -em 0.0000496 7 4 0.0 -em 0.0000496 1 7 17.0 -
em 0.0000496 7 1 0.0 -ej 0.0013275 7 1 -en 0.0013275 1
0.000727 -em 0.0013275 1 7 0.0 -em 0.0013275 7 1 0.0
```

Command Line 4. The model inferred from *G-PhoCS* but with only one event of gene flow, from the ancestor of dogs and wolves to golden jackal:

```
./macs 13 30000000 -t 1 -r 0.92 -I 7 1 2 2 2 2 2 2 -n 1
0.000010 -n 2 0.000106 -n 3 0.000077 -n 4 0.001044 -n 5
0.000457 -n 6 0.000217 -n 7 0.000778 -m 2 4 0.0 -m 4 2 0.0
-m 3 6 0.0 -m 6 3 0.0 -m 4 7 0.0 -m 7 4 0.0 -ej 0.0000403 2
1 -en 0.0000403 1 0.000032 -em 0.0000403 1 4 0.0 -em
0.0000403 4 1 0.0 -em 0.0000403 2 4 0.0 -em 0.0000403 4 2
0.0 -ej 0.0000427 3 1 -en 0.0000427 1 0.000080 -em
0.0000427 1 6 0.0 -em 0.0000427 6 1 0.0 -em 0.0000427 3 6
0.0 -em 0.0000427 6 3 0.0 -ej 0.0000446 5 4 -en 0.0000446 4
0.000056 -em 0.0000446 1 4 0.0 -em 0.0000446 4 1 0.0 -em
0.0000446 4 7 0.0 -em 0.0000446 7 4 0.0 -ej 0.0000449 6 4 -
en 0.0000449 4 0.000505 -em 0.0000449 1 4 0.0 -em 0.0000449
4 1 0.0 -em 0.0000449 4 7 0.0 -em 0.0000449 7 4 0.0 -ej
0.0000496 4 1 -en 0.0000496 1 0.001800 -em 0.0000496 1 4
0.0 -em 0.0000496 4 1 0.0 -em 0.0000496 4 7 0.0 -em
0.0000496 7 4 0.0 -em 0.0000496 1 7 0.0 -em 0.0000496 7 1
746.0 -ej 0.0013275 7 1 -en 0.0013275 1 0.000727 -em
0.0013275 1 7 0.0 -em 0.0013275 7 1 0.0
```

Command Line 5. The model inferred from *G-PhoCS* but with only one event of gene flow, from Israeli wolf to golden jackal:

```
./macs 13 30000000 -t 1 -r 0.92 -I 7 1 2 2 2 2 2 2 -n 1
0.000010 -n 2 0.000106 -n 3 0.000077 -n 4 0.001044 -n 5
0.000457 -n 6 0.000217 -n 7 0.000778 -m 2 4 0.0 -m 4 2 0.0
-m 3 6 0.0 -m 6 3 0.0 -m 4 7 0.0 -m 7 4 1162.0 -ej
0.0000403 2 1 -en 0.0000403 1 0.000032 -em 0.0000403 1 4
0.0 -em 0.0000403 4 1 0.0 -em 0.0000403 2 4 0.0 -em
0.0000403 4 2 0.0 -ej 0.0000427 3 1 -en 0.0000427 1
0.000080 -em 0.0000427 1 6 0.0 -em 0.0000427 6 1 0.0 -em
0.0000427 3 6 0.0 -em 0.0000427 6 3 0.0 -ej 0.0000446 5 4 -
en 0.0000446 4 0.000056 -em 0.0000446 1 4 0.0 -em 0.0000446
4 1 0.0 -em 0.0000446 4 7 0.0 -em 0.0000446 7 4 0.0 -ej
0.0000449 6 4 -en 0.0000449 4 0.000505 -em 0.0000449 1 4
0.0 -em 0.0000449 4 1 0.0 -em 0.0000449 4 7 0.0 -em 0.0000449
7 4 0.0 -ej 0.0000496 4 1 -en 0.0000496 1
0.001800 -em 0.0000496 1 4 0.0 -em 0.0000496 4 1 0.0 -em
0.0000496 4 7 0.0 -em 0.0000496 7 4 0.0 -em 0.0000496 1 7
0.0 -em 0.0000496 7 1 0.0 -ej 0.0013275 7 1 -en 0.0013275 1
0.000727 -em 0.0013275 1 7 0.0 -em 0.0013275 7 1 0.0
```

Command Line 6. *ms* command line that uses the demographic history estimated from

G- PhoCS.

```
./ms 7 1 -t 1000 -r 920 1000 -I 7 1 1 1 1 1 1 1 -n 1
0.000010 -n 2 0.000106 -n 3 0.000077 -n 4 0.001044 -n 5
0.000457 -n 6 0.000217 -n 7 0.000778 -m 2 4 4505.0 -m 4 2
1840.0 -m 3 6 573.0 -m 6 3 942.0 -m 4 7 58.0 -m 7 4 1162.0
-ej 0.0000403 2 1 -en 0.0000403 1 0.000032 -em 0.0000403 1
4 0.0 -em 0.0000403 4 1 0.0 -em 0.0000403 2 4 0.0 -em
0.0000403 4 2 0.0 -ej 0.0000427 3 1 -en 0.0000427 1
0.000080 -em 0.0000427 1 6 0.0 -em 0.0000427 6 1 0.0 -em
0.0000427 3 6 0.0 -em 0.0000427 6 3 0.0 -ej 0.0000446 5 4 -
en 0.0000446 4 0.000056 -em 0.0000446 1 4 0.0 -em 0.0000446
4 1 0.0 -em 0.0000446 4 7 0.0 -em 0.0000446 7 4 0.0 -ej
0.0000449 6 4 -en 0.0000449 4 0.000505 -em 0.0000449 1 4
0.0 -em 0.0000449 4 1 0.0 -em 0.0000449 4 7 0.0 -em
0.0000449 7 4 0.0 -ej 0.0000496 4 1 -en 0.0000496 1
0.001800 -em 0.0000496 1 4 0.0 -em 0.0000496 4 1 0.0 -em
0.0000496 4 7 0.0 -em 0.0000496 7 4 0.0 -em 0.0000496 1 7
17.0 -em 0.0000496 7 1 746.0 -ej 0.0013275 7 1 -en
0.0013275 1 0.000727 -em 0.0013275 1 7 0.0 -em 0.0013275 7
1 0.0
```

Command Line 7. Model where the dogs and wolves are each a separate clade, identical to Command Line 1, except for the simulation of smaller (2Mb) genomic regions.

```
./macs 13 2000000 -t 1 -r 0.92 -I 7 1 2 2 2 2 2 2 -n 1
0.000010 -n 2 0.000106 -n 3 0.000077 -n 4 0.001044 -n
5 0.000457 -n 6 0.000217 -n 7 0.000778 -m 2 4 4505.0 -
m 4 2 1840.0 -m 3 6 573.0 -m 6 3 942.0 -m 4 7 58.0 -m
7 4 1162.0 -ej 0.0000403 2 1 -en 0.0000403 1 0.000032
-em 0.0000403 1 4 0.0 -em 0.0000403 4 1 0.0 -em
0.0000403 2 4 0.0 -em 0.0000403 4 2 0.0 -ej 0.0000427
3 1 -en 0.0000427 1 0.000080 -em 0.0000427 1 6 0.0 -em
0.0000427 6 1 0.0 -em 0.0000427 3 6 0.0 -em 0.0000427
6 3 0.0 -ej 0.0000446 5 4 -en 0.0000446 4 0.000056 -em
0.0000446 1 4 0.0 -em 0.0000446 4 1 0.0 -em 0.0000446
4 7 0.0 -em 0.0000446 7 4 0.0 -ej 0.0000449 6 4 -en
0.0000449 4 0.000505 -em 0.0000449 1 4 0.0 -em
0.0000449 4 1 0.0 -em 0.0000449 4 7 0.0 -em 0.0000449
7 4 0.0 -ej 0.0000496 4 1 -en 0.0000496 1 0.001800 -em
0.0000496 1 4 0.0 -em 0.0000496 4 1 0.0 -em 0.0000496 4 7
0.0 -em 0.0000496 7 4 0.0 -em 0.0000496 1 7 17.0 -
em 0.0000496 7 1 746.0 -ej 0.0013275 7 1 -en 0.0013275
1 0.000727 -em 0.0013275 1 7 0.0 -em 0.0013275 7 1 0.0
```

Command Line 8. Regional domestication model.

```
./macs 13 2000000 -t 1 -r 0.92 -I 7 1 2 2 2 2 2 -n 1
0.000010 -n 2 0.000128 -n 3 0.000032 -n 4 0.000889 -n
5 0.000565 -n 6 0.000171 -n 7 0.000771 -m 1 2 20054 -m
2 1 59 -m 1 3 3459 -m 3 1 9560 -m 2 3 51 -m 3 2 7618 -
m 4 5 5276 -m 5 4 48 -m 4 6 19 -m 6 4 4958 -m 5 6 26 -
m 6 5 5312 -m 4 7 182.0 -m 7 4 1207.0 -ej 0.0000478 4
2 -en 0.0000478 2 0.000437 -em 0.0000478 1 2 0.0 -em
0.0000478 2 1 0.0 -em 0.0000478 2 3 0.0 -em 0.0000478
3 2 0.0 -em 0.0000478 4 5 0.0 -em 0.0000478 5 4 0.0 -
em 0.0000478 4 6 0.0 -em 0.0000478 6 4 0.0 -em
0.0000478 4 7 0.0 -em 0.0000478 7 4 0.0  -ej 0.0000614
5 1 -en 0.0000614 1 0.000162 -em 0.0000478 1 2 0.0 -em
0.0000478 2 1 0.0 -em 0.0000478 1 3 0.0 -em 0.0000478
3 1 0.0 -em 0.0000478 4 5 0.0 -em 0.0000478 5 4 0.0 -
em 0.0000478 5 6 0.0 -em 0.0000478 6 5 0.0 -ej
0.0000617 6 3 -en 0.0000617 3 0.000017 -em 0.0000478 3
2 0.0 -em 0.0000478 2 3 0.0 -em 0.0000478 1 3 0.0 -em
0.0000478 3 1 0.0 -em 0.0000478 6 5 0.0 -em 0.0000478
5 6 0.0 -em 0.0000478 4 6 0.0 -em 0.0000478 6 4 0.0 -
ej 0.0000618 2 1 -en 0.0000618 1 0.000252 -ej
0.0000626 3 1 -en 0.0000626 1 0.001790 -em 0.0000626 1
7 3.0 -em 0.0000626 7 1 782.0 -ej 0.0013859 7 1 -en
0.0013859 1 0.000682 -em 0.0013859 1 7 0.0 -em
0.0013859 7 1 0.0
```

Command Line 9. Origin of dogs from the Israeli wolf.

```
./macs 13 2000000 -t 1 -r 0.92 -I 7 1 2 2 2 2 2 -n 1
0.000010 -n 2 0.000103 -n 3 0.000076 -n 4 0.000894 -n
5 0.000445 -n 6 0.000221 -n 7 0.000765 -m 2 4 5032.0 -
m 4 2 1196.0 -m 3 6 865.0 -m 6 3 524.0 -m 4 7 142.0 -m
7 4 1063.0 -ej 0.0000401 2 1 -en 0.0000401 1 0.000025
-em 0.0000401 1 4 0.0 -em 0.0000401 4 1 0.0 -em
0.0000401 2 4 0.0 -em 0.0000401 4 2 0.0 -ej 0.0000419
3 1 -en 0.0000419 1 0.000029 -em 0.0000419 1 6 0.0 -em
0.0000419 6 1 0.0 -em 0.0000419 3 6 0.0 -em 0.0000419
6 3 0.0 -ej 0.0000444 4 1 -en 0.0000444 1 0.000186 -em
0.0000444 1 4 0.0 -em 0.0000444 4 1 0.0 -em 0.0000444
4 7 0.0 -em 0.0000444 7 4 0.0 -ej 0.0000447 5 1 -en 0.0000447
1 0.000229 -em 0.0000447 1 4 0.0 -em
0.0000447 4 1 0.0 -em 0.0000447 4 7 0.0 -em 0.0000447
7 4 0.0 -ej 0.0000450 6 1 -en 0.0000450 1 0.001801 -em
0.0000450 1 4 0.0 -em 0.0000450 4 1 0.0 -em 0.0000450
4 7 0.0 -em 0.0000450 7 4 0.0 -em 0.0000450 1 7 5.0 -
em 0.0000450 7 1 778.0 -ej 0.0013954 7 1 -en 0.0013954
1 0.000663 -em 0.0013954 1 7 0.0 -em 0.0013954 7 1 0.0
```

References

1. Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A (2011) Bayesian inference of ancient human demography from individual genome sequences. Nature Genetics 43: 1031- U1151.

2. Felsenstein J (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5: 164. 3. Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature 475: 493-U484.

4. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438: 803-819.

5. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 38: 904-909.

6. Chen GK, Marjoram P, Wall JD (2009) Fast and flexible simulation of DNA sequence data. Genome Research 19: 136-142.

7. Wong AK, Ruhe AL, Dumont BL, Robertson KR, Guerrero G, et al. (2010) A Comprehensive Linkage Map of the Dog Genome. Genetics 184: 595-U436.

8. Wall JD, Cox MP, Mendez FL, Woerner A, Severson T, et al. (2008) A novel DNA sequence database for analyzing human demographic history. Genome Research 18: 1354-1361.

9. Kurtz S, Narechania A, Stein JC, Ware D (2008) A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. Bmc Genomics 9.

10. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337-338.

11. vonHoldt BM, Pollinger JP, Lohmueller KE, Han EJ, Parker HG, et al. (2010) Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. Nature 464: 898-U109.

12. Durand EY, Patterson N, Reich D, Slatkin M (2011) Testing for Ancient Admixture

between Closely Related Populations. Molecular Biology and Evolution 28: 2239-2252.

13. Efron B (1981) Nonparametric Estimates of Standard Error - the Jackknife, the Bootstrap and Other Methods. Biometrika 68: 589-599.

14. Rasmussen M, Guo XS, Wang Y, Lohmueller KE, Rasmussen S, et al. (2011) An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia. Science 333: 94-98.

S9 Demographic Analysis Using *G-PhoCS*

Ilan Gronau[1], Adam H. Freedman[2], Robert K. Wayne[2], John Novembre[2], Adam Siepel[1]

[1]Cornell University,

Department of Biological Statistics and Computational Biology Ithaca, New York, United States of America

[2]University of California, Los Angeles

Department of Ecology and Evolutionary Biology Los Angeles, California, United States of America

S9.1 Overview of *G-PhoCS*

Our main demographic analysis is based on the Generalized Phylogenetic Coalescent Sampler (*G-PhoCS*) developed by Gronau *et al.* [1]. *G-PhoCS* performs demographic inference conditioned on a given population phylogeny augmented by a collection of migration bands (see Fig. S9.1.1). Migration bands describe scenarios of post-divergence gene flow in the demographic model, and are defined by ordered pairs of branches in the population phylogeny, allowing different rates to be associated with the two directions of gene flow. *G-PhoCS* infers demographic parameters associated with the population phylogeny (i.e., ancestral population sizes, population divergence times, and migration rates) based on inferred genealogies at thousands of neutrally evolving loci along the genome. To estimate these genealogies, *G-PhoCS* receives as input a collection of multiple sequence alignments of individual genomes at a given set of genomic loci, selected to reduce the effects of selection and sequencing error (see Section S9.2.1). Each genome in the input set is associated with a certain sampled population (terminal branch of the input phylogeny). *G-PhoCS* can analyze haploid genomes, such as the boxer reference genome (CanFam3), as well as diploid genomes, such as the six genomes sequenced in this study. Heterozygous genotypes are given in an unphased manner, and the likelihood computation analytically sums over all possible phasings.

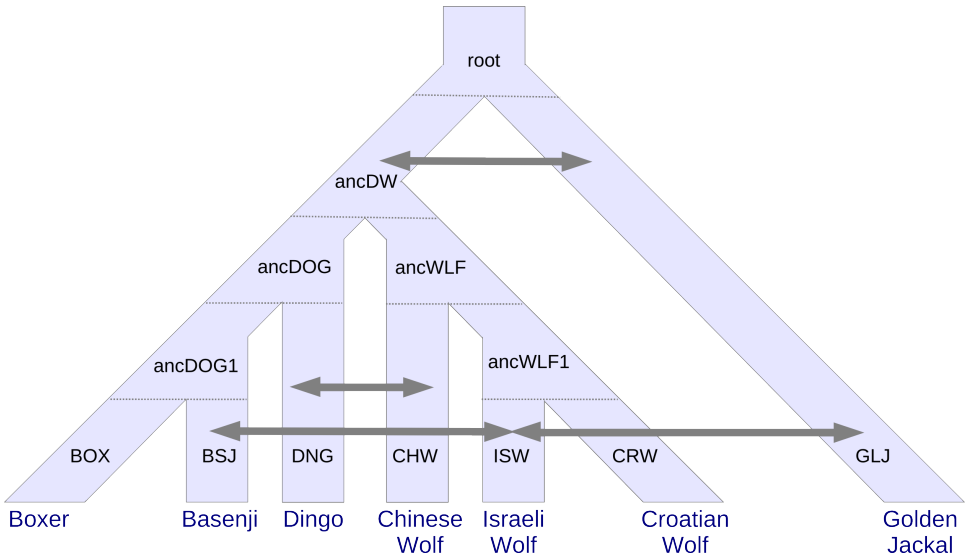


Figure S9.1.1: Population phylogeny assumed in main *G-PhoCS* demographic inference. The six genome samples and the reference genome (boxer) are indicated at the tips of the tree. Each branch in the tree is shown with its population label. The population phylogeny consists of a dog clade, a wolf clade and a jackal outgroup. Within the dog clade, boxer and basenji are assumed to be sister taxa, and within the wolf clade, the Croatian wolf and Israeli wolf are assumed to be sister taxa. The topology of the tree was inferred by Neighbor-Joining using average pairwise genomic divergences (Fig. 4; see also Text S8). Alternative topologies were considered as well (see Section S9.6). Parameters of the demographic model include effective population sizes for all branches in the tree, divergence times for all internal nodes, and migration rates for the migration bands assumed in the analysis. Bidirectional arrows represent the eight migration bands assumed in our main analysis (see Section S9.3).

S9.2 Sequence Data and Analysis Setup

S9.2.1 Alignments at Putative Neutral Regions

We followed a similar procedure to that described by Gronau *et al.* [1] in defining the set of loci on which to run the demographic analysis. We first filtered out regions covered by the genomic filter GF2 (see Text S4), namely, regions in the CanFam3 genome with assembly gaps, repeats, low mappabilty, and regions where none of the six sequenced genomes had reliable sequence data (Table S9.1.1). In addition, we removed regions of the genome that were likely to have evolved under the effect of strong natural selection. In particular, we filtered out exons of protein coding genes and the 10 kilobases (kb) flanking them on each side, as well as conserved non- coding elements (CNEs) and the 100 bases on each side of these elements. CNEs were defined using a conservation track for eleven euarchontoglire mammals computed using the 30-way genome alignment with mouse reference downloaded from the UCSC Genome Browser (see Text S7). Removing flanking regions around genes and CNEs reduces potential biases from selection at linked sites (e.g., background selection and hitchhiking) on our analysis

(see also Section S9.7.3). After filtering, 31.3% of the CanFam3 genome remained, from which we selected 1 kb loci located at least 30 kb apart. We chose a locus length of 1 kb, because it is expected to result in small amounts of intra-locus recombination in the time scale of dog and wolf evolution (see Section S9.7.2). The inter-locus distance of 30 kb was chosen to ensure sufficient inter-locus recombination to reduce the correlation between the local genealogies at different loci.

We identified a collection of 16,434 loci that obey these criteria, and extracted multiple sequence alignments for these loci using sequence data from the six individual genomes in addition to the boxer reference (CanFam3). We further masked each genome individually for positions where there was no confident genotype call (SF filter; see Text S4). In order to avoid biases from hypermutable CpGs, we masked out all position pairs having a "CG" dinucleotide in any of the six genomes or the boxer reference genome sequence [1]. To avoid possible ancestral CpGs, we also masked out position pairs with a C* dinucleotide in one genome and *G in another. Our main set of estimates was obtained by jointly analyzing the full set of 16,434 loci. However, to expedite the supporting analyses presented in this supplement, we used a subset of 5,478 loci obtained by selecting every third locus in the original set.

Table S9.1.1. Data Filters used in *G-PhoCS* analysis.

Filter name	Type	Genome % [a]	Description
mappability	mappability	2.7%	Consecutive pairs of 50 bp blocks with mean mappability score > 2
repeatMasker25	mappability	24.7%	Regions with RepeatMasker score <= 25
refGaps	assembly gaps	5.3%	Sites identified as gaps in the CanFam3 assembly
maskIntersection	missing data	18.6%	Sites with no confident genotype in any of the six sequenced genomes [b]
genesAndFlanks10kb	non-neutral	42.6%	Exons of protein coding genes (see Text S7) and 10kb flanking each exon on each side
phastConsAndFlanks100b	non-neutral	12.5%	phastCons elements computed for eleven euarchontoglire mammals in the the 30-way alignment for the mouse reference, and the 100 bp flanking each element on each side
allFilters		68.7%	Union of all filters

[a] Percent of the CanFam3 genome covered by this filter.
[b] Individual genomes are filtered using the SF filter (see Text S4).

S9.2.2 MCMC Setup for *G-PhoCS*

All MCMC runs were executed using the same setup, unless otherwise indicated. The prior distribution over model parameters was defined by a product of Gamma distributions. We used the default settings chosen by Gronau *et al.* [1]: a Gamma distribution with α=1.0 and β=10,000 for the mutation-scaled population sizes and divergence times, and a Gamma distribution with α=0.002 and β=0.00001 for the

mutation-scaled migration rates. Each Markov Chain was run for 100,000 burn-in iterations, after which parameter values were sampled for 200,000 iterations every 10 iterations, resulting in a total of 20,001 samples from the approximate posterior. Convergence was inspected manually for each run. The finetune parameters of the sampling procedure were set automatically during the first 10,000 burn-in iterations (using the 'find- finetunes TRUE' option in the *G-PhoCS* control file).

S9.2.3 Parameter Calibration

Parameters in the probabilistic model of *G-PhoCS* are scaled by mutation rate μ. Effective population sizes are given by $\theta=4N_e\mu$, and divergence times are given by $\tau=T\mu/g$, where N_e is the absolute effective population size (in number of individuals), g is the average generation time (in years), and T is the absolute divergence time (in years). Following Lindblad-Toh *et al.* (2005), we assumed an average mutation rate of $\mu=1.0\text{x}10^{-8}$ mutations per site per generation, and an average generation time of g=3 years. Throughout this section, we follow the convention of discussing the calibrated estimates (N_e and T) in the text and showing both the raw estimates and calibrated values in figures and tables. For better readability, we scale up the raw estimates (τ and θ) by an additional factor of 10^4, and scale down the calibrated estimates (N_e and T) by a factor of 10^{-3}. The probablistic model of *G-PhoCS* also uses a scaled version of migration rate, $M=m/\mu$, where m is the probability of migration across a given band in a single generation. The level of gene flow across a given migration band is measured by the *total migration rate*, which is the migration rate scaled by the time span of the migration band (τ_m): $m^{tot} = M\tau_m$. If m^{tot} is sufficiently small ($m^{tot}<0.5$), then it approximately equals the probability that a given lineage will migrate through the band. By scaling the rate M with the time span τ_m, we obtain a measure that is independent of our assumptions on mutation rate. The time span of a migration band is defined using the start and end times of the two populations that define it. For example, the time span of the migration band from BSJ to ISW is $\min\{\tau_{ancWLF1}, \tau_{ancDOG1}\}$, and the time span of the migration band from GLJ to the ancestral population ancDW is $\tau_{root} - \tau_{ancDW}$.

S9.3 Inferring Gene Flow

The unique advantage of *G-PhoCS* is its capability to detect and measure gene flow throughout the history of the sampled populations by introducing migration bands to the demographic model. A limitation of this approach is that demographic models with large numbers of migration bands often have identifiability issues that can lead to spurious inference of migration events. To address the challenge of detecting the significant signals of gene flow in the data, we followed a strategy of examining a large number of

migration bands by partitioning them across seven separate *G-PhoCS* analyses. Each of these separate analyses was conducted on the set of 5,478 neutral loci described in Section S9.2.1 using the settings described in Section S9.2.2. A migration band was inferred to have significant gene flow if the 95% Bayesian credible interval of the total migration rate for that band did not include 0, or if the total migration rate was estimated to be greater than 0.03 with posterior probability greater than 50%. We used this somewhat lax criterion for significance to ensure that we accounted for all scenarios of gene flow that have some support in the data. We then executed an additional *G-PhoCS* analysis incorporating all migration bands with significant gene flow, as well as migration bands in the opposite direction.

S9.3.1 Identifying Migration Bands with Significant Gene Flow

First, we examined gene flow between dogs and wolves by considering the 18 directional migration bands between one of the three sampled dog populations (BSJ, BOX, and DNG) and one of the three sampled wolf populations (ISW, CRW, and CHW). We conducted six separate analyses labeled according to the six sampled dog and wolf populations: the analysis labeled by population X contained the six migration bands that contain population X. Note that each of the 18 migration bands is covered in two separate *G-PhoCS* runs: the run labeled by the dog population in that band, and the run labeled by the wolf population. Thus, for each of the nine dog-wolf pairs, we recorded four migration intensities: two for the dog-to-wolf migration band, and two for the band in the opposite direction (Fig. S9.3.1A). Significant gene flow was inferred for the two migration bands between ISW and BSJ and the migration band DNG-to-CHW, consistently in both runs that included each of these migration bands. The migration band BOX- to-ISW was inferred to have a significant total rate of 0.1 (0.045–0.155) in the 'BOX' analysis, but not in the 'ISW' analysis. This observation is consistent with gene flow from BSJ to ISW, which, in the absence of a migration band between BSJ and ISW, is likely to be inferred as gene flow from BOX to ISW. Migration bands CHW-to-DNG, and ISW-to-DNG were inferred to have nonnegligible (but insignificant) total rates of 0.021 (0–0.055) and 0.023 (0–0.058) (resp.) in one of the runs that contained each of them. We conclude that significant gene flow occurred between Israeli wolf and basenji (in both directions), and from dingo to Chinese wolf. Note that our findings are consistent with the non-parametric ABBA/BABA tests for gene flow (see Text S8), but the ability to consider several migration bands in a single analysis allowed us to explain the positive ABBA/BABA signal observed for boxer and Israeli wolf as a result of gene flow from basenji to Israeli wolf.

Using the migration model of *G-PhoCS*, we were also able to model gene flow between the jackal outgroup and each of the other six samples. We conducted another analysis

with 14 additional directional migration bands: twelve between GLJ and the other six sampled populations, and two between GLJ and the population, ancDW, ancestral to all dogs and wolves (Fig. S9.3.1B). We inferred a very high total migration rate of 1.02 (0.89–1.14) for the ancDW- to-GLJ migration band, and a smaller, but significant total rate of 0.033 (0.018–0.049) for the ISW-to-GLJ migration band.

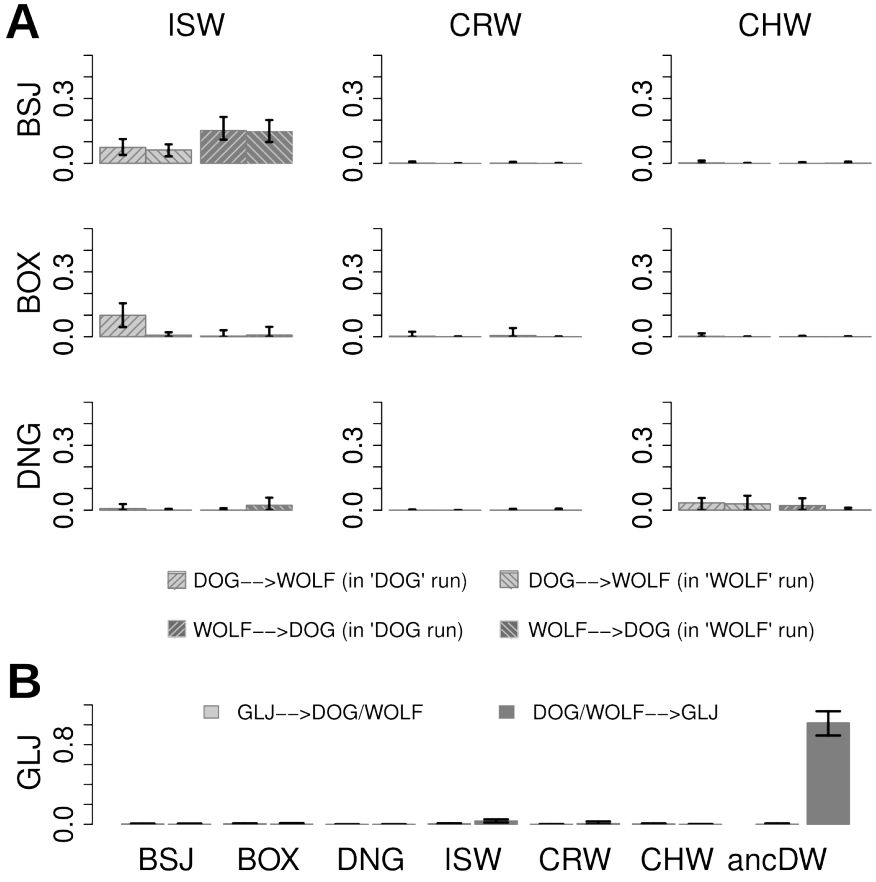


Figure S9.3.1: Identifying Migration Bands with Significant Gene Flow. A collection of 32 migration bands was examined in seven separate *G-PhoCS* analyses of the set of 5,478 neutral loci defined in Section S9.2.1 using the default MCMC settings described in Section S9.2.2. (A) Total migration rates estimated for the 18 migration bands between the sampled populations of dogs and wolves are shown with 95% Bayesian credible intervals. For each pair of dog and wolf sampled population, the left pair of bars corresponds to the DOG-to-WOLF migration band, and the right pair corresponds to the band in the opposite direction. For each migration band, the left bar indicates the total rate inferred in the analysis containing bands associated with the dog population, and the right bar corresponds to the analysis associated with the wolf population. We find significant evidence for gene flow along migration bands BSJ-to-ISW, ISW-to-BSJ, and DNG-to-CHW (see text). (B) Total migration rates inferred for 14 migration bands with GLJ. For each of the seven populations considered, rates are shown for the migration band from GLJ to that population (left) and the band in the opposite direction (right). We find significant evidence for gene flow along migration bands ISW-to-GLJ and ancDW-to-GLJ.

S9.3.2 The effect of Gene Flow on Parameter Estimates

We found evidence for significant gene flow between four pairs of populations in our

demographic model: (ISW,BSJ), (CHW,DNG), (GLJ,ISW), and (GLJ,ancDW). For all pairs other than (ISW,BSJ), significant gene flow was inferred only in one direction. However, to ensure we account for all plausible scenarios of gene flow, we kept all eight directional migration bands associated with these four pairs in our subsequent analysis. In order to test the effect of gene flow on estimates of population divergence times and effective population sizes, we compared between sets of estimates obtained in four additional analyses: an analysis without any migration band, an analysis with the four bands corresponding to (ISW,BSJ) and (CHW,DNG) population pairs, an analysis with the four bands corresponding to (GLJ,ISW) and (GLJ,ancDW) population pairs, and an analysis with all eight bands. The four sets of parameter estimates are presented in Figure S9.3.2. Modeling gene flow with the golden jackal reduced the estimated effective size for the ancestral root population (N_{root}) from 41,000 to 17,000, and the effective size of the population ancestral to dogs and wolves (N_{ancDW}) from 47,000 to 45,000. The divergence times associated with these ancestral populations consequently increased from 163 thousand years ago (kya) to 415 kya (T_{root}) and from 11.7 kya to 13.1 kya (T_{ancDW}). Modeling gene flow between dogs and wolves had no significant effect on the ancestral effective population sizes, but it did result in an increase in the estimate of the dog-wolf divergence time (T_{ancDW}=13.6 kya). Our full model of gene flow with eight migration bands resulted in further increase of this divergence time to 14.9 kya.

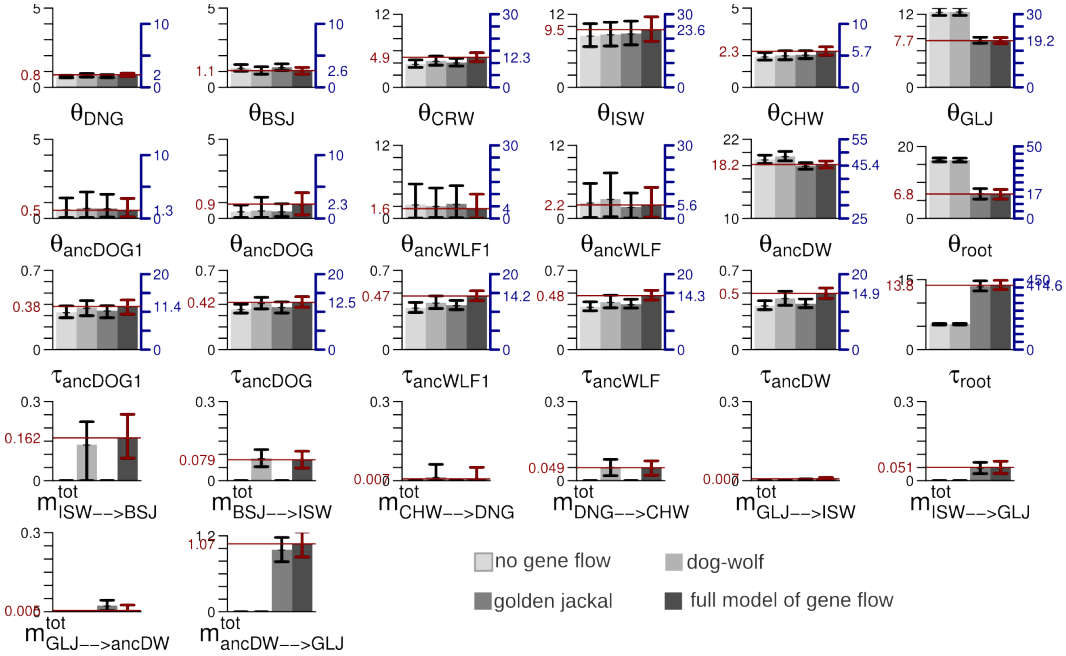


Figure S9.3.2: Parameter estimates under different scenarios of gene flow. Estimates and 95% Bayesian credible intervals for the 26 demographic parameters were obtained assuming four different scenarios of gene flow (left to right): (1) no gene flow; (2) gene flow between populations (ISW,BSJ) and between populations (CHW,DNG); (3) gene flow

between populations (GLJ,ISW) and between populations (GLJ,ancDW); and (4) gene flow along all eight migration bands (highlighted in red). All four analyses were conducted on the set of 5,478 loci defined in Section S9.2.1 with the MCMC settings as described in Section S9.2.2. Raw estimates, scaled by mutation rate ($x10^4$), are shown (left axis) next to calibrated estimate (right axis). Calibrated divergence times are given in 1,000 years and calibrated population sizes are given in thousands of individuals (see Section S9.2.3 for details).

S9.4 Main Set of Estimates for All Demographic Parameters

The main set of parameter estimates reported in our study is based on a single *G-PhoCS* analysis of the 16,434 neutral loci defined in Section S9.2.1, assuming the population phylogeny with eight migration bands shown in Fig. S9.1.1. Parameter estimates are described in Supplementary Table S12. See Section S9.2.3 for details on calibration of the raw parameter estimates. In the following sections we validate the robustness of this inferred demographic model to various factors:

. In Section S9.5 we compare the demographic model inferred by *G-PhoCS* to the one implied by the ancestral effective population sizes inferred by the pairwise sequentially Markovian coalescent (PSMC) method of Li and Durbin [2] (see Text S8).

. In Section S9.6 we examine several other plausible topologies for the population phylogeny associated with alternative hypotheses for dog domestication.

. In Section S9.7 we demonstrate the robustness of our estimates to assumptions made in the construction of the collection of neutral loci we used in the analysis.

S9.5 Comparison with Estimates from PSMC Analysis

The demographic history of dogs and wolves as inferred by *G-PhoCS* is fairly consistent with the history inferred by separately analyzing the six diploid genomes using the pairwise sequentially Markovian coalescent (PSMC) method of Li and Durbin [2] (see Text S8). Both analyses infer similar ancestral population sizes, with a parallel decline in sizes observed for dogs as well as wolves. However, whereas *G-PhoCS* infers that dogs and wolves diverged roughly 15 kya, the ancestral effective population sizes inferred from the two dog genomes by PSMC diverge from those inferred from the three wolf genomes at a time point roughly 40-50 kya. Li and Durbin note that their method is likely to interpret abrupt changes in population sizes as gradual changes that started earlier in time. Thus, if dogs and wolves experienced strong population bottlenecks, their inferred ancestral sizes would appear to diverge before the ancestral populations diverged. We confirmed this observation by showing that PSMC produces a similar pattern of early

divergence when run on data simulated according to the demographic model inferred by *G- PhoCS* (Supplementary Fig. S2; see also subsection S8.2.2 in Text S8).

As additional validation of the more recent divergence inferred by *G-PhoCS*, we conducted the reciprocal experiment in which *G-PhoCS* was run on data simulated according to a demographic model implied by the PSMC estimates. In these simulations, we assumed the population phylogeny inferred by neighbor joining (Fig. 4) without the boxer population (since the haploid boxer genome was not analyzed by PSMC). Divergence times (in years) were set to T_{ancDOG} = 13,000, T_{ancWLF} = $T_{ancWLF1}$ = 42,800, and T_{ancDW} = 47,500, according to approximate times

associated with divergence of the ancestral effective population sizes inferred by PSMC. We simulated gradual change in effective population size, as inferred by PSMC; for the current populations BSJ, DNG, ISW, CRW, CHW, and GLJ we used ancestral sizes inferred for the appropriate genome, for the ancestral population ancDOG we used ancestral sizes inferred from the basenji genome, and for the ancestral populations ancWLF1, ancWLF, ancDW, and root we used ancestral sizes inferred from the genome of the Israeli wolf. All parameters were scaled assuming an average mutation rate of 1.0×10^{-8} mutations per site per generation, and an average generation time of 3 years (see Section S9.2.3).

In order to examine the potential effects of intra-locus recombination on our estimates, we simulated data under three levels of recombination: r = 0.0 cM/Mb, r = 0.25cM/Mb, and r = 0.92 cM/Mb. The lower recombination rate (r = 0.25 cM/Mb) was based on the estimate from the PSMC analysis (see Text S8), and the higher rate (r = 0.92 cM/Mb) was based on the mean recombination rate estimated in the dog genome from a linkage map generated using microsatellites [3]. We generated four replicate data sets for each recombination rate using the MS simulation software [4] , each with 5,000 alignments of length 1 kb, and ran *G-PhoCS* on these data sets using the same settings as in our main analysis (including migration bands). Estimates of divergence times were highly concordant with the values used in generation of the data across the twelve data sets, regardless of recombination. Recombination appears mostly to influence the estimates for the effective population size and divregence time at the root (N_{root} and

T_{root}), due to the recombination events that occurred since divergence from golden jackal.

The parameter estimates obtained on these 4x3=12 simulated data sets are described in Supplementary Figure S3. This experiment shows that *G-PhoCS* accurately infers

population divergence times in demographic histories with gradual changes in ancestral population sizes, even in the presence of a small amount of intra-locus recombination. Because the divergence times *G-PhoCS* infered from real data were very different from the ones it infered from data simulated under the PSMC-based model, we conclude that the PSMC-based model with deep divergence does not fit the data. Additionally, the reciprocal experiment where PSMC was run on data generated according to the demographic model inferred by *G-PhoCS* (Supplementary Fig. S2) suggests that the deep divergences observed in the PSMC estimates are consistent with the model inferred by *G-PhoCS.*

S9.6 Alternative Topologies of the Population Phylogeny

Our demographic analysis is conditioned on a given topology for the population phylogeny. In our main analysis, we assumed the topology of the neighbor joining tree (Fig. 4). This tree describes dogs and wolves as evolving in two separate clades. We examined plausible alternative topologies in two series of analysis, to ensure that our estimates were not strongly affected by our assumptions on the tree topology.

S9.6.1 Regional Origin

One alternative scenario for the joint history of dogs and wolves is that dogs were domesticated separately in different geographic regions. To test this hypothesis, we considered three alternative topologies for the population phylogeny, in which each geographic region–Middle East (MEA), East Asia (EAS), and Europe (EUR)–corresponds to an ancestral population with two daughter populations: dog and wolf (Supplementary Fig. S4A). Each of the three alternative topologies is determined by the order of geographic divergence events. We conducted demographic inference conditioned on each of these three topologies, once assuming no gene flow between populations, and once with 16 migration bands: all bands between sampled dog populations, all bands between sampled wolf populations, and bands between GLJ and ISW and the population ancestral to all dogs and wolves (ancDW).

When no post-divergence gene flow is allowed in the model, the estimated divergence times decrease to levels lower than our original estimate of the divergence between bansenji and boxer (Supplementary Fig. S4B; T_{ancDW} = 9,000 (8,600-10,200) across the three runs). This likely

reflects poor fit of these models to the data, as a consequence of the similarity between the dog genomes. When we introduced post-divergence gene flow between dogs and between wolves into the model, the estimated divergence times increase significantly.

However, migration rates were estimated to be very high, with total rates near 1.0 for the BSJ-to-BOX migration band, and total rates near 0.5 for the BSJ-to-DNG migration band. We conclude that in order to accommodate a hypothesis of separate regional domestication of dogs, there had to have been very high levels of post-divergence gene flow between dog (and wolf) populations from different geographic regions. This is in contrast to our default model with separate clades for dogs and wolves, which can be fit to data with considerably less post-divergence gene flow.

S9.6.2 Alternative Origins for the Dog Clade

Another alternative is that dogs were domesticated once, and thus form a distinct clade in the phylogeny, but the origin of domestication is not the population ancestral to all wolves. Assuming the topology of the wolf subphylogney is ((ISW,CRW),CHW), there are five possible origins for the dog clade, corresponding to the five branches of that phylogeny (Supplementary Fig. S5A). We conducted demographic inference conditioned on each of the four alternative topologies with the eight migration bands assumed in our original analysis. Overall, estimates of all parameters were very similar to our original estimates (Supplementary Fig. S5B). In all five analyses, the difference between the three divergence times T_{ancWLF1}, T_{ancWLF}, and T_{ancDW} were very small, but they were markedly higher when the original topology was assumed: $|\Delta\tau| = |\tau_{\text{ancDW}} - \tau_{\text{ancWLF}}|$ = 597 years (42–1,416) in our original analysis compared to $|\Delta\tau|$= 81 years (0–

643) across the other four analyses. We conclude that the data does not significantly support a particular origin for dogs, but regardless of our assumptions on the identity of the ancestral lineage from which dogs were domesticated, this lineage diverged from other wolf lineages considered in this study at roughly the same time they diverged from each other (14–15 kya).

S9.7 Alternative Sets of Neutral Loci

The parameter estimates obtained by *G-PhoCS* depend on the collection of neutral loci used in the analysis (see Section S9.2.1). Certain assumptions made in the construction of these loci determined locus length, distance from coding exons, and even random subsetting, all of which can potentially influence the resulting estimates. *G-PhoCS* has been shown by Gronau *et al.* [1] to be robust to these factors in the analysis of individual human genomes. In this section we present similar validation experiments conducted on the individual canid genomes analyzed in this study.

S9.7.1 Subsetting of Loci

We compared parameter values inferred for the full set of 16,434 loci to values inferred for each of three disjoint equally-sized subsets of that set, obtained by selecting every third locus in the original set. Estimates of all parameters show high levels of agreement across these four analyses (Fig. S9.7.1). As expected, Bayesian credible intervals were smaller when all 16,434 loci were analyzed.

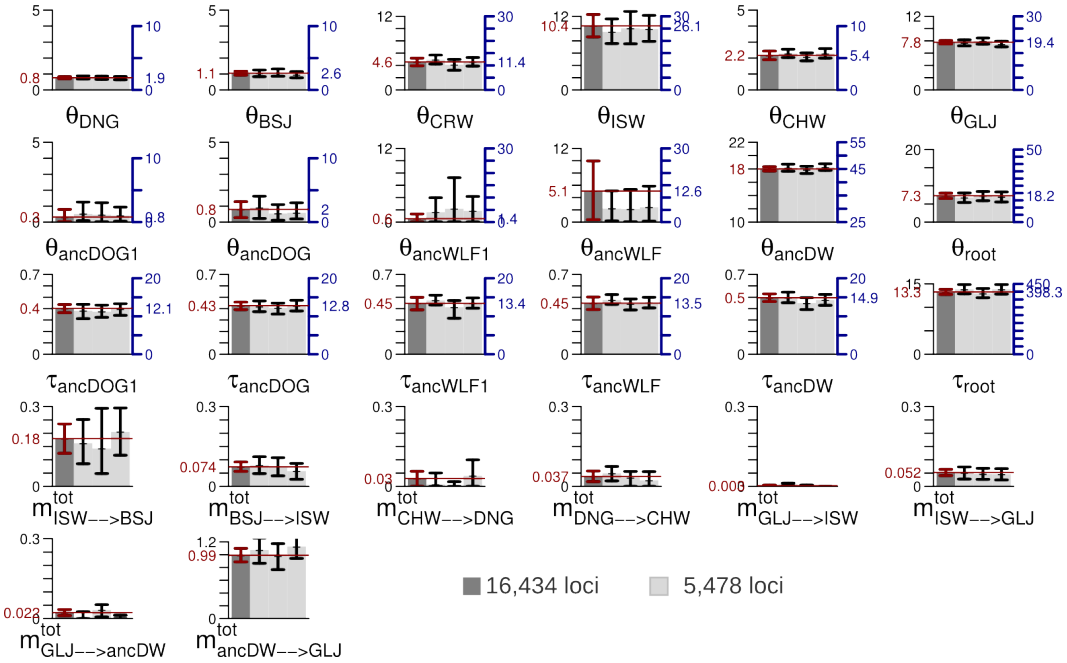


Figure S9.7.1: Parameter estimates for different sets of neutral loci. Estimates and 95% Bayesian credible intervals for the 26 demographic parameters were obtained using four different sets of neutral loci: the full set of 16,434 loci (dark gray; see Section S9.2.1), and three equally-sized disjoint subsets of that set (light gray) obtained by selecting every third locus in the original set. Raw estimates, scaled by mutation rate (x10^4), are shown (left axis) next to calibrated estimate (right axis) (see Section S9.2.3 for details on calibration).

S9.7.2 Locus Length and Intra-locus Recombination

A locus size of 1 kb was chosen for our main analysis in order to ensure small amounts of intra- locus recombination, while maintaining a reasonable number of informative sites within each locus. In order to validate the robustness of our parameter estimates for the potential effects of intra-locus recombination, we redid the analysis for different sets of loci with different lengths. To this end, we computed a set of 7,297 neutral loci, 2 kb long, from our collection of filtered neutral sites (see Table S9.1.1) with an inter-locus distance of at least 30 kb. By partitioning each locus in this set to two non-overlapping blocks of size 1kb, we constructed two non-overlapping collections of 1 kb loci, and by further partitioning each 1 kb locus to two 500 bp blocks, we constructed four collections of 500 bp loci.

We analyzed each of these seven different collections of 7,297 loci using *G-PhoCS* with the population phylogeny shown in Fig. S9.1.1, including eight migration bands. Overall, estimates obtained from loci of length 1 kb were very similar to the ones obtained from the shorter 500 bp loci (Fig. S9.7.2). Importantly, estimates of migration rates along the eight migration bands did not appear to be substantially affected by locus length. Recombination events that occurred since divergence of dogs and wolves within the analyzed loci would tend to increase the estimated divergence time (T_{ancDW}). However, estimates of T_{ancDW} obtained from the 1 kb loci and 500 bp loci were highly concordant with our original estimate of T_{ancDW}= 14.9 kya (13.9–15.9 kya). On the other hand, the estimate obtained from the collection of 2 kb loci increased to 21 kya (19–23 kya), most likely owing to a substantial increase in the number of intra-locus recombination events in these longer loci.

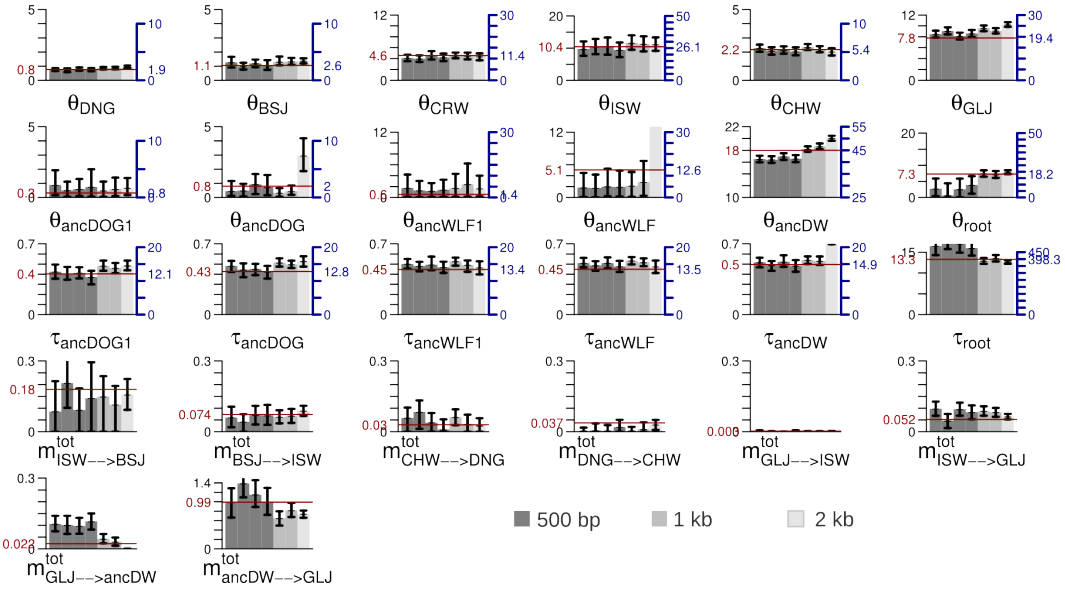


Figure S9.7.2: Effect of intra-locus recombination on parameter estimates. Estimates and 95% Bayesian credible intervals for the 26 demographic parameters were obtained using seven different sets of neutral loci at different lengths: 500 bp, 1 kb and 2 kb. Each data set contains 7,297 neutral loci. The horizontal red line marks the estimate obtained by our main analysis of 16,434 loci of length 1 kb. Raw estimates, scaled by mutation rate (x10^4), are shown (left axis) next to calibrated estimate (right axis) (see Section S9.2.3 for details on calibration).

S9.7.3 Distance from Coding Exons and Effect of Selection at Linked Sites

Another factor that could potentially affect our estimates is natural selection acting on linked sites (e.g., background selection or hitchhiking), which is known to reduce levels of genomic diversity around genes [5]. For this reason, we chose our neutral loci in regions that are located at least 10 kb away from the closest gene. In order to ensure that this approach was sufficiently conservative, we computed alternative sets of loci using different thresholds for this distance: 1, 2, 5, 20, 50, and 100 kb. We applied the

same pipeline described in Section S9.2.1 to compute the alternative sets of loci (using alternative thresholds for distances to genes). We subsampled a collection of 5,478 loci from each set, to match the number of loci in our original analysis, and ran *G-PhoCS* on each of these six alternative data sets (Fig. S9.7.3). None of the parameters showed a strong trend in estimated values as a function of distance from genes, implying that our analysis is not sensitive to selection at linked sites.

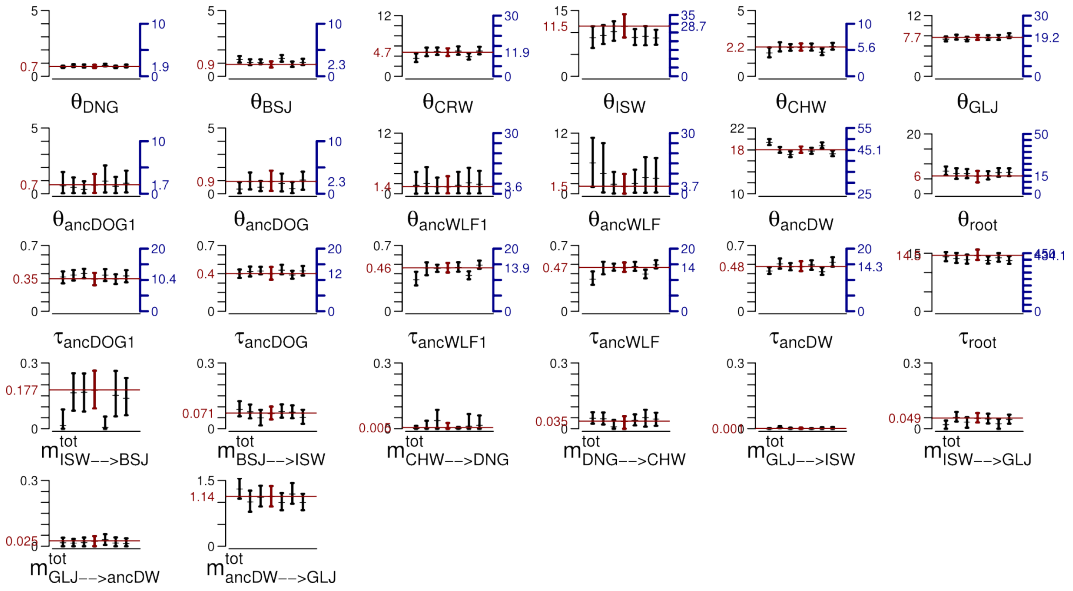


Figure S9.7.3: Effect of distance from genes on parameter estimates. Estimates and 95% Bayesian credible intervals for the 26 demographic parameters were obtained using seven different sets of neutral loci computed using different thresholds for distance from coding exons (left to right; in kb): 1, 2, 5, 10, 20, 50, 100. Each data set contains 5,478 loci of length 1 kb. The horizontal red line marks the estimate obtained using the default threshold of 10 kb. Raw estimates, scaled by mutation rate (x10^4), are shown (left axis) next to calibrated estimate (right axis) (see Section S9.2.3 for details on calibration).

References

1. Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A (2011) Bayesian inference of ancient human demography from individual genome sequences. Nat Genet 43(10):1031–1034.

2. Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature 475(7357):493–496.

3. Wong AK, Ruhe AL, Dumont BL, Robertson KR, Guerrero G, et al. (2010) A comprehensive linkage map of the dog genome. Genetics 184(2):595–605..

4. Hudson R (2002) Generating samples under a Wright-Fisher neutral model of genetic

variation. Bioinformatics 18(2) 337–338.

5. McVicker G, Gordon D, Davis C, Green P (2009) Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet 5(5)

S10 Copy Number Status of the Amylase Gene (*AMY2B*) on CFA6 in 12 Dog Breeds

Holly Beale[1], Elaine Ostrander[1]

[1]National Institutes of Health

Cancer Genetics Branch National Human Genome Research Institute Bethesda, Maryland, United States of America.

Copy number at the variant near AMY2B on Chr6 was measured in whole genome sequencing of 12 dogs, each belonging to a different breed, in a manner similar to the measurement in the wild canids. Specifically, a 7Kb CNV was identified with the approximate CanFam 3.1 coordinates 46,948,800- 46,956,325. Two of the exons of the human AMY2B transcript ENST00000361355 are syntenic to the variant (Table S10.1). Depth plots suggest that the copy is not continuous across the variant, and consequently a subregion spanning the syntenic exons was selected for the purposes of measuring copy number (Figure S6). Copy number was calculated as the average fold increase of sequence depth in the selected region divided by the average read depth (5.9-11.4x) in the surrounding 1 Mb. The expected average value for diploid single-copy regions is approximately 2.

The copy counts are consistent with the AMY2B results reported in Axelsson et al. [1]. All dogs have a duplication (Figure S6). The lowest copy number in these 12 samples is 3 or 4 copies (Husky). The copy number in other dogs ranges from 8 (Beagle) to 30 (Saluki).

Table S10.1. Syntenic positions of Human *AMY2B* exons on Chr6 (CanFam3.1 coordinates).

ID	Start	End
ENSE00001940684	4,6954,362	46,954,570
ENSE00001712096	46,955,754	46,955,872

References

1. Axelsson E, Ratnakumar A, Arendt M-J, Maqbool K, Webster MT, et al. (2013) The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature 495: 360-364.

S11 Comparison of Golden Jackal Sample to Jackals and Wolves

***Carles Vilà*[1]*, Robert K Wayne*[2]

[1]***Estación Biológica de Doñana***

Department of Integrative Ecology Sevilla, Spain

[2]***University of California, Los Angeles***

Department of Ecology and Evolutionary Biology Los Angeles, California, United States of America

Recent reports have indicated that some North African golden jackals possess gray wolf-like mitochondrial DNA sequences [1,2]. The authors have suggested that those individuals may represent a previously undescribed gray wolf lineage present in North Africa, thus supporting previous claims based on the usual morphology and size of some canid specimens from this region [3]. Conceivably, the recent divergence estimated with *G-PhoCS* between golden jackals and gray wolves/dogs compared to previous estimates [4] could be due to a presence of wolf-like individuals in Israel rather than the more basal golden jackal lineage [5]. However, the mitochondrial DNA of the sequenced jackal was not wolf-like (unpublished results) and critically, analysis of previously published SNP data suggest Israeli golden jackals, including the specimen used in this paper, represent a lineage distinct from those in North Africa (Figure S11.1). Also, an analysis of 25 exonic sequences (16,180 bp) across a large panel of over 150 jackals from outside Africa did not indicate close proximity to the gray wolf / dog clade (Koepfli *et al.* unpublished data).

The divergence time of wolves and golden jackals has been estimated as about 1.5 million years from fossil data [4]. The molecular data presented here and results in preparation suggest a much more recent divergence time, indicating that fossil remains thought to be golden jackals are not directly ancestral to living forms. The difficult of deducing affinities of a generalized canid such as the golden jackal are exemplified by the problems in assigning recent specimens to one species or another [3].

Methods

Allele Sharing Distance/Neighbor Joining Tree Reconstruction A subset of golden jackal samples from Kenya and Israel were previously genotyped using the Affymetrix Canine Mapping SNP Array 2.0 (127K SNPs) [6,7] and Illumina CanineHD array (170K SNPs),

along with a number of other species in the *Canis* genus. To provide additional resolution and insight into the phylogenetic relationships of the African and Middle Eastern golden jackals with other canids, a neighbor-joining (NJ) tree based on allele-sharing distances was generated with a subset of SNPs from a panel of two Kenyan golden jackals, two Israeli golden jackals representing both mitochondrial haplotype clades (Cau-like and Clu-like), two ancient and two modern dog breeds, gray wolves from the Middle East, Asia, Europe and North America, Coyotes (*C. latrans*), Ethiopian wolves (*C. simensis*), and a Black-backed jackal (*C. adjustus*) and Side-striped jackal (*C. mesomelas*). We applied allele conversions to a subset of 14,695 nuclear SNPs distributed across the genome to generate a combined dataset, based on a comparative analysis of the 52,329 SNPs that overlap between the two different arrays (kindly provided by Suiyuan Zhang of NHGRI).

The allele-sharing distance was calculated as one minus the proportion of alleles shared for the 14,695 SNP panel, using the program MICROSAT [8] with 1,000 bootstrap replicates. Neighbor joining trees for each replicate of the resulting pairwise matrices of allele sharing distance were calculated using the program NEIGHBOR and a consensus tree was generated using the majority rule option in the program CONSENSE, both from the PHYLIP package [9]. The resulting tree was visualized as unrooted (Figure S11.1) using DENDROSCOPE [10].

References

1. Rueness EK, Asmyhr MG, Sillero-Zubiri C, Macdonald DW, Bekele A, et al. (2011) The Cryptic African Wolf: Canis aureus lupaster Is Not a Golden Jackal and Is Not Endemic to Egypt. PLoS ONE 6: e16385.

2. Gaubert P, Bloch C, Benyacoub S, Abdelhamid A, Pagani P, et al. (2012) Reviving the African Wolf Canis lupus lupaster in North and West Africa: A Mitochondrial Lineage Ranging More than 6,000 km Wide. PLoS ONE 7: e42740.

3. Ferguson WW (1981) The Systematic Position of Canis-Aureus-Lupaster (Carnivora, Canidae) and the Occurrence of Canis-Lupus in North-Africa, Egypt and Sinai. Mammalia 45: 459-465.

4. Perini FA, Russo CAM, Schrago CG (2010) The evolution of South American endemic canids: a history of rapid diversification and morphological parallelism. J Evol Biol 23: 311-322.

5. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, et al. (2005)

Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438: 803-819.

6. vonHoldt BM, Pollinger JP, Earl DA, Knowles JC, Boyko AR, et al. (2011) A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. Genome Res 21: 1294-1305.

7. vonHoldt BM, Pollinger JP, Lohmueller KE, Han EJ, Parker HG, et al. (2010) Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. Nature 464: 898-902.

8. Minch E, Ruiz-Linares A, Goldstein D, Feldman M, Cavalli-Sforza L (2009) MICROSAT: A computer program for calculating various statistics on microsatellite allele data. Palo Alto, CA: Stanford University. pp. MICROSAT: A computer program for calculating various statistics on microsatellite allele data.

9. Felsenstein J (1993) PHYLIP (Phylogency Inference Package) version 3.5c. Seattle, WA: Department of Genetics, University of Washington.

10. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, et al. (2007) Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics 8: 460.

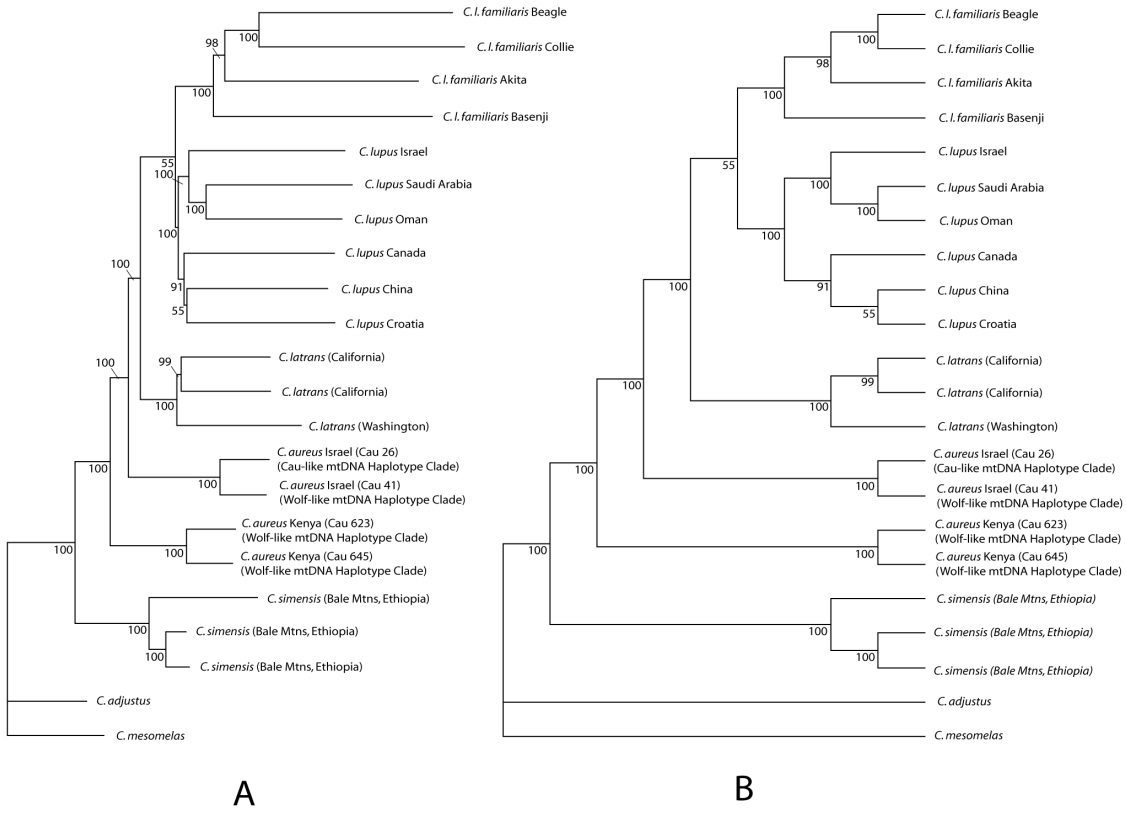


Figure S11.1. Neighbor-joining allele-sharing distance based phylogram (A) and cladogram (B) utilizing the 14,695 SNP dataset from canine SNP array assays. Bootstrap support based on 1,000 replicates is shown. Sample numbers and mtDNA haplotype clades for golden jackals are indicated for comparison with other analysis results.

Supplementary Material for Paper IV - Wordlwide Patterns of Genomic Variation and Admixture in Gray Wolves

(formatted as published)

Supplementary Figures

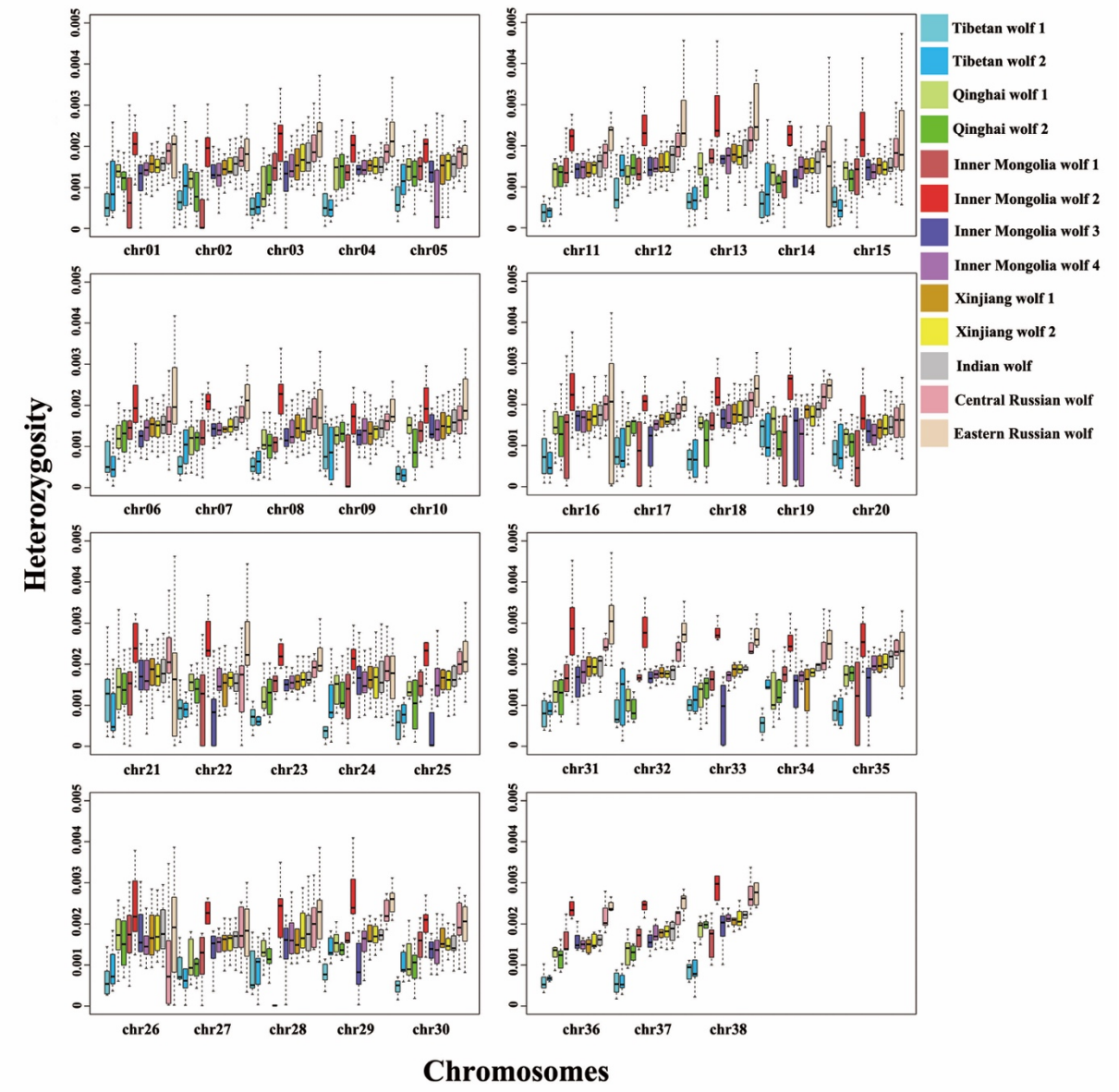


Figure S1. The box plot of heterozygosity from 5 mb non-overlapping windows across all the 38 autosomes. These are all the Asian wolves.

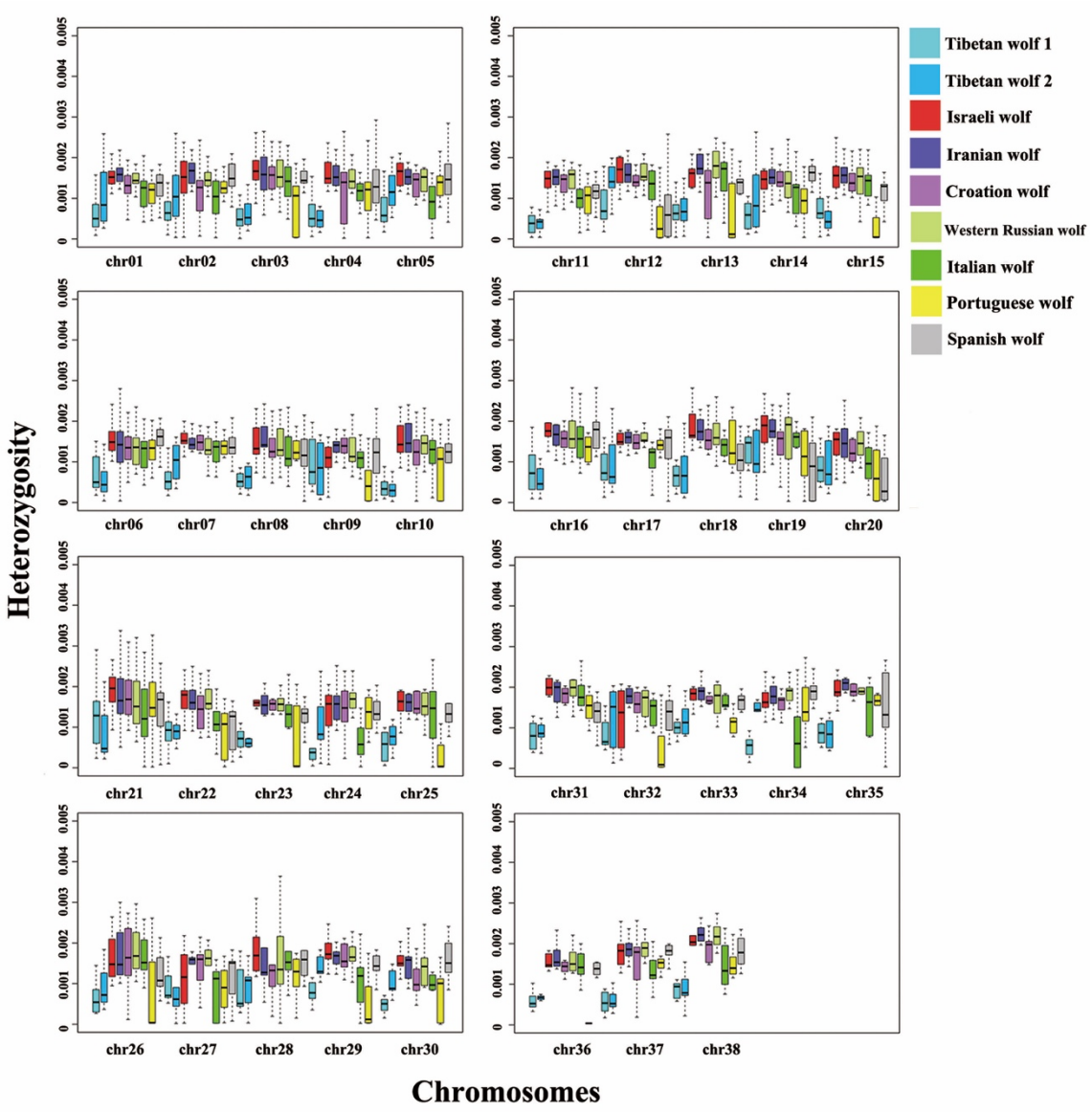


Figure S2. The box plot of heterozygosity from 5 mb non-overlapping windows across all the 38 autosomes. These are all the European and Middle East wolves, and two Tibetan wolves.

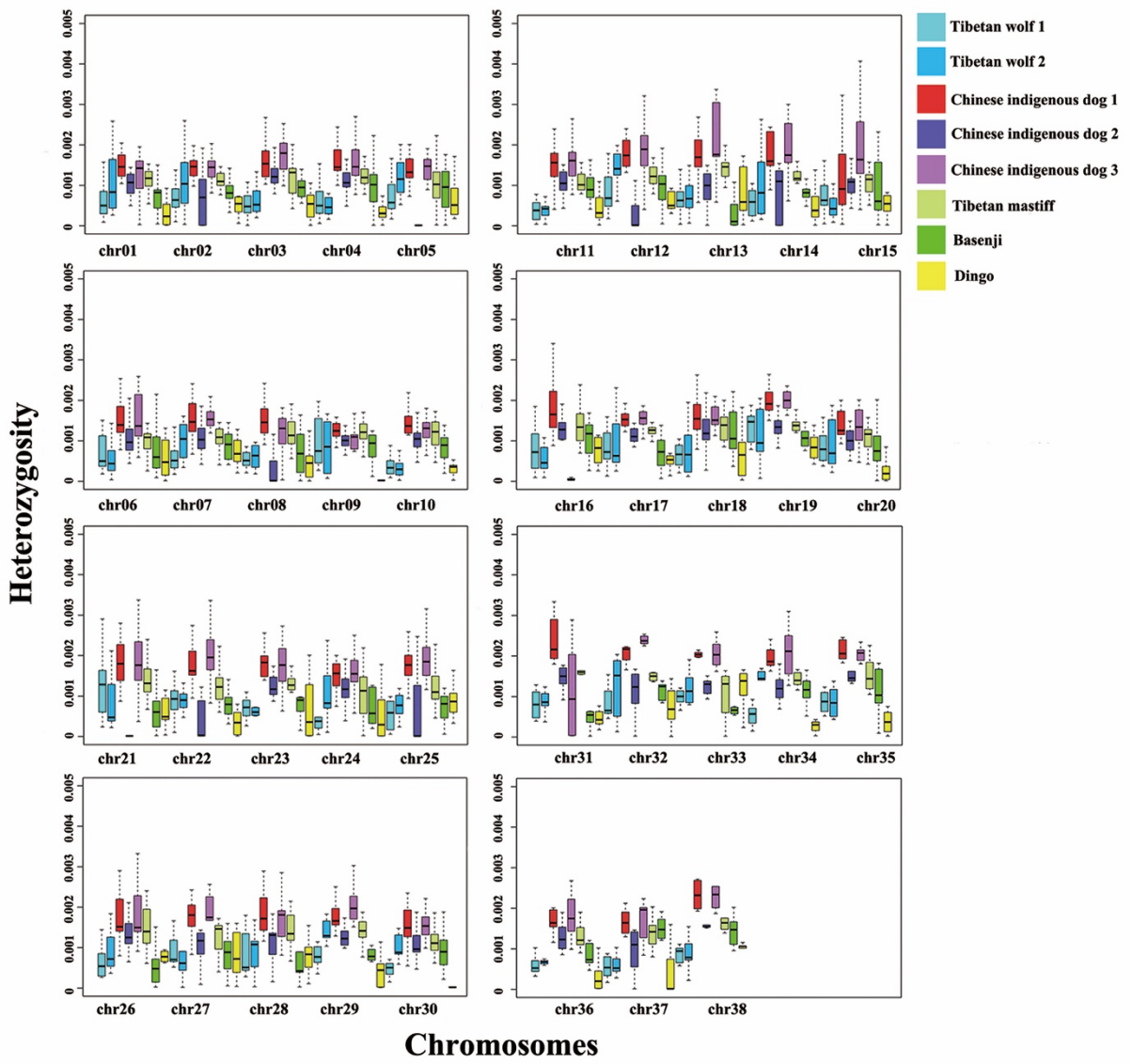


Figure S3. The box plot of heterozygosity from 5 mb non-overlapping windows across all the 38 autosomes. These are all the dogs and two Tibetan wolves.

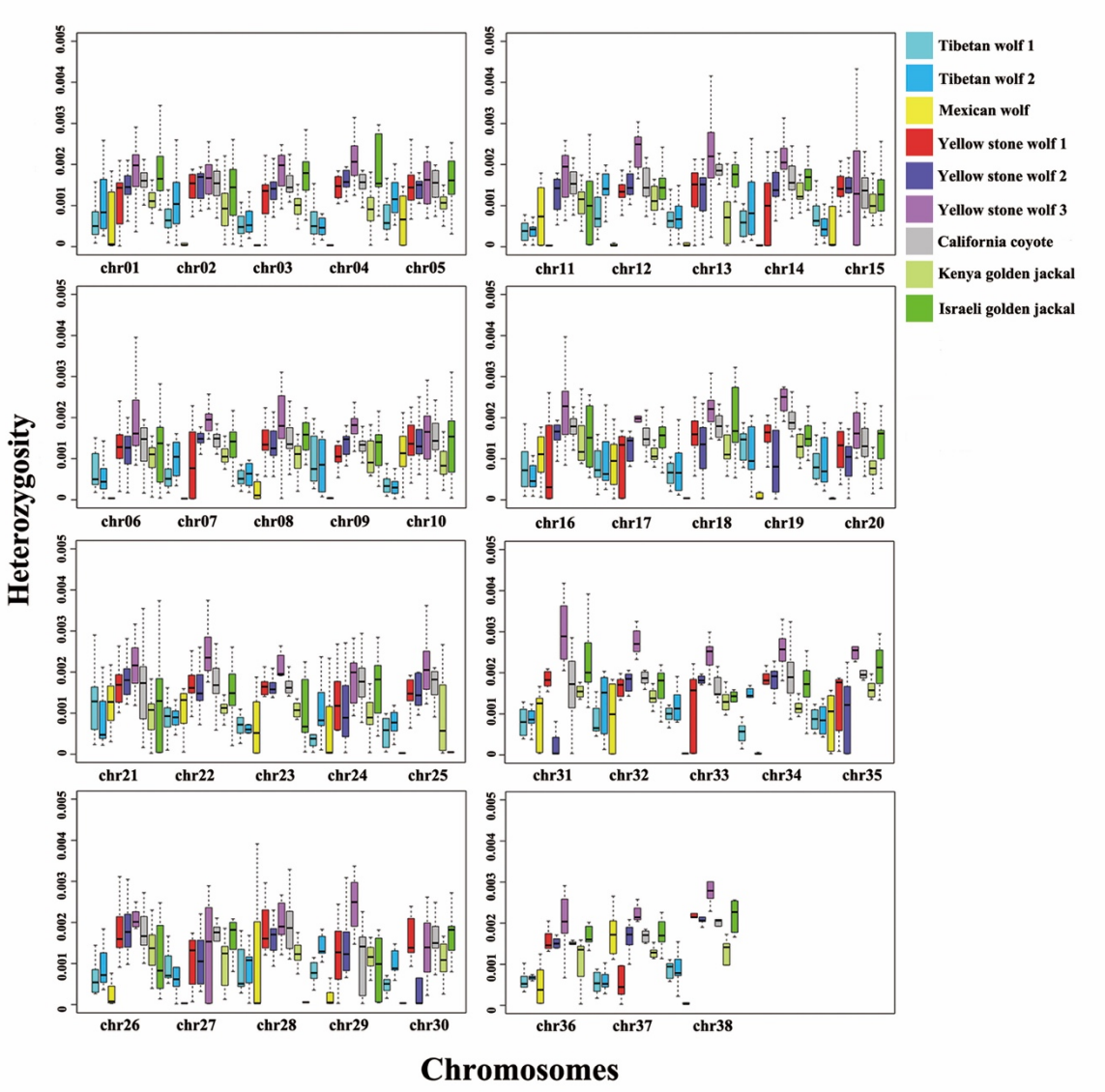


Figure S4. The box plot of heterozygosity from 5 mb non-overlapping windows across all the 38 autosomes. These are all the North American wolves, two Tibetan wolves and three outgroups.

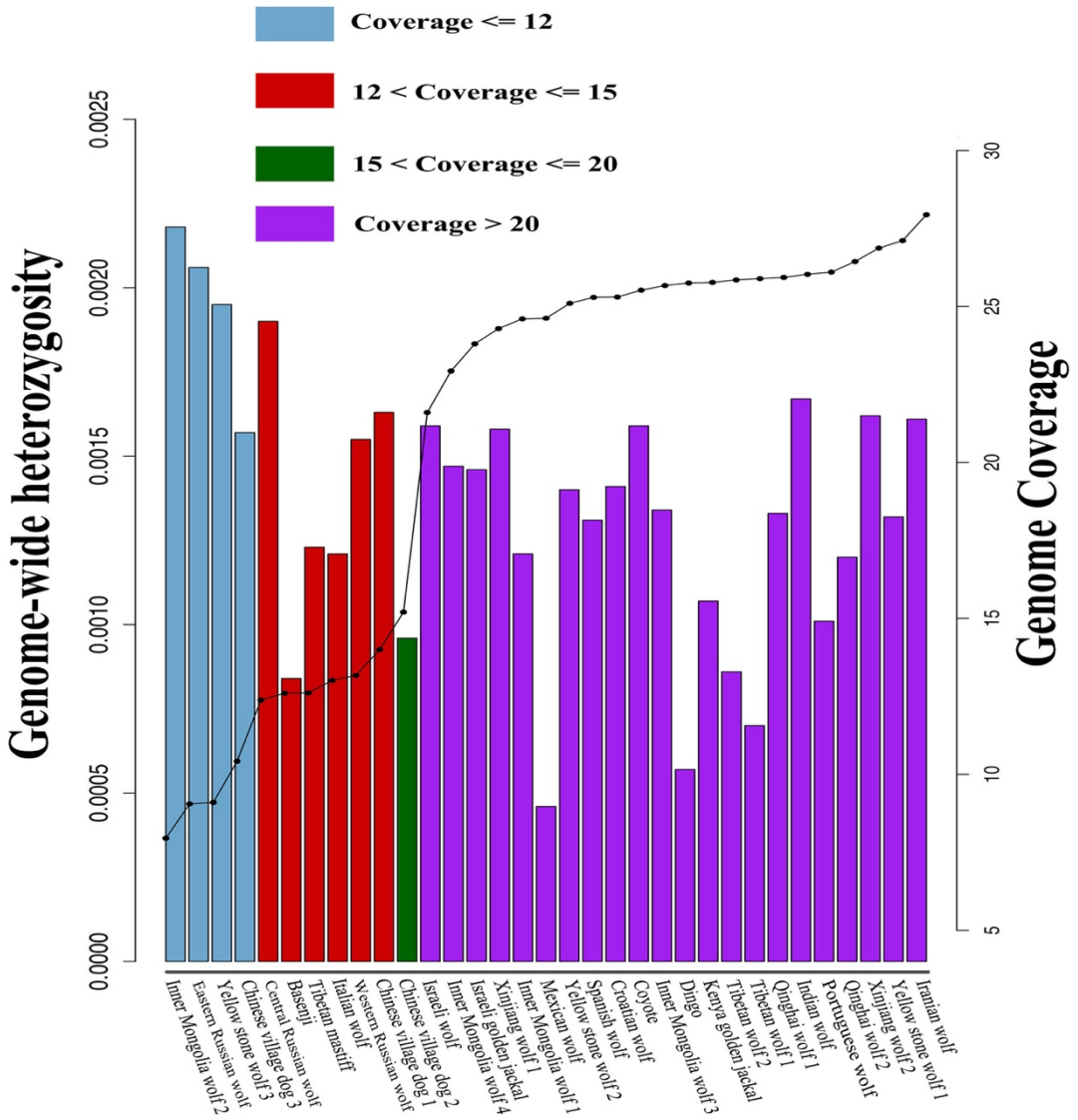


Figure S5. The genome wide heterozygosity and genome coverage. The bar is the genome wide heterozygosity in each sample, and the black plot is the genome coverage.

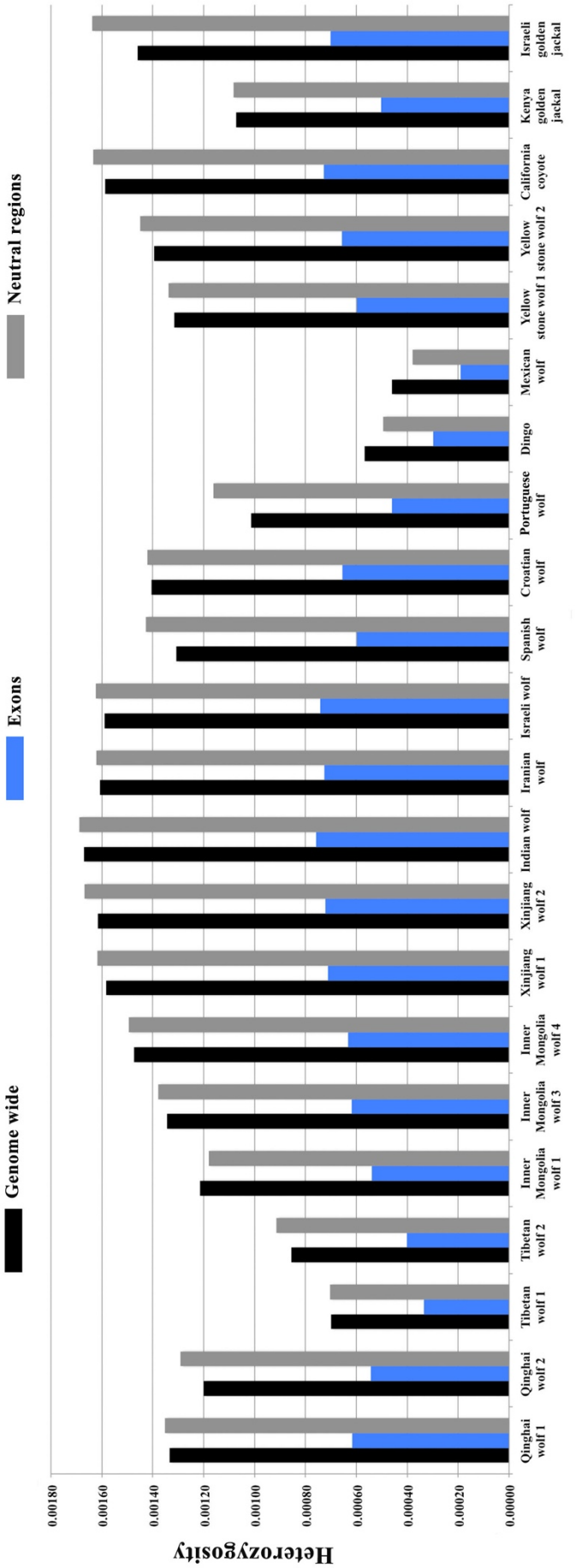


Figure S6. The heterozygosity of different genomic classes. The black bar is the genome wide heterozygosity, the blue bar is the heterozygosity of exons, and the grey bar is the heterozygosity of neutral regions. We only used samples with > 20-fold genome coverage.

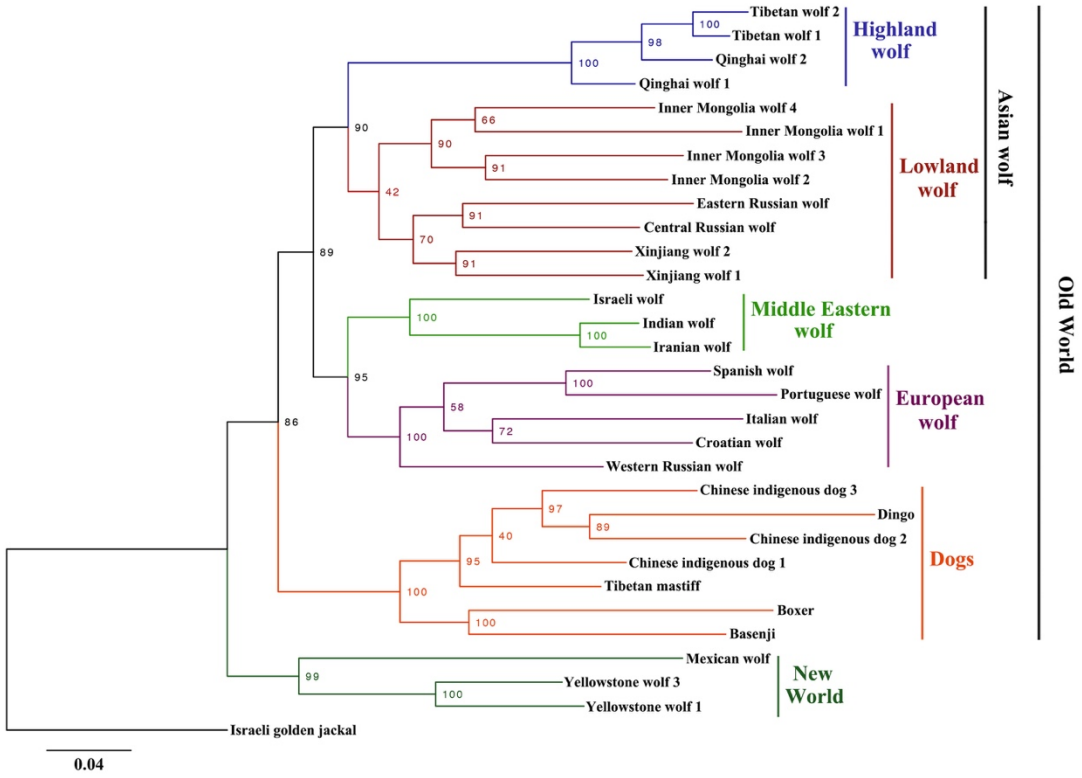


Figure S7. The maximum likelihood tree of 31 sequences. Numbers represent node supports inferred from 100 bootstrap repetitions. We excluded the Yellowstone wolf 2 because it is the offspring of Yellowstone wolf 1 (mother) and Yellowstone wolf 3 (father). The reference genome boxer was included. The Israeli golden jackal is the outgroup.

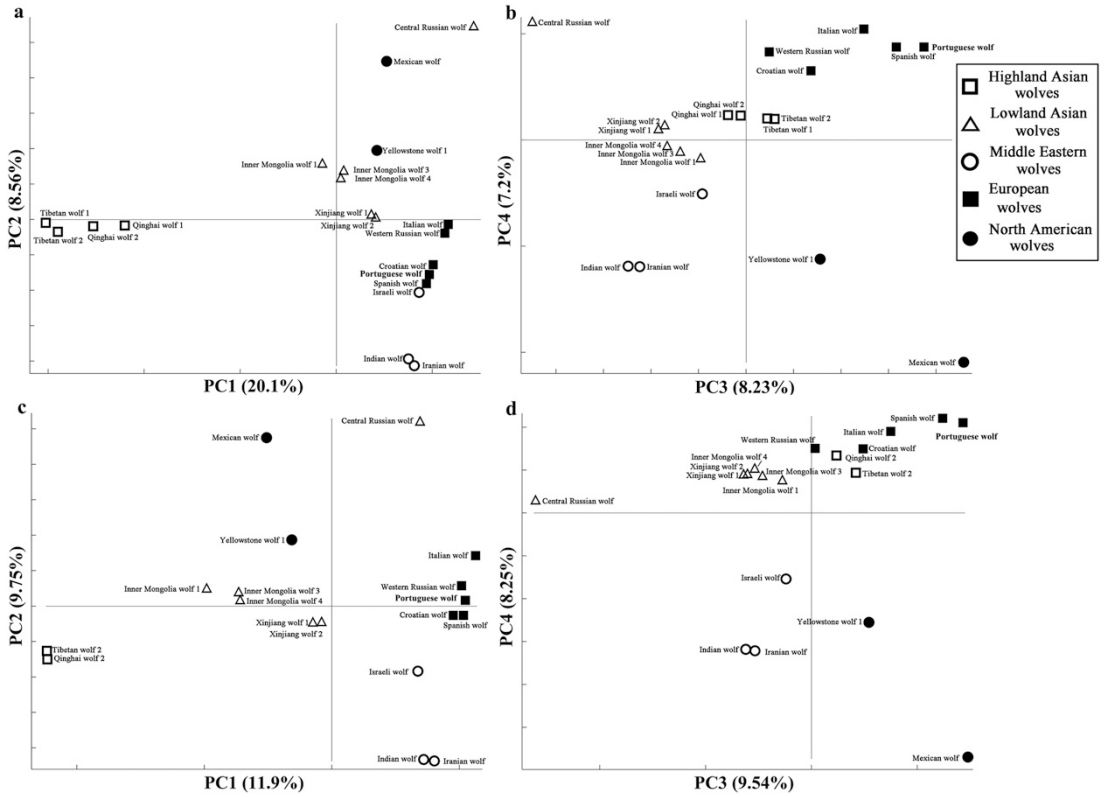


Figure S8. Principal component analyses of complete genome data for only wolves. Inner Mongolia wolf 2, Eastern Russian wolf, and Yellowstone wolf 3 were also removed due to their potential high genotype error (see Figure S5). We also excluded the Yellowstone wolf 2 as in Figure 3. (a) PC1 and PC2 of 20 wolves; (b) PC3 and PC4 of 20 wolves; (c) PC1 and PC2 of 18 wolves, excluding the Tibetan wolf 1 and Qinghai wolf 1; (d) PC3 and PC4 of 18 wolves, excluding the Tibetan wolf 1 and Qinghai wolf 1.

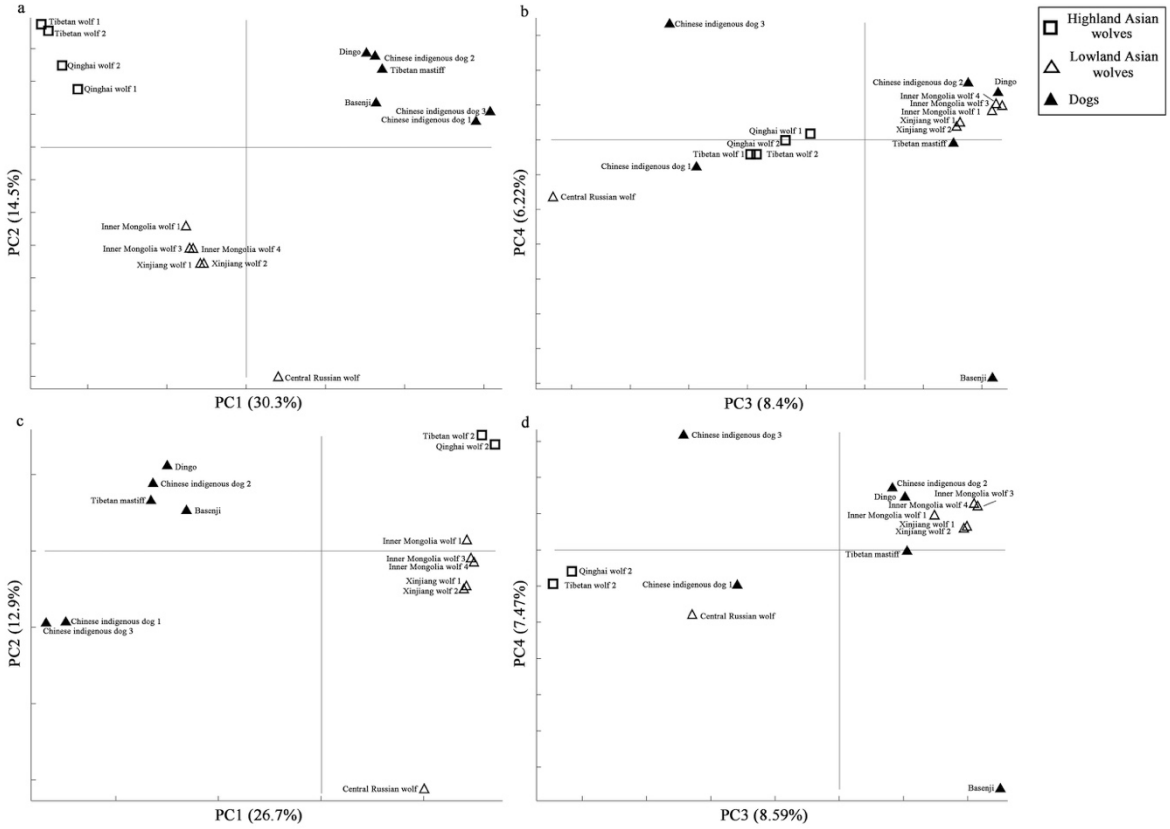


Figure S9. Principal component analysis (PCA) of complete genome data for Asian wolves and all the dogs. The results from PC1 to PC4 are shown. (a and b) Asian wolves and all dogs; (c and d) Asian wolves without Tibetan wolf 1 and Qinghai wolf 1 and all dogs

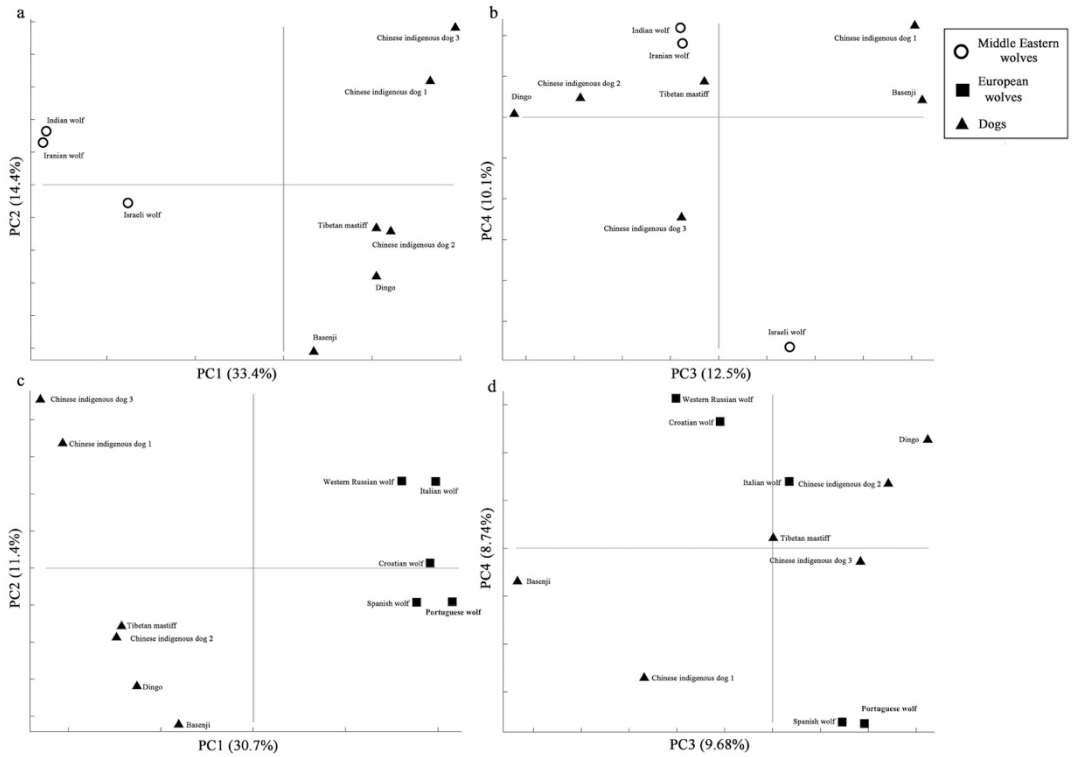


Figure S10. Principal component analysis (PCA) of complete genome data for Middle Eastern wolves, European wolves and all the dogs. The results from PC1 to PC4 are shown. (a and b) Middle Eastern wolves and all dogs. Indian wolf is grouped into Middle Eastern wolf here; (c and d) European wolves and all dogs.

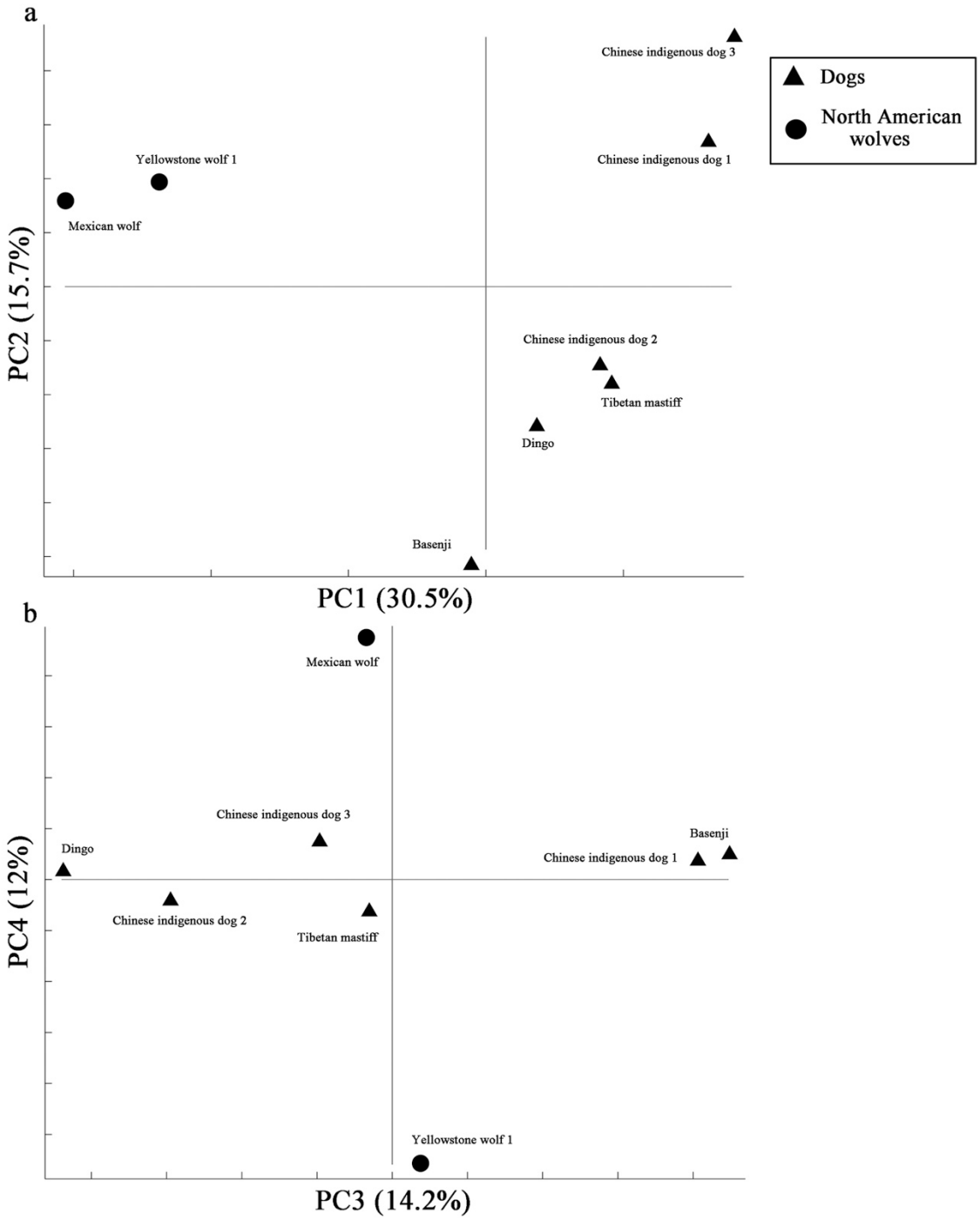


Figure S11. Principal component analysis (PCA) of complete genome data for Mexican wolf, Yellowstone wolf and all the dogs. The results from PC1 to PC4 are shown. Only one Yellowstone wolf (Yellowstone wolf 1) is used here.

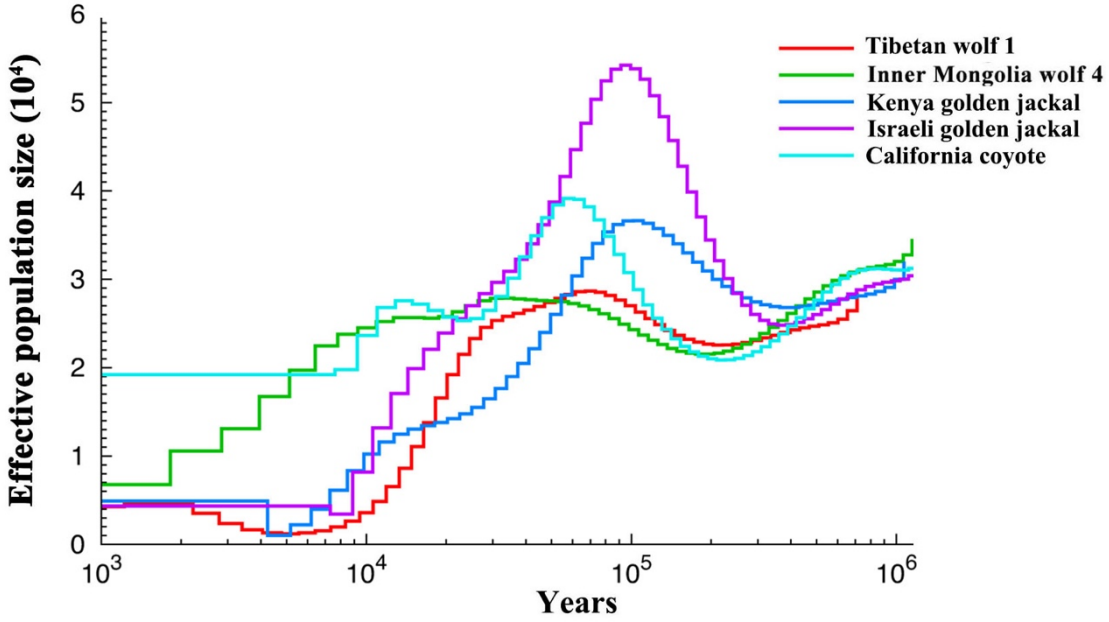


Figure S12. Demographic history of three outgroups in this study reconstructed from the their genomes analyzed with the genomic distribution of heterozygous sites using the pairwise sequential Markovian coalescent (PSMC). Followed Freedman et al. (2014) and Zhang et al. (2014), generation time = 3 and mutation rate = 1.0×10^{-8} per generation are applied. Tibetan wolf 1 and Inner Mongolia wolf 4 are exhibited here.

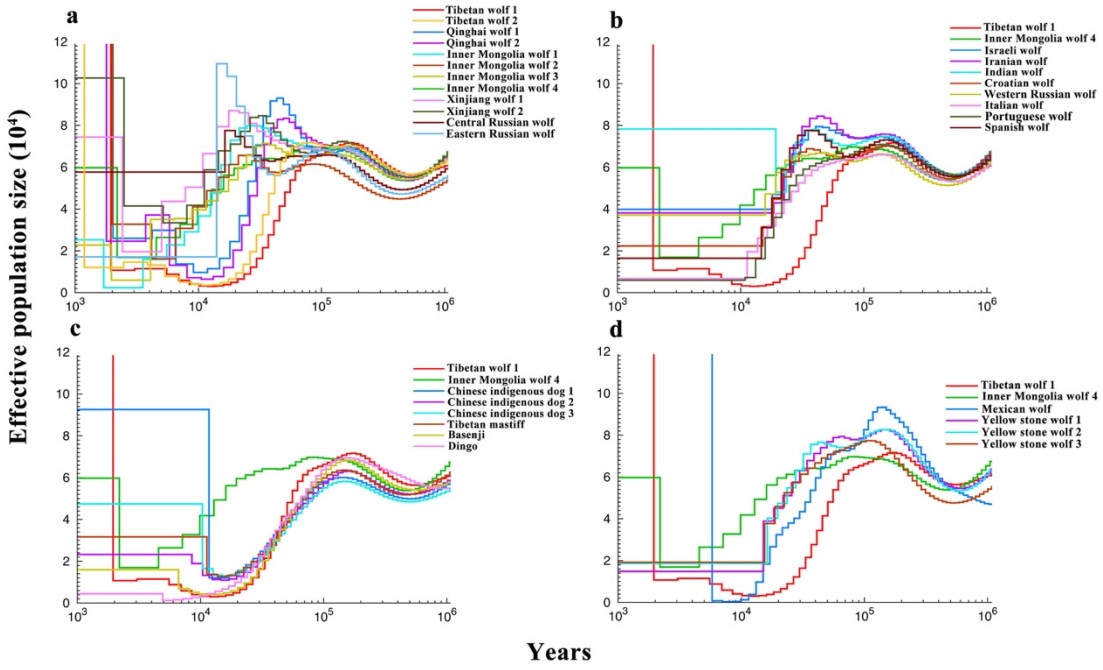


Figure S13. Demographic history of the candis in this study reconstructed from the their genomes analyzed with the genomic distribution of heterozygous sites using the pairwise sequential Markovian coalescent (PSMC). Followed Freedman et al. (2014) and Zhang et al. (2014), generation time = 3 is applied. Skoglund et al. (2015)'s reported a slower mutation rate (0.4×10^{-8} per generation), thus we applied this mutation rate here. Tibetan wolf 1 and Inner Mongolia wolf 4 are exhibited in all the plots. (a) all the Asian wolves; (b) all the European wolves and Middle Eastern wolves and Indian wolf; (c) dogs; (d) Mexican wolf and Yellowstone wolves.

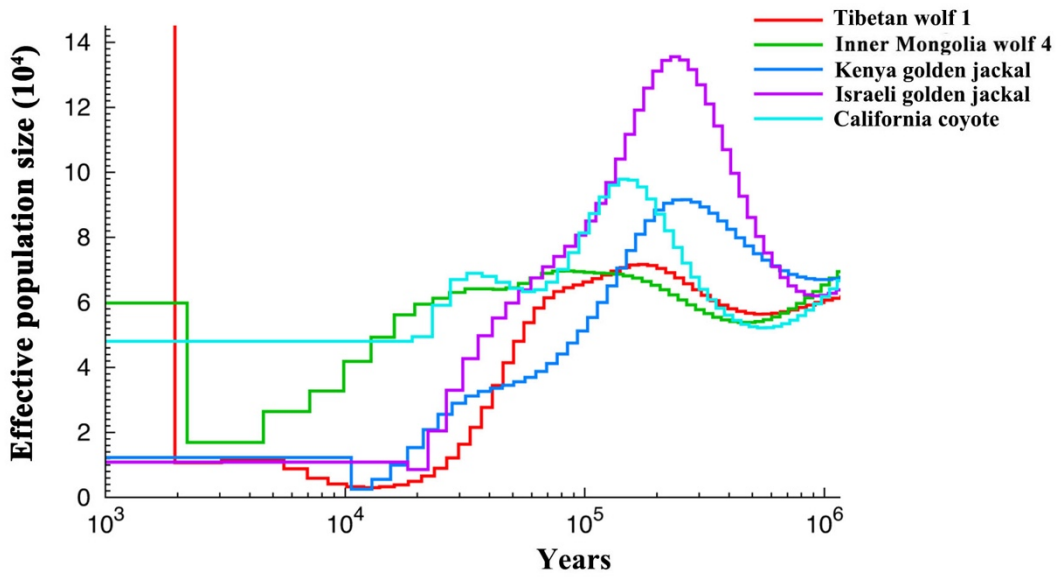


Figure S14. Demographic history of three outgroups in this study reconstructed from the their genomes analyzed with the genomic distribution of heterozygous sites using the pairwise sequential Markovian coalescent (PSMC). Followed Freedman et al. (2014) and Zhang et al. (2014), generation time = 3 is applied. Skoglund et al. (2015)'s reported a slower mutation rate (0.4×10^{-8} per generation), thus we applied this mutation rate here. Tibetan wolf 1 and Inner Mongolia wolf 4 are exhibited in all the plots.

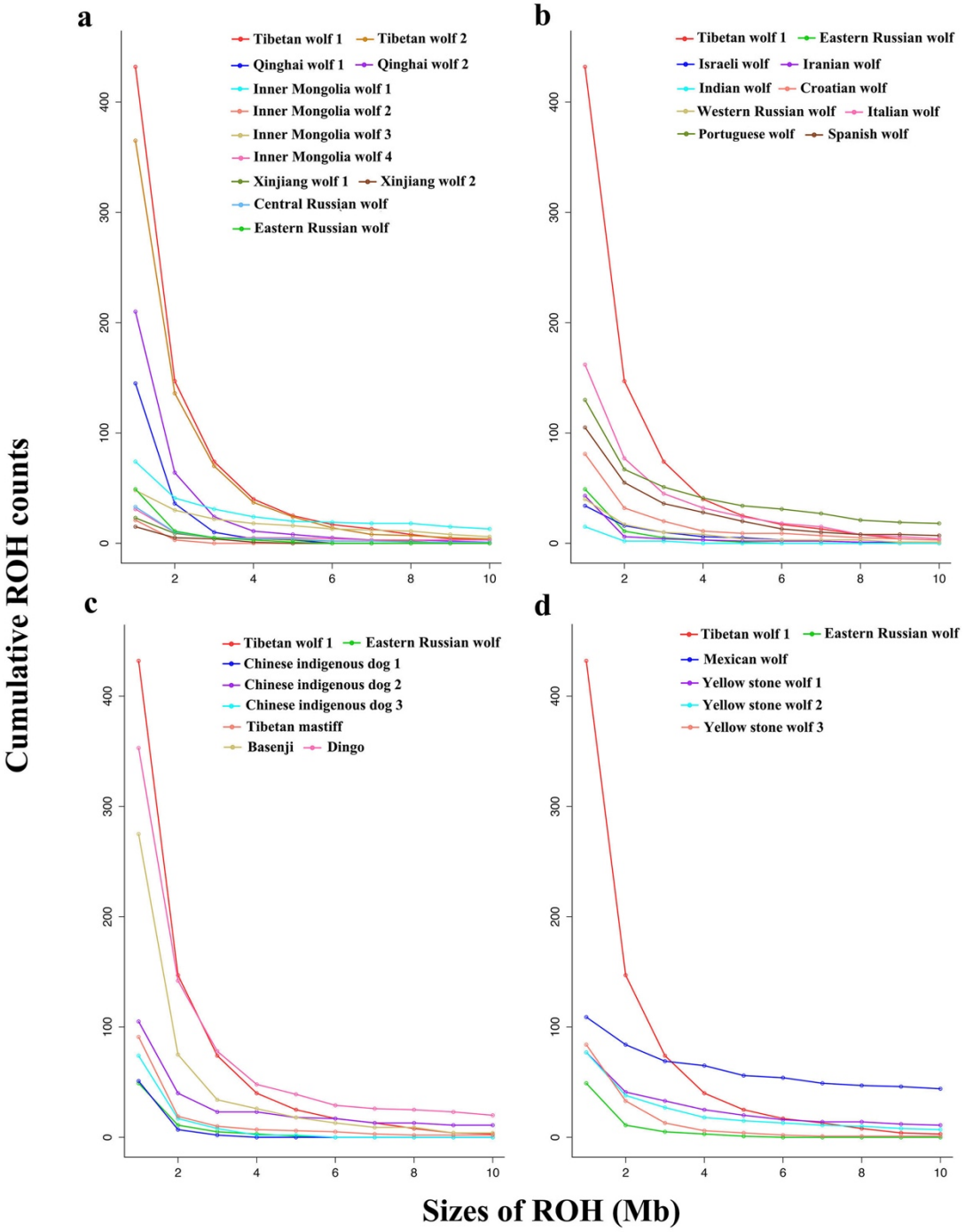


Figure S15. Autozygosity frequency distribution of runs of homozygosity (ROH). Tibetan wolf 1 and Russian wolf (east) are exhibited in all the plots. (a) Asian wolf; (b) European wolf and Middle Eastern wolf; (c) Dogs; (d) Mexican wolf and Yellowstone wolves.

Table S1. The detail information of samples in this study, which including species name, population, code and genome coverage

Name	Species	Population	Location	Genome coverage
Chinese indigenous dog 1[$]	Dog	Chinese indigenous dog	Xi'an, China	14
Chinese indigenous dog 2[$]	Dog	Chinese indigenous dog	Simao, China	15.2
Chinese indigenous dog 3[$]	Dog	Chinese indigenous dog	Ya'an, China	10.42
Tibetan mastiff [$]	Dog	Tibetan mastiff	Lijiang, China	12.61
Basenji [#]	Dog	Basenji	Bethesda, MD, USA	12.6
Dingo [#]	Dog	Dingo	Bargo Dingo Sanctuary, Australia	25.75
Spanish wolf [*]	Gray wolf	Europe	Spain	25.29
Croatian wolf [#]	Gray wolf	Europe	Perković, Croatia	25.3
Italian wolf [*]	Gray wolf	Europe	Italy	13.01
Portuguese wolf [*]	Gray wolf	Europe	Portugal (N of Douro)	26.1
Mexican wolf [*]	Mexican Gray wolf	North America	Captive	24.62
Yellow stone wolf 1 [*]	Gray wolf	North America	Yellowstone NP	27.11
Yellow stone wolf 2 [*]	Gray wolf	North America	Yellowstone NP	25.1
Yellow stone wolf 3 [*]	Gray wolf	North America	Yellowstone NP	9.1
Inner Mongolia wolf 1 [#]	Gray wolf	Asia	San Diego Zoo; Maybe Inner Mongolia	24.6
Inner Mongolia wolf 2 [$]	Gray wolf	Asia	Inner Mongolia, China	7.95
Inner Mongolia wolf 3 [@]	Gray wolf	Asia	Inner Mongolia, China	25.67
Inner Mongolia wolf 4 [@]	Gray wolf	Asia	Inner Mongolia, China	22.93
Xinjiang wolf 1 [@]	Gray wolf	Asia	Xinjiang, China	24.29
Xinjiang wolf 2 [@]	Gray wolf	Asia	Xinjiang, China	26.87
Qinghai wolf 1 [@]	Gray wolf	Asia	Qinghai, China	25.93
Qinghai wolf 2 [@]	Gray wolf	Asia	Qinghai, China	26.44
Tibet wolf 1 [@]	Gray wolf	Asia	Tibet, China	25.89
Tibet wolf 2 [@]	Gray wolf	Asia	Tibet, China	25.85
Indian wolf [*]	Gray wolf	Middle East	India (Koln Zoo)	26.03
Central Russian wolf [$]	Gray wolf	Asia	Altai, Russia	12.38
Eastern Russian wolf [$]	Gray wolf	Asia	Chukotka, Russia	9.05
Western Russian wolf [$]	Gray wolf	Europe	Bryansk, Russia	13.17
Iranian wolf [*]	Gray wolf	Middle East	Iran	27.94
Israeli wolf [#]	Gray wolf	Middle East	Neve Ativ, Golan Heights, Israel	21.6
Kenya golden jackal [*]	Golden jackal	Africa	Kenya	25.77
Israeli golden jackal [#]	Golden jackal	Middle East	Tel Aviv, Israel	23.8
California coyote [*]	Coyote	North America (Western)	California	25.52

* Generated in this study,$ Raw data were downloaded from Wang et al. 2013; @ Raw data were downloaded from Zhang et al. 2014; # Genotype results were obtained from the authors of Freedman et al. 2014

Table S2. The PCR duplicates, genome coverage, and sites information

Name	PCR duplicates (%)	Genome coverage	Non-variant sites	Numbers of SNPs			Total sites	Heterozygosity	SNP rate
				Heterozygous SNPs	Homozygous SNPs	Total SNPs			
Chinese indigenous dog 1	20.82	14	803248629	1309187	513017	1822204	805070833	0.00163	0.00226
Chinese indigenous dog 2	22.39	15.2	1272599310	1217405	944618	2162023	1274761333	0.00096	0.00170
Chinese indigenous dog 3	12.99	10.42	750804488	1181229	562823	1744052	752548540	0.00157	0.00232
Tibetan mastiff	5.76	12.61	1216398524	1494005	718085	2212090	1218610614	0.00123	0.00182
Basenji	28.5	12.6	1131597756	952848	949823	1902671	1133500427	0.00084	0.00168
Dingo	28.5	25.75	1308195288	742600	1416283	2158883	1310354171	0.00057	0.00165
Spanish wolf	20.62	25.29	1365732542	1789063	1351763	3140826	1368873368	0.00131	0.00229
Croatian wolf	16	25.3	1288847615	1815538	1183760	2999298	1291846913	0.00141	0.00232
Italian wolf	55.08	13.01	1210244651	1473009	1360614	2833623	1213078274	0.00121	0.00234
Portuguese wolf	12.57	26.1	1365562467	1387157	1661711	3048868	1368611335	0.00101	0.00223
Mexican wolf	4.72	24.62	1375906180	633161	2138138	2771299	1378677479	0.00046	0.00201
Yellow Stone wolf 1	6.42	27.11	1378055183	1817301	1541004	3358305	1381413488	0.00132	0.00243
Yellow Stone wolf 2	4.77	25.1	1375726990	1924746	1489098	3413844	1379140834	0.00140	0.00248
Yellow Stone wolf 3	64.35	9.1	819854963	1604333	858833	2463166	822318129	0.00195	0.00300
Inner Mongolia wolf 1	6.69	24.6	1369359640	1665253	1520771	3186024	1372545664	0.00121	0.00232
Inner Mongolia wolf 2	6.19	7.95	668228310	1458901	527345	1986246	670214556	0.00218	0.00296
Inner Mongolia wolf 3	11.49	25.67	1367267191	1841660	1393626	3235286	1370502477	0.00134	0.00236
Inner Mongolia wolf 4	15.96	22.93	1352264802	1996707	1278791	3275498	1355540300	0.00147	0.00242
Xinjiang wolf 1	6.93	24.29	1363690421	2164974	1237352	3402326	1367092747	0.00158	0.00249
Xinjiang wolf 2	6.82	26.87	1367249609	2215265	1225522	3440787	1370690396	0.00162	0.00251
Qinghai wolf 1	10.62	25.93	1371964002	1835047	1459374	3294421	1375258423	0.00133	0.00240

Qinghai wolf 2	8.62	26.44	1365287769	1641837	1557576	3199413	1368487182	0.00120	0.00234
Tibet wolf 1	10.06	25.89	1363362087	954748	1938855	2893603	1366255690	0.00070	0.00212
Tibet wolf 2	9.41	25.85	1365396492	1170059	1807144	2977203	1368373695	0.00086	0.00218
Indian wolf	6.49	26.03	1377349627	2305572	1290241	3595813	1380945440	0.00167	0.00260
Central Russian wolf	8.25	12.38	911425890	1732836	824677	2557513	913983403	0.00190	0.00280
Eastern Russian wolf	8.08	9.05	683598733	1412388	718504	2130892	685729625	0.00206	0.00311
Western Russian wolf	4.2	13.17	1204008617	1868427	974059	2842486	1206851103	0.00155	0.00236
Iranian wolf	11.31	27.94	1369176495	2206398	1309775	3516173	1372692668	0.00161	0.00256
Israeli wolf	20.2	21.6	1335801804	2128929	1100567	3229496	1339031300	0.00159	0.00241
kenya golden jackal	6.57	25.77	1363012971	1466825	2773320	4240145	1367253116	0.00107	0.00310
Israeli golden jackal	51.5	23.8	1310197591	1919318	2833078	4752396	1314949987	0.00146	0.00361
California coyote	8.98	25.52	1374588619	2187884	2101381	4289265	1378877884	0.00159	0.00311

Table S3. ABBA-BABA test (D-statistic) using Israeli golden jackal as outgroup. Different dogs were put in P3 in the tests. For Asian wolves, three Chinese indigenous dogs, Tibetan mastiff, and dingo were included. For European and Middle Eastern wolves, basenji and boxer were included. Z score with absolute value bigger than 3 was considered significant.

P1	P2	P3	O	D Statistic	Standard Error	Z Score	significant and Direction
Inner Mongolia wolf 1	Indian wolf	Chinese indigenous dog 1	Israeli golden jackal	-0.033837787	0.004290113	-7.887389154	Inner Mongolia wolf 1 and Chinese indigenous dog 1
Inner Mongolia wolf 2	Indian wolf	Chinese indigenous dog 1	Israeli golden jackal	-0.080461207	0.003828093	-21.01861494	Inner Mongolia wolf 2 and Chinese indigenous dog 1
Inner Mongolia wolf 3	Indian wolf	Chinese indigenous dog 1	Israeli golden jackal	-0.042909418	0.003597074	-11.92897931	Inner Mongolia wolf 3 and Chinese indigenous dog 1
Inner Mongolia wolf 4	Indian wolf	Chinese indigenous dog 1	Israeli golden jackal	-0.039192967	0.003616004	-10.83875055	Inner Mongolia wolf 4 and Chinese indigenous dog 1
Xinjiang wolf 1	Indian wolf	Chinese indigenous dog 1	Israeli golden jackal	-0.033127795	0.003501125	-9.462043611	Xinjiang wolf 1 and Chinese indigenous dog 1
Xinjiang wolf 2	Indian wolf	Chinese indigenous dog 1	Israeli golden jackal	-0.031020981	0.003477823	-8.919654457	Xinjiang wolf 2 and Chinese indigenous dog 1
Qinghai wolf 1	Indian wolf	Chinese indigenous dog 1	Israeli golden jackal	-0.010879248	0.00439615	-2.474721738	None significant
Qinghai wolf 2	Indian wolf	Chinese indigenous dog 1	Israeli golden jackal	-0.001438628	0.0051327	-0.280286798	None significant
Tibetan wolf 1	Indian wolf	Chinese indigenous dog 1	Israeli golden jackal	0.017714628	0.005744828	3.083578602	Indian wolf and Chinese indigenous dog 1
Tibetan wolf 2	Indian wolf	Chinese indigenous dog 1	Israeli golden jackal	0.01594739	0.005647844	2.82362422	None significant
Central Russian wolf	Indian wolf	Chinese indigenous dog 1	Israeli golden jackal	-0.062259313	0.004206438	-14.80095858	Central Russian wolf and Chinese indigenous dog 1
Eastern Russian wolf	Indian wolf	Chinese indigenous dog 1	Israeli golden jackal	-0.071502295	0.004002141	-17.8660102	Eastern Russian wolf and Chinese indigenous dog 1
Indian wolf	Inner Mongolia wolf 1	Chinese indigenous dog 1	Israeli golden jackal	0.032988425	0.004147514	7.953782964	Inner Mongolia wolf 1 and Chinese indigenous dog 1
Indian wolf	Inner Mongolia wolf 2	Chinese indigenous dog 1	Israeli golden jackal	0.081523036	0.003805297	21.42356454	Inner Mongolia wolf 2 and Chinese indigenous dog 1
Indian wolf	Inner Mongolia wolf 3	Chinese indigenous dog 1	Israeli golden jackal	0.042232363	0.003649846	11.57099802	Inner Mongolia wolf 3 and Chinese indigenous dog 1
Indian wolf	Inner Mongolia wolf 4	Chinese indigenous dog 1	Israeli golden jackal	0.038060023	0.003596996	10.58105727	Inner Mongolia wolf 4 and Chinese indigenous dog 1
Indian wolf	Xinjiang wolf 1	Chinese indigenous dog 1	Israeli golden jackal	0.033868987	0.00362571	9.341339164	Xinjiang wolf 1 and Chinese indigenous dog 1
Indian wolf	Xinjiang wolf 2	Chinese indigenous dog 1	Israeli golden jackal	0.033962847	0.00360429	9.422894223	Xinjiang wolf 2 and Chinese indigenous dog 1
Indian wolf	Qinghai wolf 1	Chinese indigenous dog 1	Israeli golden jackal	0.00913177	0.004387272	2.081423399	None significant

Indian wolf	Qinghai wolf 2	Chinese indigenous dog 1	Israeli jackal	golden	0.000398113	0.005143362	0.077403212	None significant
Indian wolf	Tibetan wolf 1	Chinese indigenous dog 1	Israeli jackal	golden	-0.016001387	0.005712667	-2.801036124	None significant
Indian wolf	Tibetan wolf 2	Chinese indigenous dog 1	Israeli jackal	golden	-0.017022582	0.005746549	-2.962226701	None significant
Indian wolf	Central Russian wolf	Chinese indigenous dog 1	Israeli jackal	golden	0.063416637	0.004184561	15.1549092	Central Russian wolf and Chinese indigenous dog 1
Indian wolf	Eastern Russian wolf	Chinese indigenous dog 1	Israeli jackal	golden	0.071234053	0.004008459	17.77093252	Eastern Russian wolf and Chinese indigenous dog 1
Inner Mongolia wolf 1	Indian wolf	Chinese indigenous dog 2	Israeli jackal	golden	-0.045340984	0.004168975	-10.87581151	Inner Mongolia wolf 1 and Chinese indigenous dog 2
Inner Mongolia wolf 2	Indian wolf	Chinese indigenous dog 2	Israeli jackal	golden	-0.078562173	0.003796456	-20.69355527	Inner Mongolia wolf 2 and Chinese indigenous dog 2
Inner Mongolia wolf 3	Indian wolf	Chinese indigenous dog 2	Israeli jackal	golden	-0.053592305	0.003696926	-14.49644947	Inner Mongolia wolf 3 and Chinese indigenous dog 2
Inner Mongolia wolf 4	Indian wolf	Chinese indigenous dog 2	Israeli jackal	golden	-0.049780632	0.003544237	-14.04551241	Inner Mongolia wolf 4 and Chinese indigenous dog 2
Xinjiang wolf 1	Indian wolf	Chinese indigenous dog 2	Israeli jackal	golden	-0.039225811	0.003705269	-10.58649445	Xinjiang wolf 1 and Chinese indigenous dog 2
Xinjiang wolf 2	Indian wolf	Chinese indigenous dog 2	Israeli jackal	golden	-0.040543284	0.003613705	-11.21931328	Xinjiang wolf 2 and Chinese indigenous dog 2
Qinghai wolf 1	Indian wolf	Chinese indigenous dog 2	Israeli jackal	golden	-0.023242606	0.004148588	-5.602533541	Qinghai wolf 1 and Chinese indigenous dog 2
Qinghai wolf 2	Indian wolf	Chinese indigenous dog 2	Israeli jackal	golden	-0.010217057	0.004867276	-2.099132665	None significant
Tibetan wolf 1	Indian wolf	Chinese indigenous dog 2	Israeli jackal	golden	0.005606652	0.005303789	1.057103062	None significant
Tibetan wolf 2	Indian wolf	Chinese indigenous dog 2	Israeli jackal	golden	0.004953521	0.005394274	0.918292586	None significant
Central Russian wolf	Indian wolf	Chinese indigenous dog 2	Israeli jackal	golden	-0.057368254	0.004052674	-14.1556531	Central Russian wolf and Chinese indigenous dog 2
Eastern Russian wolf	Indian wolf	Chinese indigenous dog 2	Israeli jackal	golden	-0.068239153	0.003987197	-17.11456925	Eastern Russian wolf and Chinese indigenous dog 2
Indian wolf	Inner Mongolia wolf 1	Chinese indigenous dog 2	Israeli jackal	golden	0.045034018	0.004260349	10.57050098	Inner Mongolia wolf 1 and Chinese indigenous dog 2
Indian wolf	Inner Mongolia wolf 2	Chinese indigenous dog 2	Israeli jackal	golden	0.077837733	0.003852949	20.20211623	Inner Mongolia wolf 2 and Chinese indigenous dog 2
Indian wolf	Inner Mongolia wolf 3	Chinese indigenous dog 2	Israeli jackal	golden	0.054966906	0.003647502	15.06974163	Inner Mongolia wolf 3 and Chinese indigenous dog 2
Indian wolf	Inner Mongolia wolf 4	Chinese indigenous dog 2	Israeli jackal	golden	0.049097582	0.003617889	13.57078255	Inner Mongolia wolf 4 and Chinese indigenous dog 2

Indian wolf	Xinjiang wolf 1	Chinese indigenous dog 2	Israeli jackal	golden	0.042190234	0.003674743	11.48113769	Xinjiang wolf 1 and Chinese indigenous dog 2
Indian wolf	Xinjiang wolf 2	Chinese indigenous dog 2	Israeli jackal	golden	0.040334428	0.003512355	11.48358597	Xinjiang wolf 2 and Chinese indigenous dog 2
Indian wolf	Qinghai wolf 1	Chinese indigenous dog 2	Israeli jackal	golden	0.024418846	0.004203483	5.809193911	Qinghai wolf 1 and Chinese indigenous dog 2
Indian wolf	Qinghai wolf 2	Chinese indigenous dog 2	Israeli jackal	golden	0.010536996	0.004875735	2.161109195	None significant
Indian wolf	Tibetan wolf 1	Chinese indigenous dog 2	Israeli jackal	golden	-0.006409387	0.00530184	-1.208898564	None significant
Indian wolf	Tibetan wolf 2	Chinese indigenous dog 2	Israeli jackal	golden	-0.003922505	0.005460372	-0.718358459	None significant
Indian wolf	Central Russian wolf	Chinese indigenous dog 2	Israeli jackal	golden	0.057125934	0.004121544	13.86032464	Central Russian wolf and Chinese indigenous dog 2
Indian wolf	Eastern Russian wolf	Chinese indigenous dog 2	Israeli jackal	golden	0.070349455	0.003913779	17.97481643	Eastern Russian wolf and Chinese indigenous dog 2
Inner Mongolia wolf 1	Indian wolf	Chinese indigenous dog 3	Israeli jackal	golden	-0.042249004	0.00408022	-10.35459093	Inner Mongolia wolf 1 and Chinese indigenous dog 3
Inner Mongolia wolf 2	Indian wolf	Chinese indigenous dog 3	Israeli jackal	golden	-0.083375664	0.003790967	-21.99324287	Inner Mongolia wolf 2 and Chinese indigenous dog 3
Inner Mongolia wolf 3	Indian wolf	Chinese indigenous dog 3	Israeli jackal	golden	-0.050140251	0.003662101	-13.69166162	Inner Mongolia wolf 3 and Chinese indigenous dog 3
Inner Mongolia wolf 4	Indian wolf	Chinese indigenous dog 3	Israeli jackal	golden	-0.047177538	0.003422731	-13.78359588	Inner Mongolia wolf 4 and Chinese indigenous dog 3
Xinjiang wolf 1	Indian wolf	Chinese indigenous dog 3	Israeli jackal	golden	-0.036604403	0.003591494	-10.19197221	Xinjiang wolf 1 and Chinese indigenous dog 3
Xinjiang wolf 2	Indian wolf	Chinese indigenous dog 3	Israeli jackal	golden	-0.039421178	0.003690929	-10.68055572	Xinjiang wolf 2 and Chinese indigenous dog 3
Qinghai wolf 1	Indian wolf	Chinese indigenous dog 3	Israeli jackal	golden	-0.016289495	0.004143488	-3.931348415	Qinghai wolf 1 and Chinese indigenous dog 3
Qinghai wolf 2	Indian wolf	Chinese indigenous dog 3	Israeli jackal	golden	-0.008649655	0.005002051	-1.729221867	None significant
Tibetan wolf 1	Indian wolf	Chinese indigenous dog 3	Israeli jackal	golden	0.008730885	0.005486926	1.591216189	None significant
Tibetan wolf 2	Indian wolf	Chinese indigenous dog 3	Israeli jackal	golden	0.006638829	0.005602848	1.184902538	None significant
Central Russian wolf	Indian wolf	Chinese indigenous dog 3	Israeli jackal	golden	-0.062046957	0.004093477	-15.15752005	Central Russian wolf and Chinese indigenous dog 3
Eastern Russian wolf	Indian wolf	Chinese indigenous dog 3	Israeli jackal	golden	-0.070733463	0.003998951	-17.68800217	Eastern Russian wolf and Chinese indigenous dog 3
Indian wolf	Inner Mongolia wolf 1	Chinese indigenous dog 3	Israeli jackal	golden	0.043983265	0.004108508	10.70541023	Inner Mongolia wolf 1 and Chinese indigenous dog 3

Indian wolf	Inner Mongolia wolf 2	Chinese indigenous dog 3	Israeli jackal	golden	0.08281032	0.003837119	21.58137996	Inner Mongolia wolf 2 and Chinese indigenous dog 3
Indian wolf	Inner Mongolia wolf 3	Chinese indigenous dog 3	Israeli jackal	golden	0.050521838	0.003626176	13.93254044	Inner Mongolia wolf 3 and Chinese indigenous dog 3
Indian wolf	Inner Mongolia wolf 4	Chinese indigenous dog 3	Israeli jackal	golden	0.048278329	0.003423661	14.10137606	Inner Mongolia wolf 4 and Chinese indigenous dog 3
Indian wolf	Xinjiang wolf 1	Chinese indigenous dog 3	Israeli jackal	golden	0.036750288	0.003732461	9.846126636	Xinjiang wolf 1 and Chinese indigenous dog 3
Indian wolf	Xinjiang wolf 2	Chinese indigenous dog 3	Israeli jackal	golden	0.040294296	0.003649837	11.04002513	Xinjiang wolf 2 and Chinese indigenous dog 3
Indian wolf	Qinghai wolf 1	Chinese indigenous dog 3	Israeli jackal	golden	0.018562788	0.004195908	4.424021271	Qinghai wolf 1 and Chinese indigenous dog 3
Indian wolf	Qinghai wolf 2	Chinese indigenous dog 3	Israeli jackal	golden	0.008926909	0.005005912	1.783273362	None significant
Indian wolf	Tibetan wolf 1	Chinese indigenous dog 3	Israeli jackal	golden	-0.008217292	0.005471299	-1.50189039	None significant
Indian wolf	Tibetan wolf 2	Chinese indigenous dog 3	Israeli jackal	golden	-0.008293511	0.005678328	-1.460555106	None significant
Indian wolf	Central Russian wolf	Chinese indigenous dog 3	Israeli jackal	golden	0.062368266	0.004130293	15.10020429	Central Russian wolf and Chinese indigenous dog 3
Indian wolf	Eastern Russian wolf	Chinese indigenous dog 3	Israeli jackal	golden	0.072244581	0.004063814	17.77753317	Eastern Russian wolf and Chinese indigenous dog 3
Inner Mongolia wolf 1	Indian wolf	Tibetan Mastiff	Israeli jackal	golden	-0.033529906	0.004083634	-8.210801471	Inner Mongolia wolf 1 and Tibetan Mastiff
Inner Mongolia wolf 2	Indian wolf	Tibetan Mastiff	Israeli jackal	golden	-0.073126755	0.003724104	-19.63606851	Inner Mongolia wolf 2 and Tibetan Mastiff
Inner Mongolia wolf 3	Indian wolf	Tibetan Mastiff	Israeli jackal	golden	-0.040783939	0.003321307	-12.27948611	Inner Mongolia wolf 3 and Tibetan Mastiff
Inner Mongolia wolf 4	Indian wolf	Tibetan Mastiff	Israeli jackal	golden	-0.037918606	0.003435479	-11.03735705	Inner Mongolia wolf 4 and Tibetan Mastiff
Xinjiang wolf 1	Indian wolf	Tibetan Mastiff	Israeli jackal	golden	-0.032853941	0.003600387	-9.125113451	Xinjiang wolf 1 and Tibetan Mastiff
Xinjiang wolf 2	Indian wolf	Tibetan Mastiff	Israeli jackal	golden	-0.033639996	0.003477758	-9.672896635	Xinjiang wolf 2 and Tibetan Mastiff
Qinghai wolf 1	Indian wolf	Tibetan Mastiff	Israeli jackal	golden	-0.011085664	0.004223245	-2.624916276	None significant
Qinghai wolf 2	Indian wolf	Tibetan Mastiff	Israeli jackal	golden	8.69E-05	0.004990967	0.017405666	None significant
Tibetan wolf 1	Indian wolf	Tibetan Mastiff	Israeli jackal	golden	0.015575895	0.005449474	2.858238046	None significant
Tibetan wolf 2	Indian wolf	Tibetan Mastiff	Israeli jackal	golden	0.012282863	0.005570031	2.205169793	None significant

Central Russian wolf	Indian wolf	Tibetan Mastiff	Israeli jackal	golden	-0.053977101	0.003748519	-14.39958218	Central Russian wolf and Tibetan Mastiff
Eastern Russian wolf	Indian wolf	Tibetan Mastiff	Israeli jackal	golden	-0.062778509	0.003940158	-15.93299097	Eastern Russian wolf and Tibetan Mastiff
Indian wolf	Inner Mongolia wolf 1	Tibetan Mastiff	Israeli jackal	golden	0.034275692	0.004045238	8.473096363	Inner Mongolia wolf 1 and Tibetan Mastiff
Indian wolf	Inner Mongolia wolf 2	Tibetan Mastiff	Israeli jackal	golden	0.073367508	0.00371928	19.72626543	Inner Mongolia wolf 2 and Tibetan Mastiff
Indian wolf	Inner Mongolia wolf 3	Tibetan Mastiff	Israeli jackal	golden	0.043034004	0.003304099	13.02442839	Inner Mongolia wolf 3 and Tibetan Mastiff
Indian wolf	Inner Mongolia wolf 4	Tibetan Mastiff	Israeli jackal	golden	0.038861655	0.003428659	11.33435907	Inner Mongolia wolf 4 and Tibetan Mastiff
Indian wolf	Xinjiang wolf 1	Tibetan Mastiff	Israeli jackal	golden	0.031537035	0.003554942	8.87132091	Xinjiang wolf 1 and Tibetan Mastiff
Indian wolf	Xinjiang wolf 2	Tibetan Mastiff	Israeli jackal	golden	0.035978996	0.003489597	10.31035705	Xinjiang wolf 2 and Tibetan Mastiff
Indian wolf	Qinghai wolf 1	Tibetan Mastiff	Israeli jackal	golden	0.012472483	0.0042495	2.935046977	None significant
Indian wolf	Qinghai wolf 2	Tibetan Mastiff	Israeli jackal	golden	0.000181535	0.005016359	0.036188541	None significant
Indian wolf	Tibetan wolf 1	Tibetan Mastiff	Israeli jackal	golden	-0.015945107	0.005496804	-2.900795844	None significant
Indian wolf	Tibetan wolf 2	Tibetan Mastiff	Israeli jackal	golden	-0.01111313	0.005607364	-1.981881304	None significant
Indian wolf	Central Russian wolf	Tibetan Mastiff	Israeli jackal	golden	0.052482409	0.003741891	14.02563643	Central Russian wolf and Tibetan Mastiff
Indian wolf	Eastern Russian wolf	Tibetan Mastiff	Israeli jackal	golden	0.063504238	0.003970299	15.99482625	Eastern Russian wolf and Tibetan Mastiff
Inner Mongolia wolf 1	Indian wolf	Dingo	Israeli jackal	golden	-0.054326806	0.004865246	-11.16630128	Inner Mongolia wolf 1 and Dingo
Inner Mongolia wolf 2	Indian wolf	Dingo	Israeli jackal	golden	-0.07473861	0.004560311	-16.38892995	Inner Mongolia wolf 2 and Dingo
Inner Mongolia wolf 3	Indian wolf	Dingo	Israeli jackal	golden	-0.05140046	0.004328838	-11.873963	Inner Mongolia wolf 3 and Dingo
Inner Mongolia wolf 4	Indian wolf	Dingo	Israeli jackal	golden	-0.054162027	0.003620937	-14.95801595	Inner Mongolia wolf 4 and Dingo
Xinjiang wolf 1	Indian wolf	Dingo	Israeli jackal	golden	-0.036348833	0.004431507	-8.202363853	Xinjiang wolf 1 and Dingo
Xinjiang wolf 2	Indian wolf	Dingo	Israeli jackal	golden	-0.041169464	0.003370556	-12.21444375	Xinjiang wolf 2 and Dingo
Qinghai wolf 1	Indian wolf	Dingo	Israeli jackal	golden	-0.029562247	0.004240235	-6.971841437	Qinghai wolf 1 and Dingo

Qinghai wolf 2	Indian wolf	Dingo	Israeli jackal	golden	-0.019878602	0.005208526	-3.816550238	Qinghai wolf 2 and Dingo
Tibetan wolf 1	Indian wolf	Dingo	Israeli jackal	golden	-0.001177835	0.005481328	-0.21488122	None significant
Tibetan wolf 2	Indian wolf	Dingo	Israeli jackal	golden	-0.004782687	0.005538752	-0.863495446	None significant
Central Russian wolf	Indian wolf	Dingo	Israeli jackal	golden	-0.052738579	0.004812766	-10.95805956	Central Russian wolf and Dingo
Eastern Russian wolf	Indian wolf	Dingo	Israeli jackal	golden	-0.065875016	0.004195071	-15.70295772	Eastern Russian wolf and Dingo
Indian wolf	Inner Mongolia wolf 1	Dingo	Israeli jackal	golden	0.053689858	0.004852218	11.0650132	Inner Mongolia wolf 1 and Dingo
Indian wolf	Inner Mongolia wolf 2	Dingo	Israeli jackal	golden	0.075334549	0.004593578	16.39997143	Inner Mongolia wolf 2 and Dingo
Indian wolf	Inner Mongolia wolf 3	Dingo	Israeli jackal	golden	0.050619804	0.004272268	11.84846251	Inner Mongolia wolf 3 and Dingo
Indian wolf	Inner Mongolia wolf 4	Dingo	Israeli jackal	golden	0.052022774	0.003734972	13.92855647	Inner Mongolia wolf 4 and Dingo
Indian wolf	Xinjiang wolf 1	Dingo	Israeli jackal	golden	0.03618321	0.004487441	8.063216634	Xinjiang wolf 1 and Dingo
Indian wolf	Xinjiang wolf 2	Dingo	Israeli jackal	golden	0.039244259	0.003290093	11.92800771	Xinjiang wolf 2 and Dingo
Indian wolf	Qinghai wolf 1	Dingo	Israeli jackal	golden	0.029097784	0.00424551	6.853778309	Qinghai wolf 1 and Dingo
Indian wolf	Qinghai wolf 2	Dingo	Israeli jackal	golden	0.018736098	0.005256195	3.564574532	Qinghai wolf 2 and Dingo
Indian wolf	Tibetan wolf 1	Dingo	Israeli jackal	golden	0.000325263	0.005539451	0.058717477	None significant
Indian wolf	Tibetan wolf 2	Dingo	Israeli jackal	golden	0.007079563	0.005541481	1.277558081	None significant
Indian wolf	Central Russian wolf	Dingo	Israeli jackal	golden	0.052137127	0.004800123	10.86162326	Central Russian wolf and Dingo
Indian wolf	Eastern Russian wolf	Dingo	Israeli jackal	golden	0.066808418	0.004172548	16.01142051	Eastern Russian wolf and Dingo
Israeli wolf	Indian wolf	Basenji	Israeli jackal	golden	-0.074611422	0.005945534	-12.54915392	Israeli wolf and Basenji
Iranian wolf	Indian wolf	Basenji	Israeli jackal	golden	-0.005700737	0.004100655	-1.390201675	None significant
Croatian wolf	Indian wolf	Basenji	Israeli jackal	golden	-0.022189273	0.003629812	-6.113064883	Croatian wolf and Basenji
Western Russian wolf	Indian wolf	Basenji	Israeli jackal	golden	-0.043594786	0.005301261	-8.223475165	Western Russian wolf and Basenji

Italian wolf	Indian wolf	Basenji	Israeli jackal	golden	-0.021235173	0.00391942	-5.417938046	Italian wolf and Basenji
Portuguese wolf	Indian wolf	Basenji	Israeli jackal	golden	-0.017167963	0.00401505	-4.275902874	Portuguese wolf and Basenji
Spanish wolf	Indian wolf	Basenji	Israeli jackal	golden	-0.06558685	0.005757246	-11.39205369	Spanish wolf and Basenji
Indian wolf	Israeli wolf	Basenji	Israeli jackal	golden	0.072787624	0.005877721	12.38364786	Israeli wolf and Basenji
Indian wolf	Iranian wolf	Basenji	Israeli jackal	golden	0.00506298	0.00414127	1.222567001	None significant
Indian wolf	Croatian wolf	Basenji	Israeli jackal	golden	0.022072956	0.00362086	6.096053581	Croatian wolf and Basenji
Indian wolf	Western Russian wolf	Basenji	Israeli jackal	golden	0.043898635	0.005194967	8.450224512	Western Russian wolf and Basenji
Indian wolf	Italian wolf	Basenji	Israeli jackal	golden	0.020306015	0.003956135	5.132791487	Italian wolf and Basenji
Indian wolf	Portuguese wolf	Basenji	Israeli jackal	golden	0.017102765	0.00407328	4.198769881	Portuguese wolf and Basenji
Indian wolf	Spanish wolf	Basenji	Israeli jackal	golden	0.067399247	0.005713216	11.79707597	Spanish wolf and Basenji
Israeli wolf	Indian wolf	Boxer	Israeli jackal	golden	-0.094261019	0.006314451	-14.92782492	Israeli wolf and Boxer
Iranian wolf	Indian wolf	Boxer	Israeli jackal	golden	-0.005646785	0.003966797	-1.423512629	None significant
Croatian wolf	Indian wolf	Boxer	Israeli jackal	golden	-0.076276273	0.004061924	-18.77835822	Croatian wolf and Boxer
Western Russian wolf	Indian wolf	Boxer	Israeli jackal	golden	-0.101832465	0.00539321	-18.8816055	Western Russian wolf and Boxer
Italian wolf	Indian wolf	Boxer	Israeli jackal	golden	-0.064488295	0.004162783	-15.49163008	Italian wolf and Boxer
Portuguese wolf	Indian wolf	Boxer	Israeli jackal	golden	-0.040015442	0.004271833	-9.36727692	Portuguese wolf and Boxer
Spanish wolf	Indian wolf	Boxer	Israeli jackal	golden	-0.112290148	0.007070975	-15.88043351	Spanish wolf and Boxer
Indian wolf	Israeli wolf	Boxer	Israeli jackal	golden	0.095127507	0.006190939	15.36560288	Israeli wolf and Boxer
Indian wolf	Iranian wolf	Boxer	Israeli jackal	golden	0.005896425	0.003911854	1.50732259	None significant
Indian wolf	Croatian wolf	Boxer	Israeli jackal	golden	0.073995784	0.004101091	18.04295149	Croatian wolf and Boxer
Indian wolf	Western Russian wolf	Boxer	Israeli jackal	golden	0.099491546	0.005446134	18.26828743	Western Russian wolf and Boxer

Indian wolf	Italian wolf	Boxer	Israeli jackal	golden	0.066369181	0.004167535	15.92528493	Italian wolf and Boxer
Indian wolf	Portuguese wolf	Boxer	Israeli jackal	golden	0.038985292	0.004291769	9.083734909	Portuguese wolf and Boxer
Indian wolf	Spanish wolf	Boxer	Israeli jackal	golden	0.112129584	0.007039953	15.92760465	Spanish wolf and Boxer

Table S4. The proportion dog ancestry in the genomes of Old World wolves. Only the wolves showed significant gene flow with dogs in ABBA-BABA tests were included. For Asian wolf, we estimated the proportion of Chinese indigenous dogs ancestry in their genomes. For the European and Middle Eastern wolf, we estimated the proportion of basenji and boxer ancestry in their genomes.

Asian wolf	Proportion of Chinese indigenous dog 1 ancestry (%)	Proportion of Chinese indigenous dog 2 ancestry (%)	Proportion of Chinese indigenous dog 3 ancestry (%)	Averaged proportion (%)
Inner Mongolia wolf 1	9.04	12.06	11.47	10.86
Inner Mongolia wolf 2	22.34	20.85	21.60	21.59
Inner Mongolia wolf 3	11.57	14.72	13.18	13.16
Inner Mongolia wolf 4	10.43	13.15	12.59	12.06
Xinjiang wolf 1	9.28	11.30	9.58	10.05
Xinjiang wolf 2	9.31	10.80	10.51	10.21
Central Russian wolf	17.38	15.30	16.26	16.31
Eastern Russian wolf	19.52	18.84	18.84	19.07

European and Middle Eastern wolf	Proportion of basenji and boxer ancestry (%)
Israeli wolf	23.54
Croatian wolf	13.76
Western Russian wolf	20.37
Italian wolf	12.41
Portuguese wolf	7.97
Spanish wolf	25.32

Table S5 Population sizes inferred from G-PhoCS.

Population sizes (individuals)	
$\theta_{\text{New World-Old world}}$	45,100 (44,400 – 45,900) [1]
$\theta_{\text{Wolf-Dog}}$	8,000 (3,400 – 16,100) [1]
$\theta_{\text{New World}}$	17,300 (13,000-21,700) [1]
$\theta_{\text{Tibetan wolf}}$	2,500 (2,300-2,700) [2]
$\theta_{\text{Inner Mongolian wolf}}$	9,400 (8,400-10,600) [2]
$\theta_{\text{Xinjiang Wolf}}$	20,800 (17,500-24,600) [3]
$\theta_{\text{Qinghai wolf}}$	93,700 (75,300 – 116,300) [2]
$\theta_{\text{Russian wolf}}$	13,500 (11,400 – 15,700) [2]
$\theta_{\text{Croatian wolf}}$	4,600 (3,900-5,300) [4]
$\theta_{\text{Israeli wolf}}$	16,600 (12,700 – 21,900) [4]
$\theta_{\text{Indian wolf-Iranian wolf}}$	6,200 (5,400-7,000) [4]
$\theta_{\text{Mexican wolf}}$	600 (400-700) [1]
$\theta_{\text{Yellowstone wolf}}$	3,500 (2,600-4,300) [1]
θ_{Dog}	2,000 (1,400 – 2,700) [1]
θ_{Basenji}	1,600 (1,500 – 1,800) [1]
θ_{Dingo}	1,000 (900 – 1,100) [1]
$\theta_{\text{Chinese indigenous dog}}$	26,100 (18,600 – 36,700) [1]

[1] estimated in 'global' run of G-PhoCS
[2] estimated in 'asian' run of G-PhoCS with Inner Mongolian wolf
[3] estimated in 'asian run of G-PhoCS with Xinjiang wolf
[4] estimated in 'european' run of G-PhoCS

Table S6. Samples used in the demographic inference of G-PhoCS

Name	Population symbol	Runs used
Tibetan wolf 1	TIW	Global, Asian1, Asian2
Tibetan wolf 2	TIW	Asian1, Asian2
Inner Mongolia wolf 3	IMW	Global, Asian1
Inner Mongolia wolf 4	IMW	Asian1
Xinjiang wolf 1	XJW	Asian2
Xinjiang wolf 2	XJW	Asian2
Qinghai wolf 1	QHW	Asian1, Asian2
Qinghai wolf 2	QHW	Asian1, Asian2
Central Russian wolf	RUW	Asian1, Asian2
Western Russian wolf	RUW	Asian1, Asian2
Indian wolf	INW	European
Iranian wolf	IRW	European
Israeli wolf	ISW	Global, European
Croatian wolf	CRW	Global, European
Mexican wolf	MXW	Global
Yellow stone wolf 1	YSW	Global
Chinese indigenous dog 1	CHD	Global, Asian1, Asian2
Chinese indigenous dog 2	CHD	Asian1, Asian2
Tibetan mastiff	TMD	Global [1]
Basenji	BAS	Global, European
Dingo	DIN	Global, Asian1, Asian2
Israeli golden jackal	JAC	Global, Asian1, Asian2, European

[1] The Tibetan mastiff genome was analyzed in a separate global run, where it replaced the Chinese indigenous dog genome

Supplementary Methods

1. Samples and data

We analyzed sequences from 24 wolves, three ancient breed dogs (one Tibetan mastiff, one basenji and one dingo), three Chinese indigenous dogs, and the boxer as a modern breed. We used three outgroup species (one coyote, one Kenya golden jackal and one Israeli golden jackal). Within the Old World wolves, five were from Europe (Spain, Croatian, Italy, Portugal and Russia), two were from the Middle East (Iran and Israeli), 13 were from Asia (China (10), India (1), and Russia (2)) (Fi. 1; Supplemental Table S1). The three Russian wolves were from the Far East Asia (Chukotka), Central Asia (Altai), and Europe (Bryansk). The three Chinese indigenous dogs were also from different locations: Xi'an (Central Asia); Si'mao (close to Laos and Vietnam); and Ya'an (the eastern edge of the Tibetan Plateau) (Fig. 1). There were four New World wolves, including one Mexican wolf and three Yellowstone National Park wolves representing parents and an offspring. We only included the Yellowstone wolf with the highest genome coverage (Yellowstone wolf 1, the father) in some downstream analyses, such as phylogenetic tree.

Genomic DNA was extracted from whole blood using the standard phenol-chloroform method. The whole genome sequencing was performed using an Illumina Hiseq 2000 at Beijing Genomics Institute (BGI). For each sample, two paired-end libraries with insert sizes of ~ 300 to 500 bp were generated. Library preparation and all sequencing runs were performed according to manufacturer's protocols. Some individuals were sequenced and reported in previous studies (Supplemental Table S1; Wang et al. 2013; Freedman et al. 2014; Zhang et al. 2014). We downloaded the raw short reads of the three Russian wolves, three Chinese indigenous dogs, one Tibetan mastiff, and nine Chinese wolves (Wang et al. 2013; Zhang et al. 2014), and then processed these reads with our genotype pipeline together with the sequences new to

this study (Fig. 1, Supplemental Table S1). Freedman et al. (2014) sequenced three wolves (Croatian wolf, Israeli wolf and Chinese wolf), two dogs (basenji and dingo), and an Israeli golden jackal. For these published genomes, we obtained the genotype files as Variant Call Format (VCF) from the authors and combined the VCF files with genotype calls.

2. Post-genotype filters

We applied a series of data quality filters to improve the quality of genotype calls. These filters were designed to minimize the errors from sequencing and alignment, and to exclude regions exhibiting accelerated evolutionary rates that are not caused by positive selection, but reflect a high mutation rate (Fan et al. 2014; Freedman et al. 2014; Zhang et al. 2014). We used two levels of filters, the Genome Filters (GF) and Sample Filters (SF). The GF is based on the features of the reference genome and polymorphism across all the samples, whereas the SF is based on the genotype results of each independent sample. Thus, high quality sites in each sample should pass both GF and the corresponding SF and were the only ones used in the following analyses. Details of these filters were described in our previous studies (Fan et al. 2014; Freedman et al. 2014; Zhang et al. 2014).

3. Detection of gene flow using the *D*-statistic

We applied the ABBA-BABA test (*D*-statistic) between closely related populations by detecting differences in allele sharing between two lineages (P1 and P2) with a third lineage (P3) (Durand et al. 2011). Given an outgroup (O), two allelic configurations of P1-P2-P3-O are informative of gene flow between P3 with either P1 or P2: ABBA (P1 and O share the same allele A, while P2 and P3 share the alternative allele B) and BABA (P1 and P3 share the alternative allele B, while P2 and O share the allele A). The null hypothesis states that the genome-wide frequencies of these two configurations should be approximately equal in the absence of lineage-specific post-divergence gene flow.

Rejection of that null hypothesis implies gene flow between P3 and either P1 or P2. We quantified deviations from this null expectation using the *D*-statistic:

$$D = \frac{\sum_{i=1}^{n} C_{ABBA}(i) - \sum_{i=1}^{n} C_{BABA}(i)}{\sum_{i=1}^{n} C_{ABBA}(i) + \sum_{i=1}^{n} C_{BABA}(i)}$$

Here, $C_{ABBA}(i)$ and $C_{BABA}(i)$ take the value of 0 or 1 depending on the absence or presence of an ABBA or BABA allele configuration at the i^{th} site. For each comparison, we calculated the *D* statistic in 5Mb windows along the genome, considering only sites passing genome and sample filters and randomly selecting one allele from each genotype for each site. We estimated the standard error of the *D*-statistic with a jackknife procedure as done in Durand et al. (2011). We calculated the Z-score by dividing the value of the *D* statistic by its standard error. Z-scores with absolute values ≥3 were considered significant evidence of gene flow between the P3 and one of the two lineages P1 or P2 (P1 for negative Z-scores, P2 for positive values).

In this study, we focused on gene flow between Old World wolves and the closest dog populations in our dataset. Consequently, we tested whether Asian wolves had gene flow with the Chinese indigenous dogs, Tibetan mastiff and dingo. For the European and Middle Eastern wolves, we tested whether these populations had gene flow with the basenji and boxer. We also tested whether Mexican wolf and Yellowstone wolf had gene flow with Chinese indigenous dogs, boxer and dingo. In the calculations for all the pairwise combinations, we assigned the above wolves as P1 and the dogs as P3. We used the Israeli golden jackal as outgroup in all the runs. The Indian wolf was used as P2 in all the Old World runs. For the New World wolf runs, we set the Mexican wolf and Yellowstone wolf as P1 and P2 and then switched them. Since domestic dogs are closer to Old World wolves than to New World wolves we did not use the Indian wolf as P2

because of potential bias due to higher allele sharing between the Indian wolves and domestic dogs.

Under the assumption of one gene flow event that is recent compared to the divergence of dogs and wolves, we further used the Durand et al. (2011) equation to estimate the proportion of dog ancestry in the wolf genomes. The original equation was applied to estimate the proportion of Neanderthal ancestry in non-Africans (Green et al. 2010):

$$\text{Dog ancestry proportion in wolf} = \frac{S(\text{wolf1,wolf2,dog1,jackal})}{S(\text{wolf1,dog1,dog2,jackal})}$$

The S statistics were the average value of the different combinations of wolf 1 and wolf 2 samples. Here, wolf 1 is always the Indian wolf, and wolf 2 is the wolf population that had gene flow with dog detected from above ABBA-BABA runs. For the Asian wolf, we used three Chinese indigenous dogs as dog 1 and dog 2. For the European and Middle Eastern wolves, boxer and basenji were used as dog 1 and dog 2 as the latter represents an ancient African-Middle East lineage, and the former represents the lineage leading to modern European breeds. Additional dog references might increase the number of observed admixture events, but this would be computationally challenging to undertake.

4. Inference of population size changes through time with PSMC

In order to validate the confidence in PSMC findings, we ran 100 bootstrap replicates for each genome. To sample a bootstrap replicate, we divided the genome into segments of 5Mb, sampled with replacement from those segments until we obtained a sequence with approximately the same length as the original genome defined by using the “-b” option in the PSMC software, and re-ran the EM-based effective population size estimation procedure.

5. Demographic inference with G-PhoCS

Neutral loci were selected to be short (1 kb) interspersed (> 30 kb apart) genomic segments that are >10 kb from protein coding genes and avoiding regions with low map ability, high sequencing error rate, or CpG dinucleotides (see Freedman et al. 2014 for more details).

We assumed an exponential distribution with mean of 0.0001 for the mutation-scaled population size (ϴ) and divergence time (τ) parameters, and a Gamma (α= 0.002, β= 0.00001) prior for migration rate parameters (*m*). The Markov Chains exploring the space of parameter values were ran for 75,000 burn-in iterations and an additional 125,000 iterations, in which values of the model parameters were sampled every 50 iterations, resulting in a total of 2,501 samples from the approximate posterior distribution. For each parameter, we recorded the mean sampled value and the 95% Bayesian credible interval (CI). Population size estimates (N_e) were obtained from the mutation-scaled samples (ϴ) by assuming a mutation rate per generation of 1.0×10^{-8}, and divergence times were calibrated by assuming the same rate and an average generation time of three years. We also examined the influence of uncertainty on mutation rates on timing of key events (Skogland et al. 2015). Gene flow was measured by the total migration rate, which is the per-generation rate times the number of generations in which migration was allowed.

Given the large number of sequences and computational limits, we ran separate analyses on different subsets of sequences, which generated separate inferences that we then integrated into a unified demographic history. In each run, we assumed a population phylogeny consistent with the genome-wide ML tree (Fig. 3), and augmented this tree with various migration bands to model gene flow. To obtain a high level view of global history, we analyzed a subset of six wolf genomes from Europe, the Middle East, East Asia and North America, together with three dog genomes and the golden jackal outgroup (Supplemental Table S6). In this case, we considered also an alternative structure to the population phylogeny, in which dogs are an outgroup to all wolf

populations, in addition to the scenario implied by the genome-wide tree, in which New World wolves branched before dogs. To obtain population-specific information on effective population size and migration rates, we ran additional analyses focusing on different geographic regions. One analysis considered the four European and Middle Eastern wolves, basenji and dingo representing dogs, and the Israeli golden jackal as an outgroup. Two additional analyses were done where each considered eight East Asian wolf samples from four populations, the dingo, two Chinese indigenous dogs, and the golden jackal. The two runs differ in the samples chosen to represent lowland Chinese wolves. In all runs, we allowed gene flow between the golden jackal population and all other sampled wolf and dog populations as well as the population ancestral to all dogs and wolves. In the three local runs we allowed gene flow between all sampled dog populations and all sampled wolf populations, but not within wolf populations. In the global run we allowed gene flow between basenji and the Eurasian and Middle Eastern wolves and between dingo and Chinese indigenous dogs and the East Asian wolves. In addition, we modeled gene flow between the two East Asian wolf populations and the two West Eurasian wolf populations.

Results

1. Alignments, PCR duplicates, coverage and genotyping accuracy

The alignments were done in Bowtie2. PCR duplicates were marked in Picard, and from 4.2% to 64.35% of the reads (mean: 15.8%) represented PCR duplicates that were excluded from downstream analyses. The Italian wolf, Yellowstone wolf 3 and Israeli golden jackal had > 50% PCR duplicates, whereas the rest of the genomes had much lower rates (Supplemental Table S2). After running our genotyping pipeline (Fan et al. 2014; Zhang et al. 2014), the average coverage was 21-fold with most sequences (75%) having more than 20-fold effective coverage (Supplemental Table S1). Of the 24 gray

wolves, the Italian wolf, Yellowstone wolf 3, Inner Mongolia wolf 2, and the three Russian wolves had lower than 20-fold coverage. As sequencing was not done at the same time or by the same groups, we assessed the potential for batch effects by examining discordance in geographic clusters using principal component analysis (PCA; see Fig. 4) and more directly by comparing the genotype calls from the sequence data with those using the Illumina CanineHD BeadChip. We did not find clusters in the PCA suggesting batch effects such as a grouping of the sequences from Wang et al. (2013)(Fig. 4). The genotyping accuracy of the new genomes generated in this study was assessed with Illumina CanineHD BeadChip for the Indian wolf and Portugal wolf. Both wolves showed > 99.85% concordance. In addition, another three wolves (Israeli wolf, Croatian wolf, and Inner Mongolia wolf 1) and two dogs (basenji and dingo) were compared previously to the Illumina CanineHD BeadChip, and showed > 99.6% concordance (Freedman et al. 2014).

2. Useable sites

All the high coverage (> 20-fold) individuals had > 1.2 billion total useable sites, which covered more than 60% of the reference genome. The numbers of total SNPs varied between different canids. As expected, the number of SNPs increased with divergence from the boxer reference with dogs having the fewest at ~ 2 million SNPs (from 1,744,052 to 2,212,090, average: 2,000,320, also reflecting the domestication bottleneck, Freedman et al. 2014), wolves having ~ 3 million SNPs (from 1,986,246 to 3,595,813, average: 3,033,017) and the three outgroups having > 4 million SNPs (from 4,240,145 to 4,752,396, average: 4,427,268) (Supplemental Table S2).

3. Heterozygosity

We only used SNPs to calculate heterozygosity for the following reasons: 1) Misalignment is possible when short reads containing novel CNVs are mapped to the reference genome and can lead to false SNP calls; and 2) Short reads are prone to misalignment near indels and the local realignment around indels in our genotyping

pipeline may not fully fix this problem, thus we excluded any SNPs near indels (5bp, either up or downstream).

The lower coverage genomes, especially those with lower than 12-fold coverage (Inner Mongolia wolf 2, Russian wolf (east), Yellowstone wolf 3, and Chinese indigenous dog 3, had very high heterozygosity (Supplemental Figure S5) possibly due to more genotyping error at heterozygous sites. Consequently, we assessed the heterozygosity of exons and neutral regions only for the samples with > 20-fold genome coverage (Supplemental Figure S6). The pattern of heterozygosity of exons and neutral regions is consistent with the pattern of genome wide heterozygosity in the samples. The samples having higher genome wide heterozygosity tended to have higher heterozygosity in exons and neutral regions (Supplemental Figure S6). However, the neutral regions had the highest heterozygosity, whereas the exonic regions had consistently lower heterozygosity in all the samples (Supplemental Figure S6) consistent with the action of purifying selection. The heterozygosity of exonic regions is only 41.6% to 52.6% of the corresponding genome wide heterozygosity.

4. PCA of different geographical regions

The PCA for Asia of this dataset showed that the highland Chinese wolves were the most distinct populations (Supplemental Figure S9a and S9c). For Europe, dogs and wolves were separated on PC1, and Spanish wolf and Portugal wolf were separated from the Croatian wolf, Italian wolf and Western Russian wolf on PC2 (Supplemental Figure S10a). For the Middle Eastern wolf samples, PC1 separated the dogs and wolves, and then PC2 separated the Israeli wolf from the Indian wolf and Iranian wolf (Supplemental Figure S10c). For North American wolves, dogs and wolves separated from each other on PC1 (Supplemental Figure S11).

5. Demographic inference with G-PhoCS

Notable are the very high values of N_e inferred for Qinghai wolf (93,700 individuals), which is also suggested by the peak in their PSMC plot (Fig. 5) and might in part, reflect

ancestral population structure. The Israeli wolf had the largest N_e within the European-Middle Eastern wolves (16,600 individuals), which was 3.6-fold and 2.7-fold higher than Croatian wolf (4,600 individuals) and the ancestral population of Iranian and Indian wolf (6,200 individuals), respectively.

Supplementary Material for Paper V - The Effects of Population Structure and Sampling Scheme on Demographic Inferences from Microsatellite Data: an Empirical Test on the Iberian Wolf Population

Table S1: Signals of population bottlenecks or expansions in the simulated data with varying migration rates (*m*). Numbers indicate the number of populations inferred to have experienced a bottleneck or an expansion out of 50 or out of 5 simulated populations, for local and pooled samples, respectively. Two different sampling schemes were used for both the local and pooled samples: either 20 individuals or all 50 individuals of each population were sampled.

sampling scheme	**sample size**	**heterozygote excess (bottleneck)**				**heterozygote deficiency (expansion)**			
		***m*=0**	***m*=0.015**	***m*=0.03**	***m*=0.1**	***m*=0**	***m*=0.015**	***m*=0.03**	***m*=0.1**
local	20	7/50	10/50	4/50	2/50	0/50	1/50	1/50	1/50
	50	6/50	10/50	6/50	3/50	2/50	2/50	2/50	2/50
pooled	200 (20 each)	0/5	0/5	0/5	0/5	5/5	3/5	3/5	1/5
	500 (50 each)	0/5	0/5	0/5	0/5	5/5	3/5	3/5	1/5

Table S2: M ratio tests applied to populations simulated in an island model with different rates of migration (*m*) and sampling schemes. Two different sampling schemes were used for both the local and pooled samples: either 20 individuals or all 50 individuals of each population were sampled. The known effective sizes of simulated populations and metapopulations (N_e=50 and Ne=500, respectively) were used to calculate M_C values, also taking into account the different sample sizes. *M* values refer to average values in 10 simulation replicates. The number of populations (out of 50 for the local sampling, or out of 5 for the pooled sampling) where $M < M_C$, is indicated in parenthesis.

sampling scheme	**sample size**	***M*$_C$**	**average *M* (# *M*<*M*$_C$ / total)**			
			***m*=0**	***m*=0.015**	***m*=0.03**	***m*=0.1**
local	20	0.9819	0.9994 (0/50)	0.9507 (48/50)	0.9629 (48/50)	0.9677 (44/50)
	50	0.9855	0.9995 (0/50)	0.9558 (48/50)	0.9668 (48/50)	0.9734 (41/50)
pooled	200 (20 each)	0.9596	0.8397 (5/5)	0.9834 (0/5)	0.9829 (0/5)	0.9869 (0/5)
	500 (50 each)	0.9656	0.8459 (5/5)	0.9875 (0/5)	0.9859 (0/5)	0.9900 (0/5)

Table S3: Demographic parameters inferred by MSVAR for the total Iberian Peninsula sample (n=218).

population size change model	**current N$_e$ (*N0*)**	**past N$_e$ (*N1*)**	**time since size change (*xa*)**
exponential	55 (18-179)	14 348 (4 375-44 525)	504 (166-1 637)
linear	23 (7-73)	8 128 (2 704-26 375)	3 599 (1 248-11 995)

Table S4: Demographic parameters inferred by MSVAR for four subpopulations.

population	**population size change model**	**current N_e (*N0*)**	**past N_e (*N1*)**	**time since size change (*xa*)**
Alto Minho	exponential	1281 (201-9842)	911 (207-4213)	74 (9-611)
Castilla y León	exponential	912 (145-7029)	1283 (296-5702)	71 (8-489)
S Douro	exponential	1547 (207-12806)	695 (134-3714)	94 (10-840)
W Galicia	exponential	*poor chain mixing*		

Table S5: Demographic parameter estimates inferred by MSVAR on simulated data with varying migration rates (m).

migration rate	**current N_e (*N0*)**	**past N_e (*N1*)**	**time since size change (*xa*)**
m=0.015	1920 (339-10359)	652 (192-2248)	47 (6-292)
m=0.03	1230 (222-6879)	777 (212-2804)	59 (8-585)
m=0.1	1773 (331-9300)	817 (232-2597)	38 (6-237)