

MSC

2.º
CICLO

FCUP
2019

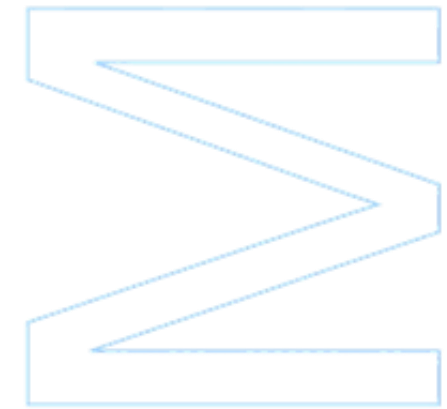


PseudoChecker: An integrated online platform for
automatised gene loss inference in mammals



PseudoChecker: An integrated online platform for automatised gene loss inference in mammals

Luís Pedro Palmares Quádrio Alves
Dissertação de Mestrado apresentada à
Faculdade de Ciências da Universidade do Porto em
Bioinformática e Biologia Computacional
2019



PseudoChecker: An integrated online platform for automatised gene loss inference in mammals

Luís Pedro Palmares Quádrio Alves

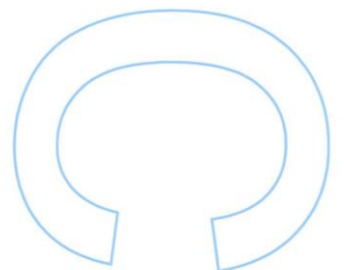
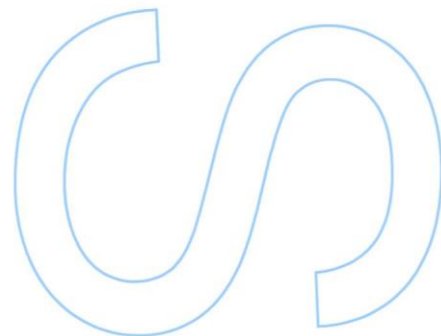
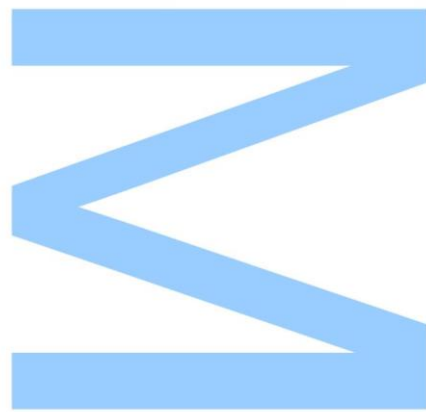
Mestrado em Bioinformática e Biologia Computacional
2019

Orientador

Luís Filipe Costa de Castro, Professor Auxiliar Convidado, Faculdade de Ciências da Universidade do Porto

Coorientador

Pedro Manuel Pinto Ribeiro, Professor Auxiliar, Faculdade de Ciências da Universidade do Porto

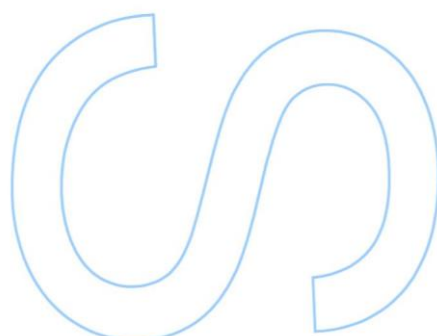
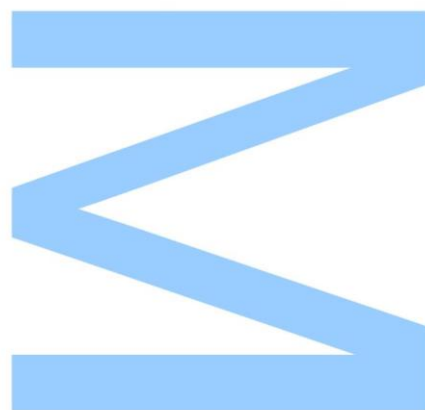




Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, ____/____/____



Author statement

The present work included in this thesis was totally executed by the candidate with close revisions by the supervising team. Leveraged by the development of the hereby explored bioinformatics tool, within the same study scope, two articles were published (first author or joint first authors are underlined):

The dopamine receptor D5 gene shows signs of independent erosion in toothed and baleen whales

PeerJ, August 2019 (Accepted for publication)

Luís Q. Alves, Juliana Alves, Rodrigo Ribeiro, Raquel Ruivo, L. Filipe C. Castro

Convergent inactivation of the skin-specific C-C motif chemokine ligand 27 in mammalian evolution

Immunogenetics, May 2019

Mónica Lopes-Marques, Luís Q. Alves, Miguel M. Fonseca, Giulia Secci-Petretto, André M. Machado, Raquel Ruivo, L. Filipe C. Castro

During the MSc work program, the candidate actively participated in other research projects, which entailed two additional publications not referenced here (check *curriculum vitae*).

Abstract

Gene duplication bursts and amino acid substitutions have often been assigned to the evolution of phenotypic traits. Much less explored is the role of gene loss as an engine of evolutionary change, either from adaptive or regressive perspectives. Remarkably, mammals provide a reference test case to address the role and importance of gene loss as a major driver in evolution. They successfully colonised a large range of habitats, including land, water, air and display tantalising phenotypic adaptations to their surrounding environment. Paradoxically, the study of gene loss across mammalian lineages is still in its early infancy, and only a few years ago, gene loss studies highlighting this *taxon* started to emerge.

The rapid increase of high-quality genome assemblies requires the introduction of automated methods capable of leveraging their comprehensive study. Particularly, for addressing and promoting the study of gene loss as a source of mammalian evolutionary innovation, a systematic, highly accurate, and computational automatised approach is ultimately required.

With this thesis culminated the development of *PseudoChecker*: the first integrated online platform for automatised gene loss inference in mammals. Amongst the few existing methods, our comparative genomics-based approach comprehends the unique displaying full automation, built-in inclusion of a graphical user interface, an index of pseudogenisation - the PseudoIndex, and finally, but not least important, displaying its full integration as a web service, not only allowing to keep *PseudoChecker* as a multi-platform and constantly updated bioinformatics tool, but facilitating its expedite and easy usability for the quick identification of gene loss. In a short amount of time, by scrutinising published lost genes in Mammalia and presumed viable genes in the same group of species, using human as reference species, on average, *PseudoChecker* was able to correctly capture over 95% of the tested mammalian gene functionality events and over 97% of the analysed gene loss events.

Importantly, *PseudoChecker*, currently in Version 0.8, still constitutes a prototype and a set of improvements, including support for polymorphic loss (population analysis), and built-in inclusion of selection analyses can be employed. Moreover, on the technical side, the implementation of multi-core processing, as well as code rewriting in more efficient, mid-level compiled languages, would also benefit from speed improvements. Finally, with minor modifications and adequate testing, the ultimate natural evolution of *PseudoChecker* passes by its expandability to non-mammalian animal lineages.

Keywords: Gene Loss, Unitary Pseudogenes, Bioinformatics, Mammals

Resumo

Eventos de duplicação de genes e substituições aminoacídicas têm sido frequentemente reconhecidos como mecanismos subjacentes à evolução de traços fenotípicos. Muito menos explorado é o papel da perda de genes como um mecanismo de mudança evolutiva, quer segundo a perspectiva adaptativa ou regressiva. De forma notável, os mamíferos constituem uma boa referência para abordar o papel e a importância da perda de genes como um dos principais condicionadores evolutivos. Estes colonizaram com sucesso uma grande variedade de habitats, incluindo terra, água, ar e exibem magníficas adaptações fenotípicas ao ambiente circundante. Paradoxalmente, o estudo da perda de genes sobre esta linhagem ainda está no seu início e apenas há alguns anos, estudos do efeito da perda de genes sobre este táxon começaram a surgir.

A rápida produção de genomas de elevada qualidade requer a introdução de metodologias adequadas capazes de alavancar abrangentemente o seu estudo. Em particular, para abordar e promover o estudo de perda de genes como fonte de inovação evolutiva em mamíferos, é necessário desenvolver uma metodologia computacional automatizada, sistemática e de elevada precisão.

A presente dissertação culminou no desenvolvimento do *PseudoChecker*: a primeira plataforma online integrada para a inferência automatizada de perda de genes em mamíferos. De entre os poucos métodos existentes, a nossa abordagem baseada em genómica comparativa é, entre todos, a única que possui automatização total, inclusão de raiz de uma interface gráfica e índice de pseudogenização - o *PseudoIndex*, e, finalmente, é a única que se encontra totalmente integrada num serviço web, permitindo não só manter o *PseudoChecker* como uma ferramenta de bioinformática constantemente atualizada e multi plataforma, mas também facilitando a sua rápida e fácil utilização, que se traduz na rápida identificação de eventos de perda de genes.

Tendo sido examinados genes publicados como perdidos em Mammalia, assim como genes presumivelmente funcionais no mesmo grupo evolutivo, em média, e em pouco tempo, o *PseudoChecker* foi capaz de identificar corretamente mais de 95% dos eventos testados de funcionalidade genética e mais de 97% dos eventos testados de perda de genes.

No entanto, é importante ressaltar que o *PseudoChecker*, atualmente na Versão 0.8, ainda constitui um protótipo, sendo que um conjunto de melhorias podem ser aplicadas, nomeadamente o suporte para o estudo de perda polimórfica de genes (análise populacional) e suporte embutido para análises de seleção sobre as sequências em análise. Para além disso, no que diz respeito ao ponto de vista técnico, a implementação de multiprocessamento, assim como a reescrita do código em linguagens compiladas de médio nível, poderão também contribuir para melhorias

a nível do tempo de execução das análises. Finalmente, com pequenas modificações e testes adequados, a evolução natural do *PseudoChecker* passa pela sua expansão a outros animais.

Palavras-chave: Perda de Genes, Pseudogenes Unitários, Bioinformática, Mamíferos

Agradecimentos

O culminar desta dissertação não poderia ter sido atingido sem o apoio de muitas pessoas que cruzaram o meu percurso pessoal e académico neste último ano.

Em primeiro lugar, devo expressar o meu agradecimento total ao Professor Filipe Castro, que pela sua dedicação, empenho, constante boa disposição e inspiradora paixão pela ciência, tornou todo este processo mais aliciante, desafiante e compensador.

Não posso igualmente deixar de lhe agradecer a oportunidade e voto de confiança que me foram concedidos aquando do início da minha colaboração no grupo AGE (Animal Genetics and Evolution) do CIIMAR. Aí iniciei a minha, ainda que pequena, produção científica, tendo igualmente aprendido imenso desde o primeiro minuto. Deixo aqui o meu agradecimento a toda a equipa do AGE, em particular ao André Machado e à Mónica Marques por toda a paciência e conhecimento que me transmitiram.

Ao Professor Pedro Ribeiro agradeço toda a sua dedicação, empenho, apoio e empatia que revelou durante todo o ano e em particular nas proveitosas reuniões que me fizeram entender que a ciência de computadores não é apenas código, mas muito para além disso. O seu apoio foi fulcral para me fazer pensar 'além da caixa', essencial nalguns pontos deste projeto. Por isto, o meu sincero obrigado.

Aos meus amigos de sempre e que me acompanham desde o secundário (eles sabem quem são) agradeço todo o seu apoio, essencial para aliviar muitos dos meus momentos de maior stress e preocupação. Esta dissertação também é vossa.

Aos meus colegas e amigos da FCUP, em particular, ao Miguel e Filipe, os meus colegas (e amigos) do '2º ano do Mestrado', o meu voto de agradecimento por tudo o que vivi com eles.

Por último, devo o meu maior agradecimento à minha Mãe, ao meu Pai e ao meu irmão João por todo o apoio e incentivo constantes ao longo de todos estes anos.

Aos meus avós Arnaldo e Helena.

Contents

Author statement	i
Abstract	iii
Resumo	v
Agradecimientos	vii
Contents	xiii
List of Tables	xvi
List of Figures	xviii
List of Algorithms	xix
List of Abbreviations	xxi
1 Introduction	1
1.1 Historical context of evolution, genetics and genomics	1
1.2 Evolution by gene duplication	5
1.3 Evolution by gene loss	8
1.4 Motivation	11
1.5 Goals and contributions	12
1.6 Thesis outline	13

2	Basic concepts	15
2.1	Expression and structure of eukaryotic protein-coding genes	15
2.2	Pseudogenes	19
2.3	Unitary pseudogenes	20
3	Related work	23
3.1	First approaches towards the identification of pseudogenes	23
3.2	Identification of gene loss	25
4	Pseudo<i>Checker</i>	31
4.1	Introduction	31
4.2	Overview of Pseudo <i>Checker</i> 's structure and pipeline	33
4.3	Pseudo <i>Checker</i> 's pipeline in detail	36
4.3.1	Coding sequence prediction	36
4.3.2	Alignment by MACSE	49
4.3.3	Report of the mutational evidence and PseudoIndex computation	55
4.4	PseudoIndex	58
4.4.1	Absent exonic content component	60
4.4.2	Shifted codons component	60
4.4.3	Truncated sequence component	62
4.5	Illustrative examples of analyses ran with Pseudo <i>Checker</i>	64
4.5.1	Running a Pseudo <i>Checker</i> analysis	64
4.5.2	Pseudo <i>Checker</i> confirms the functionality of <i>TP53</i> gene across mammals	69
4.5.3	Pseudo <i>Checker</i> corroborates the convergent loss of <i>PON1</i> gene in several marine mammals	70
4.6	Limitations	72
5	Pseudo<i>Checker</i>'s performance evaluation	75
5.1	Experimental design	75
5.2	Evaluation metrics	77

5.3	Results	78
6	Conclusions	83
6.1	Main contribution	83
6.2	Perspectives and future work	84
A	Raw data associated with Pseudo<i>Checker</i>'s performance evaluation over gene functionality events	87
B	Raw data associated with Pseudo<i>Checker</i>'s performance evaluation over gene loss events	99
	Bibliography	105

List of Tables

4.1	Comparison of the different features and requirements regarding each available gene loss inference approach, including <i>PseudoChecker</i>	32
4.2	Fixed similarity scoring schemes provided by <i>PseudoChecker</i> underlying the coding sequence prediction component of its pipeline	44
4.3	Tested combinations of match/mismatch scores concerning the best-fit similarity scoring scheme underlying the first step of <i>PseudoChecker</i> 's integrated pipeline .	45
4.4	In-depth description of the alignment elementary bonus/costs associated with the scoring function employed by MACSE	52
4.5	Default values for MACSE alignment elementary costs	53
4.6	Relation of the percentage value of absent exonic content with the attributed sub-PseudoIndex for the absent exonic content component of PseudoIndex	60
4.7	Relation underlying the percentage value of shifted codons with the attributed preliminary sub-PseudoIndex for the shifted codons component of PseudoIndex .	61
4.8	Relation underlying the percentage value of truncated sequence with the attributed preliminary sub-PseudoIndex for the truncated sequence component of PseudoIndex	62
5.1	Percentage value of discarded observations per tested similarity scoring scheme and type of analysis (involving lost genes or functional ones)	78
5.2	Percentage value of well classified gene functionality events per tested similarity scoring scheme including the obtained average value for all of these	79
5.3	Percentage value of well classified gene loss events per tested similarity scoring scheme including the obtained average value for all of these	81
A.1	Raw data underlying the 90 analysis conducted during <i>PseudoChecker</i> 's performance evaluation over presumable cases of gene functionality	88

B.1	Raw data underlying the 90 analysis conducted during <i>PseudoChecker</i> 's performance evaluation over previously known cases of gene loss events	100
-----	---	-----

List of Figures

1.1	A new view of the tree of life	2
1.2	Biological data explosion	4
1.3	Evolutionary fate of gene duplication	7
2.1	Schematisation of the central dogma of molecular biology	17
2.2	Depiction of the five main known types of alternative splicing events	18
2.3	The origin and classification of pseudogenes	20
2.4	Schematic representation of the typical ORF-abolishing mutations found at unitary pseudogenes.	22
3.1	Relativity of gene loss identification	26
4.1	Pseudo <i>Checker</i> 's conceptualisation	35
4.2	Pairwise DNA sequence alignment following three different alignment problems .	37
4.3	Illustrative example of an optimal semi-global exon alignment produced by Pseudo <i>Checker</i>	38
4.4	Illustrative example of the extension of an exon alignment concerning a simulated reference species' 3' exon	40
4.5	Illustrative example of a multiple sequence alignment executed by MACSE . . .	50
4.6	Illustrative example of a MACSE alignment displayed by Pseudo <i>Checker</i>	55
4.7	Illustrative example of a MACSE alignment not accepted by Pseudo <i>Checker</i> . . .	57
4.8	Illustrative example of several different frameshift mutational events observable within a MACSE alignment regarding a simulated Pseudo <i>Checker</i> analysis	61

4.9	Illustrative example of several mutational events concerning the presence of premature stop codons observable within a MACSE alignment underlying a simulated <i>PseudoChecker</i> analysis	63
4.10	Overview of <i>PseudoChecker</i> 's home page	65
4.11	<i>PseudoChecker</i> 's main results page underlying the demonstrative simulated analysis.	67
4.12	<i>PseudoChecker</i> 's coding sequence prediction report	68
4.13	<i>PseudoChecker</i> 's main results page underlying the <i>TP53</i> gene test case	69
4.14	<i>PseudoChecker</i> 's main results page underlying the <i>PON1</i> gene test case	71
4.15	Illustrative example of a <i>PseudoChecker</i> simulated analysis underlying a MACSE alignment comprising sequences with their reading frame disrupted by the absence of exons	72
4.16	Illustrative example of a <i>PseudoChecker</i> simulated analysis underlying a MACSE alignment presenting erroneously aligning regions mediated by the aggressive loss of coding exons	73

List of Algorithms

1	High level overview of the basic approach underlying the coding sequence prediction component of <i>PseudoChecker</i> 's pipeline	42
2	High level overview of the heuristic method applied at the first component of <i>PseudoChecker</i> 's pipeline	48
3	High level overview of the basic approach underlying the Alignment by MACSE component of <i>PseudoChecker</i> 's pipeline	54
4	High level overview of the basic approach underlying the third and final component of <i>PseudoChecker</i> 's pipeline	57

List of Abbreviations

2R WGD	Two Rounds of Whole Genome Duplication	NGS	Next Generation Sequencing
3R WGD	Three Rounds of Whole Genome Duplication	ORF	Open Reading Frame
CDS	Coding Sequence	RAM	Random Access Memory
CPU	Central Processing Unit	RNA	Ribonucleic Acid
CSS	Cascading Style Sheets	SRA	Sequence Read Archive
DNA	Deoxyribonucleic Acid	TSS	Transcription Start Site
EBI	European Bioinformatics Institute	TTS	Transcription Termination Site
ENA	European Nucleotide Archive	UTR	Untranslated Region
GO	Gene Ontology	bp	Base Pair
GUI	Graphical User Interface	cDNA	Complementary DNA
HTML	HyperText Markup Language	mRNA	Messenger RNA
NCBI	National Center for Biotechnology Information	rRNA	Ribosomal RNA
		tRNA	Transfer RNA

Chapter 1

Introduction

1.1 Historical context of evolution, genetics and genomics

Organisms on Earth are rich in diversity, with complex genetic systems being generated and constantly modified through evolution since the origin of life [164] (Figure 1.1). The first efforts to understand biodiversity were conducted by Carl Linnaeus [105] who implemented the still used today binomial nomenclature of species, and had the critical insight that the diversity of life was hierarchically nested in groups, each characterised by particular sets of traits. This remarkable observation preceded and enabled the understanding that diversity was produced through evolution by a common descent [68].

Evolution has been a fascinating subject for several years and Charles Lyell in his publication the '*Principles of Geology*' (1830) was likely the pioneer using the word 'evolution' in the same context as we know today [116]. Moreover, thanks to the contributions of Wallace [195], Lamarck [99] and Darwin with the 'The Origin of Species' by means of natural selection [39], evolution as a biological phenomenon has been being gradually accepted during the last 200 years [112]. Yet, only in 1944 with the discovery of deoxyribonucleic acid (DNA) by Oswald Avery and his colleagues [9], succeeded by the discovery of its structure by Watson, Crick and Franklin in 1953 [198], revealing to constitute a substance encoding genetic information, together with the publication of the central dogma of molecular biology in 1970 [37], big steps towards understanding the molecular mechanisms on the basis of evolution were taken.

The evolutionary course has distinctive properties, as it does not proceed along some predictable course, instead being defined by the environment surrounding a population and in the genetic variants that arise on it. Moreover, evolution of species proceeds in a tree-like, branching pattern, as modern diversity of species has been generated by the repeated splitting of lineages since the single common ancestor of life [33, 164] (Figure 1.1). As a matter of fact, the perception of evolution has changed dramatically in the last 25-30 years, owing largely to the great advance, sophistication, and reliability of phylogenetics. Phylogenetic analysis of living organisms enabled science to trace the evolution of traits, their polarity of change, convergence,

reversal, and, when calibrated, allowed to relate evolutionary events to climate and geological changes, estimate rates of character evolution, and enabled the possibility of inferring the time course of diversification including rates of speciation and extinction [51].

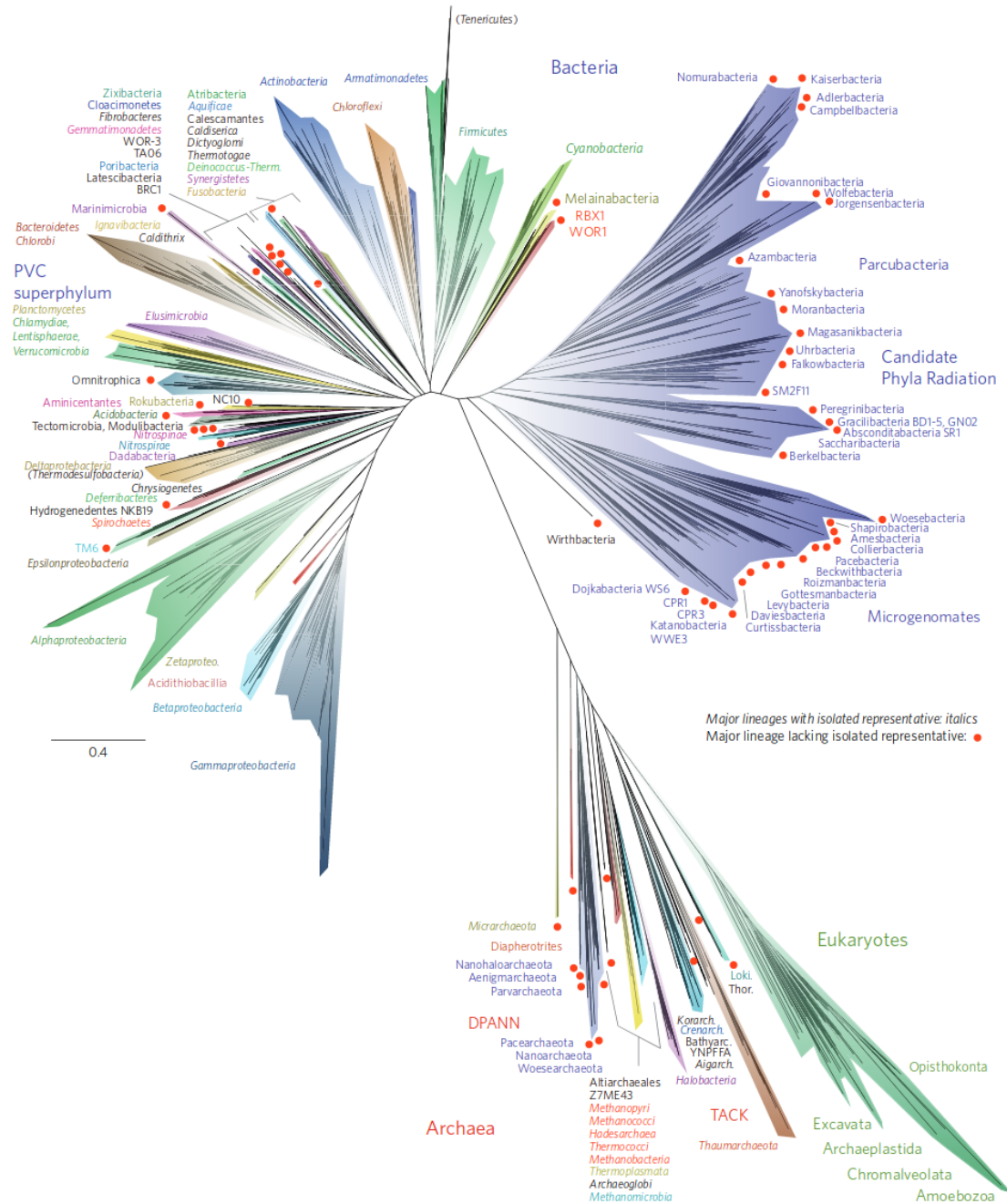


Figure 1.1: A new view of the tree of life. Retrieved from [76]. The phylogenetic tree comprises 92 bacterial phyla, 26 archaeal phyla and all the five supergroups underlying eukaryotes.

The genome of an organism, term firstly coined in 1920 by botanist Hans Winkler by joining the words 'gene' and 'chromosome' [164], consists in its complete set of DNA. Hence, evolutionary genomics focuses on studying how genome changes during evolution. In fact, since the first whole-genome sequencing of the bacteria *Haemophilus influenzae* in 1995 [49] that researchers

have been increasingly recognising the informational wealth that large-scale and high-throughput data can provide [189]. Extraordinarily, during the last few years, advances in genome sequencing have progressed at a rapid pace, with increased throughput being accompanied by decreasing costs and increased applications [121, 135, 138]. Next Generation sequencing, also named as 'Next-gen sequencing' or, simply, NGS, refers to highly parallelised processes that enable the sequencing of thousands to millions of molecules simultaneously, including DNA sequences, ribonucleic acid (RNA) sequences and even proteins [167, 178].

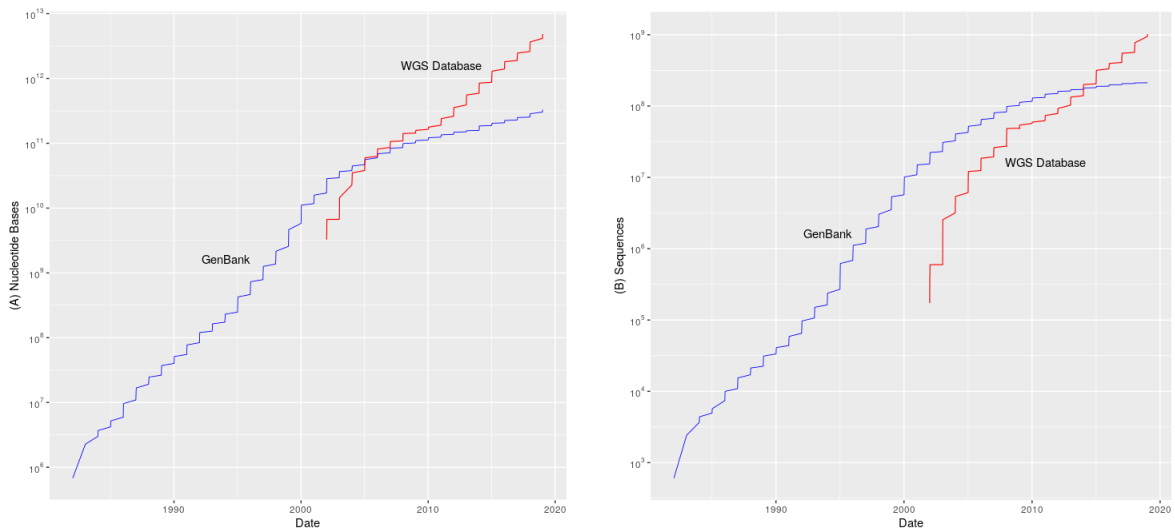
In fact, the biggest change in the paradigm of evolutionary studies occurred in the mid-2000s with the advent of cloud computing and the early days of NGS, leading to a dramatic increase in number and scale of biological datasets. Incredibly, in 2008, the sequencing cost dropped faster than what would have been expected from Moore's law, a term used in the computer industry to describe a trend that states that the number of transistors of an integrated circuit doubles approximately every 2 years [115, 167]. This revolution underlying the biological data availability was the cause of the creation of international databases for storing high-throughput sequencing data, such as the European Nucleotide Archive (ENA) governed by the European Bioinformatics Institute (EBI)¹ and the Sequence Read Archive (SRA) from the National Center of Biotechnology Information (NCBI)² [138, 167, 181]. In truth, since the creation of the Sequence Read Archive (SRA) that the number of deposited sequences has been growing significantly, containing, in June of 2019, approximately 28 petabases of biological information, translated into 10 000 terabytes of raw data (data obtained from SRA). A similar trend is observed for the Whole Genome Shotgun (WGS) and GenBank databases from NCBI (Figure 1.2), with the number and size of available nucleotide sequences largely growing over the past few years.

This revolution in biological data availability has led to increased robustness and decreased cost of the 'OMICS' approaches. These comprise big topics such as genomics, an interdisciplinary field aiming to perform a comprehensive study of whole genomes [65], transcriptomics, focused in studying both qualitatively and quantitatively the transcripts arising in a given biological sample [65, 166], and, finally, proteomics and metabolomics, that aim to describe the full molecular complexity of functional cells, tissues, organs and organisms in a holistic and integrative manner [53, 65]. Genomics, however, is still by far the most studied *omic* field, and particularly, comparative genomics, a sub-area of genomics, is nowadays one of the major approaches used in the functional annotation of genomes [131]. Comparative genomics aims to perform direct comparisons of the complete genetic material of different organisms, gaining a better understanding of how species evolved, determining the function of genes and non-coding regions in genomes, also tending to understand the forces and evolutionary mechanisms that actuate over different species [6]. Comparative analysis comprehends such an important tool to better characterise genomes that a set of 29 mammalian genomes, including 22 specifically sequenced for the project, were analysed together as means of better understanding the human genome [69, 104].

Comparative genomics analyses nowadays focus on the similarity and differences between

¹www.ebi.ac.uk

²www.ncbi.nlm.nih.gov



(a) Number of nucleotide nitrogenous bases submitted to WGS and GenBank databases over the last several years.

(b) Number of nucleotide sequences submitted to WGS and GenBank databases over the last several years.

Figure 1.2: Time scale of the number of nucleotide nitrogenous bases (a) and nucleotide sequences (b) deposited through direct submission in GenBank (blue lines) and Whole Genome Shotgun (WGS) (red lines) databases. Data retrieved from NCBI on June 2019.

the annotation or sequences of two or more genomes [69], and aim to compare the gene number, content, location, the length and number of coding regions within genes, the amount of non-coding DNA in each genome and conserved regions maintained across compared species [175]. Genome correspondence, the method of determining the correct match of chromosomal segments and functional elements across species, involves determining orthologous genes, those that diverged after a speciation event, hence, direct evolutionary counterparts, and paralogous genes, those that arose by duplication events prior to the divergence of the species compared [93].

Once genome correspondence is established, comparative genomics can be used for determining the conservation of synteny across species (the preserved physical co-localisation of genes on chromosomes of different species) [175], rearrangements of DNA that include insertions, deletions, inversions, translocations [24], and events of gene gain and gene loss [175]. Remarkably, the first comparative genomics analyses started in the end of the nineties, for which time there were about 30 bacterial, archaeal and eukaryotic available genomes to compare, with these unveiling the existence of small sets of universal gene families encoding for core components of the transcription and translation machineries underlying gene expression [93]. Just a few years later, comparisons of the human genome with the genome of the fruit fly (*Drosophila melanogaster*) revealed that approximately 60 % of genes are conserved between fly and human, with two-thirds of human genes known to be involved in cancer having counterparts in the fruit fly.

The continuing advances in speed, quality and affordability of whole-genome analyses, including whole-genome sequencing, have transitioned the comparative genomics field from the realm of comparing reference sequence assemblies to comparing assemblies of individual

genomes [96]. Evolutionary biology tends not only to explain the origin and history of species, but is also decisive for enabling the understanding of the future of species, and although biologists have long recognised the importance of studying evolution to understand the organisation of living organisms, only with the development of comparative genomics, placing genomes into an evolutionary and comparative framework, evolutionary studies have become part of their routine toolkit [3].

Researchers are now increasing the interest in exploring how interactions among genes, individuals and environments have shaped the evolutionary processes both at micro- and macrolevels, an indispensable understanding for gaining the unified view of life, and, in some manner, Biology [113], as Theodosius Dobzhansky, one of the twentieth century's most eminent evolutionary biologists remarked in one often quoted phrase, '*nothing in biology makes sense except in the light of evolution*' [44].

1.2 Evolution by gene duplication

Gene duplication has long been recognised as an important mechanism of evolution. The first phenotypic outcome from duplication was observed in 1936 by Bridges, who reported that a doubling chromosomal band in a mutant of the fruit fly *D. melanogaster* led to an extreme reduction in its eye size [20]. Nonetheless, it was in 1970 with the Book '*Evolution by Gene Duplication*' with Ohno's observations on duplication and the proposal of whole-genome duplications that a turning point in evolutionary thinking was established [144]. Based on his observations concerning genome sizes in several species, Ohno postulated that the observed differences could result from duplication and proposed that a tetraploidisation event took place with the emergence of the first vertebrate, approximately 500 million years ago [144].

Nowadays, this hypothesis is known as the 2R hypothesis, and it suggests that the vertebrate ancestor underwent two separate rounds of whole-genome duplication (2R WGD). Yet, the 2R hypothesis was not immediately accepted, it was challenged and generated a fair amount of controversy in the late 1990s and early 2000s [77, 78, 122, 176], however, even though its time of occurrence it is still discussed [72], with the increase of the genomic data availability, several studies were published supporting the 2R hypothesis [42, 139, 147, 156, 192] and the importance and prevalence of gene duplication was clearly demonstrated. Moreover, with the aid of molecular experimentation, accompanied by studies underlying population genetic modelling, the mechanisms by which duplicate genes diverge in function and contribute to evolution were further disclosed [205].

Today, it is believed that in all three domains of life (Bacteria, Archaea and Eukarya), large proportions of genes were generated by duplication. Indeed, it is assumed that these proportions correspond to underestimates, as many duplicated genes have diverged so much, that virtually no sequence similarity can be detected. [205]. Lynch and Conery estimated that in eukaryotes such as *Homo sapiens* (human), *Mus musculus* (house mouse), *D. melanogaster* (fruit

fly), *Caenorhabditis elegans* (roundworm), *Arabidopsis thaliana* (thale cress) and *Saccharomyces cerevisiae* (baker's yeast) gene duplication arises and fixes in populations over an approximate rate of 1 gene per 100 million years [118]. Paralogues form gene families, and these vary in size and number amongst species. Particularly, in *D. melanogaster* (the fruit fly) the biggest gene family belongs to the trypsin gene family, comprising 111 paralogous genes [59], whereas the biggest gene family in mammals belongs to the olfactory receptor family with approximately 1000 members [132, 208].

Depending on the mechanisms of duplication, the corresponding outcomes may be slightly different. Duplication may result from unequal crossing over during meiosis, retrotransposition, chromosomal or whole-genome duplication and finally, polyploidisation. Unequal crossing over often originates tandem gene duplication (duplicated genes linked in a chromosome), that, depending on its position of occurrence, the duplicated region might comprise the part of a given gene, its entire portion, or several ones. Retrotransposition, on the other hand, occurs when a messenger RNA (mRNA) is retrotranscribed to complementary DNA (cDNA) further being then re-inserted into the genome. Chromosomal or genome duplication probably occurs by lacking of disjunction among daughter chromosomes after DNA replication during meiosis [205], and finally, polyploidisation corresponds to an evolutionary process whereby two or more genomes are joint together into the same nucleus, usually by hybridisation followed by chromosome doubling [123], a notable evolutionary process proposed to have occurred frequently in plants, particularly in 95% of pteridophytes [57] and in 70% of angiosperms.

Nonetheless, for a duplicate gene to be retained, it has to persist in the genome until fixed in the population [79]. Upon fixation, the long-term evolutionary fate of duplication will still be determined by the functions of the duplicate genes [205]. In this sense, several evolutionary outcomes have been documented for duplicate genes (Figure 1.3). After duplication, a duplicate gene might accumulate several disrupting mutations, leading to its erosion and ultimately loss, this outcome is known as nonfunctionalisation or pseudogenisation and represents the most common outcome for a duplicate gene. Indeed, massive nonfunctionalisation phenomena occurred shortly after the two rounds of whole-genome duplication concerning the vertebrates ancestor (2R WGD), as well as following the teleost ancestor specific three independent rounds of whole-genome duplication (the latter phenomenon currently known as the 3R hypothesis) (3R WGD) [30, 74, 80].

Still, other evolutionary duplication outcomes are possible. One of which consists in the conservation of both copies, each assuming the parent gene function, yet being differently regulated, expressed in alternative tissues or developmental stages [114, 118, 119]. Alternatively, both copies might maintain the parent gene function and be both expressed in the same tissue or development stage, as the presence of redundant duplicated genes is in certain cases beneficial for providing extra amounts of protein or RNA products required in high demand, such as ribosomal RNAs (rRNAs) and histones. In these cases, however, both copies suffer slight downregulation of their transcription levels, and since a single copy cannot fulfil the required expression level, selective pressure for the preservation of both copies is created [157, 168, 205].

Yet, unless the presence of an extra amount of gene product is required, two co-expressed genes with identical functions are less likely to be kept in the genome [142]. Population genetics states that both duplicate genes can, however, be stably maintained if they differ in some aspects of their functions, which can occur by subfunctionalisation, an evolutionary outcome of gene duplication in which each gene copy adopts part of the functions of their parental gene [86, 146] or, finally, by neofunctionalisation, the most important gene duplication evolutionary outcome, by which one gene copy maintains the ancestral function while the second copy suffers large genetic changes, accumulating mutations over time [8, 206], functionally diverging and acquiring a novel function [205].

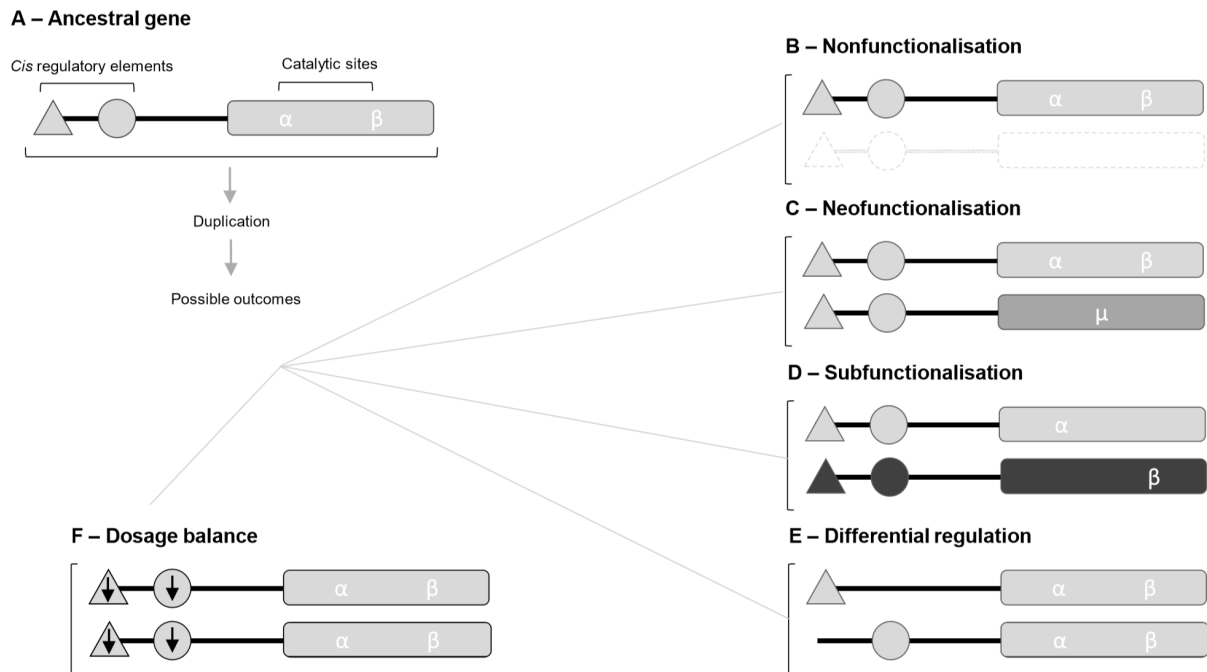


Figure 1.3: Schematic representation of the several possible evolutionary outcomes after gene duplication. Adapted from [112, 114]. (A) Representation of an ancestral gene containing two active sites, upstream regulated by two *cis*-elements that undergoes duplication; (B) Nonfunctionalisation or pseudogenisation, one duplicate accumulates deleterious mutations leading to its ultimate inactivation; (C) Neofunctionalisation, one gene copy acquires a novel function and the other maintains the ancestral function; (C) Subfunctionalisation, both copies are kept in the genome by partitioning the ancestral function; (D) Differential regulation, both copies and ancestral function are maintained, yet, these are expressed in different conditions (tissues or development stages); (F) Both copies share the ancestral function and are expressed under the same conditions, however, gene expression is downregulated.

Evolution by gene duplication has likely played a substantial role in explaining the rapid changes in organisms' complexity observed in deep evolutionary splits and diversification of more closely related species. The most valuable contribution of gene duplication to evolution consists in providing new genetic material for mutation, drift and selection to act upon, generally culminating in specialised or new gene functions. In the absence of gene duplication, however, the plasticity of a genome or species would be limited since no more than two variants per gene (alleles) exist at any *locus* within a (diploid) individual, therefore, it would be difficult, for instance, to comprehend how the vertebrate adaptive immune system, comprising dozens of duplicated immunoglobulin genes would exist without gene duplication phenomena [25, 32, 205].

Additionally, gene duplication has probably also contributed to the appearance of species-specific features via species-specific gene evolution and to the evolution of gene networks in a way that complex and sophisticated expression regulation mechanisms can be established [193, 205]. An insightful example points to the paralog resulting from the duplication of the *Pax6* gene in *Drosophila*, also known as *eyeless*, underlying the eye development control in metazoans, that now regulates the expression of *eyeless* itself [38]. Indeed, it is expected that accompanied by the genomic biological data explosion, as well as with quick advances in molecular genetic technology, the fundamental roles and mechanisms underlying gene duplication will become even more evident, sustaining evolution by gene duplication as one of the cornerstones of evolutionary biology [205].

1.3 Evolution by gene loss

Gene loss, in contrast, has mostly been neglected as an evolutionary force since it has been often associated with the loss of duplicate redundant genes without apparent functional consequences. However, with the aid of genomic data, recent studies have highlighted the crucial role of gene loss in evolution [5, 145]. Two different mechanisms have been proposed to explain the loss of a gene from a given genome. First, the loss of a gene can result from an abrupt mutational event such as unequal crossing over during meiosis or through the mobilisation of a transposable or viral element that is capable of physically removing a gene from an organism's genome. On the other hand, the loss of a gene can be the result of a slow process of accumulation of deleterious mutations following an initial loss-of-function mutation. Indeed, the term 'gene loss' is nowadays often used in a broad sense, not only referring to the absence of a gene, identified when comparing orthologous *loci* in different species, but also to any allelic variant comprising a deleterious mutation that is found fixed within a population [5].

It is believed that gene loss is pervasive across multiple lineages. The perspective on animal evolution has changed after the genome sequencing of multiple cnidarian species, as several studies revealed that the ancestral eumetazoan genome was much more complex than what was expected and that gene loss had occurred in several animal phyla [95, 97, 185]. This new evolutionary perspective led to a change in a classical paradigm that attempted to correlate the apparent increase of biological complexity in an evolutionary context with an increase in the

number of genes [183]. Remarkably, in mammals, from a total of 9 990 gene families that were inferred to be present in their most recent ancestor, 1421 gene families (14%) revealed to be fully absent in at least one extant genome [43]. Amongst primates, humans have undergone 86 gene losses whereas chimpanzees suffered 729 gene losses over the same period [43, 196]. Furthermore, alongside animals, fully sequenced genomes from a large range of species, including prokaryotes [103, 117, 125, 136, 201], fungi [149, 179], protozoa [107] and plants [40] have shown that gene loss is massively present in all life kingdoms [5].

Importantly, the unevenness of gene loss distribution over the three of life suggests that this evolutionary pattern does not occur in a stochastic way. There is, instead, a bias related to the genomic position that a given gene occupies, for which the biological significance remains somewhat unclear, and the gene function that a given gene assumes. For instance, concerning the latter case, a comparison of *Schizosaccharomyces pombe* (fission yeast) and *S. cerevisiae* (baker's yeast) revealed that not all gene ontology (GO) functional categories were equally affected by gene loss [5]. Instead, a large fraction of lost genes corresponded to functional categories including nuclear structure maintenance, post-transcriptional gene silencing, messenger RNA (mRNA) splicing, protein folding and processing [5]. In plants, for instance, genes that are involved in repairing and modifying DNA or in ancient biochemical processes are more prone to be lost than genes encoding transcription factors and protein kinases [16, 26, 120, 169]. Concerning vertebrates, genes involved in protein modification, metabolism, catabolism and peptidase activity are more susceptible to be lost in fish than in land vertebrates, whereas genes that are involved in catalytic activities reveal the opposite trend [17]. In mammals, particularly, genes from gene ontology (GO) categories related to the immune response, reproduction, transcription, chemosensation or gamete interaction are more likely to be lost than in other vertebrate lineages [43, 126].

Such biased patterns of gene loss are likely to arise from differences concerning gene dispensability of each functional category. Gene dispensability may result from environmental variability, in the sense that gene functions are not equally essential in every environment, but also from the mutational robustness of biological systems. In a simplified approach, mutational robustness can result from the presence of redundant or backup genes (usually genes that arose from duplication, paralogues), or via the presence of alternative pathways, in the sense that genes and proteins are structured in regulatory networks with a scale-free structure re-routing biological tasks through alternative pathways when one part of the network fails, hence, conferring distributed robustness into the network. These alternative pathways are also known as complementation or backup pathways [5, 194].

For understanding the impact of gene loss throughout evolution, however, it is critical to infer the potential adaptability or neutrality of nonfunctionalisation mutations leading to gene losses. Currently, there are two main evolutionary models for gene loss. The first model proposed by Olson *et al.* - the 'less is more' hypothesis - states that nonfunctionalisation represents an evolutionary adaptive response. Adaptive gene loss has been reported for a large range of species, from bacteria (where more than 200 cases of adaptive gene loss have been reported) [73] to

multicellular organisms. For instance, in plants, adaptive gene loss has been associated with pollinators: in *Petunia axillaris* (white moon petunia) the loss of *AN2* gene leads to the formation of white flowers, which has been proposed as an adaptation to pollination by nocturnal hawk moths [71], moreover, in *Ipomoea quamoclit* (red-jasmine), the loss of flavonoid 3'-hydroxylase gene leads to the appearance of red flowers, consecrating an adaptive trait for bird pollination [217]. In cosmopolitan *D. melanogaster* (fruit fly) the loss of *Desaturase 2* (*Desat2*) is associated with resistance to cold environments, and, in other flies of the same gender, the loss of some genes related with chemoreceptors such as gustatory or odorant receptors is associated with the loss of behaviours, colonisation of new ecological niches, and with the evolution of new diets [34, 56, 124].

Nevertheless, some of the most widely known cases of adaptive gene loss, were, in fact, described in humans. Remarkable adaptive gene loss concerning this species include for instance, a polymorphic 32 base pair (bp) deletion of *CCR5* gene that acts as a receptor for HIV, with homozygous individuals for that deletion being protected against the infection by the same virus [41], as well as loss-of-function mutations underlying the atypical chemokine receptor 1 (*ACKR1*, also known as *DUFFY*) culminating in natural resistance to vivax malaria [188] for the affected populations. Interestingly, several adaptive gene loss events emphasised the evolutionary origins of humans. Illustrative examples of these include the loss of *MYH16* gene, that probably became dispensable right after diet changes, reducing the reliance on powerful masticatory jaw muscles during the evolution of hominids, thus removing anatomical constraints allowing the development of the modern human brain, and the losses of *CMAH* and *CASPASE12* genes, that provided resistance to some pathogens, and protection from severe sepsis, respectively [27, 196].

The second evolutionary model for gene loss, more recently proposed, suggests that gene loss events bring no effects to the involved species. This evolutionary model, based on the 'use it or lose it' principle, is known as regressive evolution, and according to it, characteristics that have been rendered useless over time, are lost with neutral effects [5, 85]. Insightful examples of regressive evolution include the independent loss of the *GLO* gene, which encodes the enzyme responsible for the last step in vitamin C biosynthesis, in different vertebrates lineages after the adaptation to vitamin C rich diets, including some teleost fish, bats, guinea pigs, several primates and passeriform birds, as well as the neutral loss of eye development and pigmentation related genes, unnecessary for life in darkness, in several populations of the cavefish *Astyanax mexicanus* after the colonisation of dark environments [154]. Yet, although these losses regarding cave system environment adaptations were initially seen as neutral, it can be argued that these come with an energetic benefit to the involved species, as the absence of producers limit the nutrient availability, leading to the rerouting of energy previously allocated for eye development to support other vital functions [133].

1.4 Motivation

One of the most fascinating aspects of nature consists in the phenotypic diversity amongst life forms. A key challenge in both contemporary evolutionary biology and genomics passes by understanding the evolution of phenotypic traits at the molecular level. This necessarily requires identifying the underlying genomic and genetic signatures on the origin of adaptive phenotypes, which may reveal insights into the subjacent molecular and cellular mechanisms [172].

Recent improvements in DNA sequencing including linked- and long-read technologies have provided the ability to produce high-quality genomes at prices that make genome assembly an affordable experiment for the diverse laboratories across the world [7]. Moreover, different consortia aiming to produce genome assemblies on a wide scale, exemplified by the recent sequencing of 48 bird genomes [83], and ongoing projects such as Genome-10K [143] and i5k [161], intending to respectively sequence the complete genomes of 10 000 vertebrate and 5 000 arthropod lineages, have now made it possible to make use of comparative analysis to associate genomic differences with phenotypic diversity across species.

Gene duplication bursts and amino acid substitutions have often been assigned to the evolution of novel traits such as vision, hair, taste and digestion. Much less explored is the role of gene loss as an evolutionary force, either from adaptive or regressive perspectives, in response to specific environmental cues and niches [5, 172]. Mammals, particularly, provide a reference test case to address the role and importance of gene loss as a major driver of evolutionary change. They successfully colonised a large range of habitats, and display tantalising phenotypic adaptations to their surrounding environment, including specialised dentition, placentation, enlarged brains, lactation, increased sensitivity of sense organs and hair to preserve heat and skin [172].

Paradoxically, the study of gene loss in Mammalia is still in its early days. Indeed, until a few years ago, the vast majority of gene loss studies highlighting this *taxon* only covered humans or other primates. Yet, recent studies have now been unveiling associations between gene losses and several non-primate mammalian phenotypes [172]. Insightful examples point to recurrent gene losses in cetaceans that underlined their remarkable anatomical, physiological and behavioural adaptations throughout their evolutionary transition from land to water, with these including, for instance, the erosion of the melatonin gene modules, a hormone known to participate in the maintenance of circadian rhythms, paralleling their idiosyncratic sleeping and vigilant behaviours [111], the complete inactivation of sebum-producing genes associated with their loss of sebaceous glands [110], the specific loss of hair type keratin producing genes accompanying their hairless phenotype [141], and the inactivation of Odontogenic ameloblast-associated (*ODAM*) gene, expressed in dental tissue, likely explained by the simplified outer enamel surface of toothed whales [180]. Furthermore, convergent loss phenomena have been also reported in Mammalia, including, for instance, the independent gene losses underlying similar dietary adaptations of diverse herbivores and carnivores organisms [67] and the loss of the skin-specific C-C motif chemokine ligand 27 encoding gene during mammalian evolution, underlying specific skin physiologies [109].

Nonetheless, the full catalogue of mammalian gene inactivation occurrences and affected lineages is still vastly incomplete. For this reason, it is still largely unknown whether gene loss can also play a relevant role in natural phenotypic adaptations underlying other, not yet explored, mammals [5]. As we will further see in Chapter 3, several automatised and semi-automatised pipelines are currently available for the identification of redundant loss underlying either nonfunctional duplicated or retrotransposed copies of a parent gene within a given genome. Yet, few are the systematic approaches capable of truly inferring gene loss across Mammalia, and those available for that effect present strong limitations such as: exclusive application to humans or other primates [196, 213], limited design of the approach, being specifically created for addressing the questions of specific research, poor degree of automation, demanding computational resources requirements [172], or the requirement of an extensive and exhausting manual curation [109–111, 180].

1.5 Goals and contributions

It is known that contemporary research in biosciences increasingly depends upon bioinformatics for the effective analysis of biological data. Indeed, having access to the appropriate bioinformatics tools is nowadays crucial for the success of any research project [54, 92]. The rapid increase of high-quality genome assemblies requires the introduction of automated methods capable of leveraging their comprehensive study. Particularly, for addressing and promoting the study of gene loss as a source of mammalian evolutionary change, a systematic, highly accurate, and computational automatised approach is ultimately required.

Motivated by this necessity, this thesis culminated in the development of *PseudoChecker*: the first integrated online platform for automatised gene loss inference in mammals. We intended not to develop a simple command-line based computational approach, but instead to develop a fully integrated free-to-use bioinformatics web service specifically designed for accurately classifying protein-coding genes as intact or lost (with different levels of pseudogenisation) in any mammalian given species by taking advantage of the existent and emerging mammalian genomic data. *PseudoChecker* was intentionally conceptualised to constitute a general-purpose tool, thus, designed to be systematically applied in any gene loss study, being easily usable and expandable to any researcher of the field.

To produce this proposed integrated pipeline, we took into account several important aspects and established main goals underlying *PseudoChecker*'s development, namely:

- (a) Introduce the main capability of simultaneously and at a quick pace define the coding status of any protein-coding gene at any given mammalian species without restrictions regarding gene size, number of target species or genome versions underlying them;
- (b) Apply the ability to infer and display relevant information concerning mutational evidence, including non-conserved and conserved deleterious mutations across related *taxa*, hence, in

- the latter case, unveiling ancestral gene inactivating events;
- (c) Implement a pseudogenisation index, a user auxiliary metric intended to, at a glance, objectively measuring the level of erosion of a given gene at a given target species;
 - (d) Integrate the entire system as a web service not only to keep the tool always updated and allow its usability in any operative system, but also to provide a user-friendly interface that avoids any contact with command-line based interactions, consequently increasing the diversity of possible users, and dismissing any prior file(s) download(s), local software installation or consumption of local computational resources;
 - (e) Once completed the development, evaluate the tool's accuracy in classifying protein-coding genes as functional or lost, by confronting it, in an experimentally designed systematic way, with ground truth cases of published lost genes, as well as with protein-coding genes presumably functional across Mammalia.

1.6 Thesis outline

The present thesis is structured into six major chapters. A brief description of each one of them is now provided.

Chapter 1 - Introduction. Provides a general introduction of the topic being explored throughout the thesis, the main motivation, goals, contributions, and its structural organisation.

Chapter 2 - Basic concepts. Introduces the reader to the basic biological concepts regarding the structure and expression of eukaryotic protein-coding genes, nonfunctional elements known as pseudogenes, and particularly, unitary pseudogenes, the vestigial elements underlying gene loss.

Chapter 3 - Related work. Consists of a detailed review of the approaches proposed over the years for enabling the large-scale identification of mammalian pseudogenes in the broad sense, as well as the methodologies underlying the identification of gene loss in the same group of organisms.

Chapter 4 - PseudoChecker. Starts by providing a brief introduction to *PseudoChecker*, the bioinformatics tool developed during this thesis, including the source of inspiration underlying its development, as well as a brief explanation of the pipeline. Then each step of the pipeline is fully explored, and the corresponding implementation scrutinised. Next, *PseudoIndex*, the user auxiliary metric specifically created for the tool, as well as each composing criteria are explored. Finally, the chapter ends by presenting illustrative examples of *PseudoChecker* in action, further describing current *PseudoChecker*'s limitations.

Chapter 5 - PseudoChecker's performance evaluation. Addresses the *PseudoChecker*'s performance evaluation steps conducted during the work, as well as the relevant metrics used in the same evaluation.

Chapter 6 - Conclusion. This final chapter summarises and provides the final remarks for the produced work, culminating with future perspectives and tasks to be posteriorly conducted.

Chapter 2

Basic concepts

2.1 Expression and structure of eukaryotic protein-coding genes

Studying large genome sequences such as that of humans and other species has enabled an unparalleled and detailed look at the structure and organisation of both coding and non-coding information of the genomes from multiple species [163]. The genome sequence of an individual corresponds to its blueprint, dictating the set of instructions underlying many of its biological traits. Genes, in this sense, comprehend the fundamental units of inheritance and the ultimate determinants of all phenotypes within an organism. They provide the structure, function and regulation of every biological system. Particularly, protein-coding genes, in contrast to non-coding DNA sequences [90, 98], are responsible for encoding the information underlying the main functional and structural players in the cell, proteins [64, 153].

According to the central dogma of molecular biology (Figure 2.1), the real effect of a protein-coding gene is conducted during its expression. This implies having its DNA partially unwind, transcribed into a messenger RNA molecule, that is further translated into a polypeptide, posteriorly acquiring a certain structural conformation, turning into a protein, the final effector of the gene's action. The first step underlying the central dogma of molecular biology consists in transcription. This step occurs in the cell nucleus and it is controlled by the promoter of the gene in question. This control determines when and in what tissue a given gene is expressed, and it is exerted by specific nucleotide sequences in the promoter region named *cis*-acting factors and by proteins, the *trans*-acting factors that must interact with the DNA sequence of the promoter and/or RNA polymerase II for transcription occurs. Specifically, the enzyme RNA polymerase II synthesises a (single-stranded) RNA molecule that is complementary to the entire stretch of the gene's DNA template, the opposite strand from where a given protein-coding gene is encoded. Physically, this process initiates in the transcription start site (TSS) and terminates in the transcription termination site (TTS) (Figure 2.1).

Yet, before the mRNA molecule undergoes translation to yield a protein, this primary transcript needs to suffer processing before turning into a mature messenger RNA (mRNA). This processing

step includes the addition of a cap to the 5' end of the transcript, directly involved in enzymatic reactions underlying the process [158], RNA splicing, and the addition of poly-A (polyadenylation) tail to its 3' end, necessary for endorsing the stability of the final transcript [129]. The most relevant component of this process, in fact, corresponds to RNA splicing, which involves the removal of non-coding sequences, but referred to as intervening sequences, the introns, thus, culminating in an mRNA molecule consisting of the remaining sequences, called exons, connected to one another via the splicing process, most of them, directly involved in encoding protein information. [64, 153]. This means that when a given gene is transcribed, the primary transcript spans more genomic terrain than what would be indicated by the length of the gene coding sequence [163] (Figure 2.1).

The biochemical mechanism underlying RNA splicing has been studied and is now fairly well characterised. Introns are removed from the primary transcript by cleavage of conserved dinucleotides, the splice sites, found at the 5' and 3' ends of introns. This process is conducted by cellular machineries named spliceosomes, and in the majority of cases, the removed RNA sequence begins with the dinucleotide GU (or GT in the DNA alphabet) in the 5' end (constituting the 5' splice site, also known as the donor splice site), and ends with the dinucleotide AG at its 3' end (which constitutes the 3' splice site, also known as the acceptor splice site). Importantly, besides these two canonical splice sites are conserved in the vast majority of mammalian genes (in over 98% of the cases [22]) a small percentage of genes can also hold non-canonical, but still viable, GC-AG splice site pairs. As we will see in Section 2.3, these consensus dinucleotide sequences are of great importance since the disruption of these sites may result in the loss of splicing [29].

Once the transcription and RNA processing steps are completed, the mature messenger (mRNA) is transported to the cell's cytoplasm and its translation initiates. In this process, the now processed mRNA is read according to the genetic code which relates the DNA sequence of the gene to the amino acid sequence of the protein that it encodes. Each group of three nitrogenous bases in an mRNA constitutes a codon, and each codon specifies a particular amino acid sequence in proteins. Since the 64 triplets that can be formed by 4 nucleotides exceed the 20 distinct amino acids used to compose proteins, different amino acids can be encoded by several different triplets [153]. This is usually referred to as the genetic code redundancy.

In detail, with the aid of specialised ribosomal subunits that are joint together on the mRNA molecule, and particularly, with specialised RNA molecules including ribosomal RNA (rRNA) and transfer RNA (tRNA), the mRNA molecule is used as a template to assemble in an ordered way the chain of amino acids that will constitute the encoded protein [28] (Figure 2.1). The translation initiates when the translation machinery recognises the methionine-coding codon, also known as the start codon, defined by the triplet AUG (or ATG, when referring to the corresponding DNA molecule), and next, each amino acid will then be incorporated into the growing polypeptide chain according to the genetic code underlying the codon being read. The translation finally terminates when one of the three stop codons (UAG, UAA, UGA, or, in the DNA alphabet, TAG, TAA, TGA, respectively) are detected by the translation machinery, in particular by release factors that promote the dissociation of the ribosome subunits from the

mRNA molecule.

Nonetheless, not all regions of an mRNA molecule are translated into amino acids. The so-called 5' and 3' untranslated regions (UTRs) are found at the beginning and end of an mRNA molecule, respectively. These are responsible for playing crucial roles in the post-transcriptional regulation of gene expression and do not encode any amino acid sequence. Moreover, depending on where the translation begins and ends, the UTRs may be derived from more than one exon and may constitute a considerable portion of the length of the processed transcript. In fact, exons might even be classified into two different classes according to its coding nature: coding exons, those which code for an amino acid sequence, and non-coding exons, those which constitute part of an UTR, thus, not encoding any amino acid. Therefore, the coding sequence (CDS) of a protein-coding gene consists on the nucleotide sequence initiated by the start and terminated by a stop codon, respectively flanked by the end of the 5' UTR and start of 3' UTR, responsible for encoding a given protein [28, 153] (Figure 2.1). Particularly, in the cases where the coding sequence (CDS) of a gene comprises a single exon, these are often classified as single-exon, or simply, intronless genes [88].

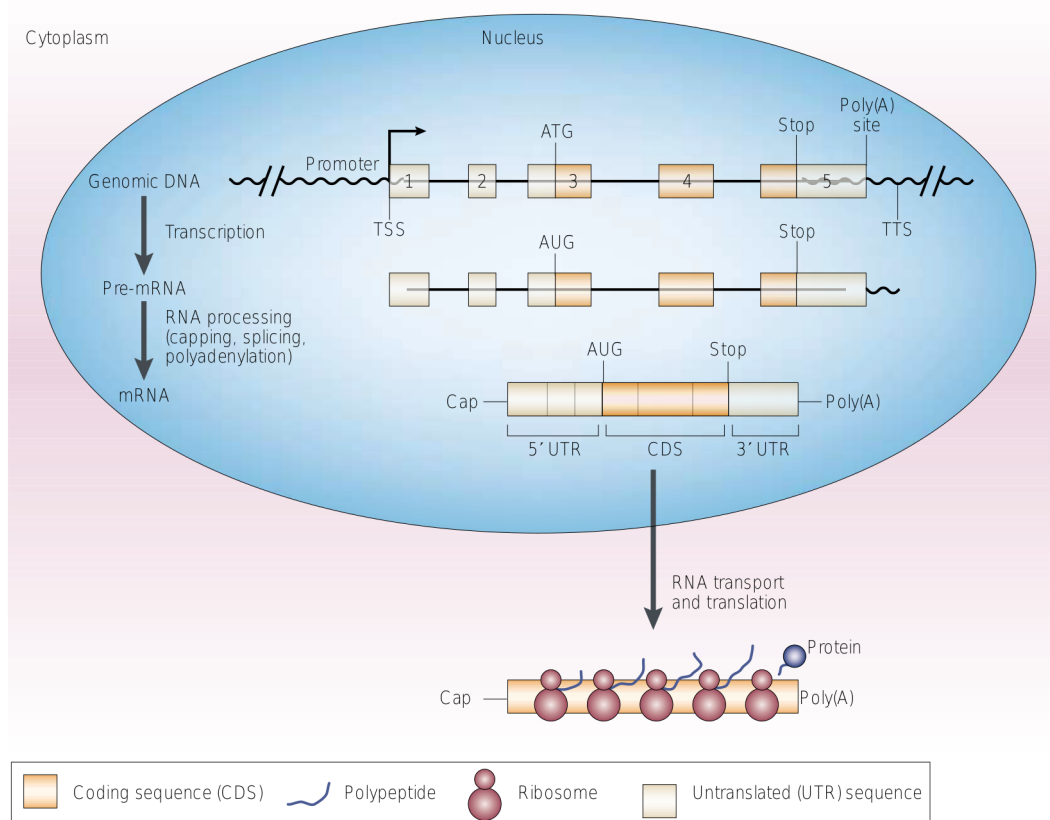


Figure 2.1: Schematisation of the central dogma of molecular biology. Retrieved from [207]. The coding sequence (CDS) of a gene consists in the nucleotide sequence flanked by the 5' and 3' untranslated regions (UTRs) and initiates with the start codon, ending with the stop codon. Coding exons are represented by orange boxes, whereas non-coding exons are represented by grey boxes.

Alternative splicing, in contrast to the previously mentioned constitutive splicing, has been recognised as an additional prevalent RNA splicing mechanism in eukaryotes and, particularly, in mammals. This phenomenon, first proposed in 1978 by Gilbert [55], has been suggested to constitute a mechanism of increasing protein complexity from a limited number of genes by making use of the distinguishable eukaryotic protein-coding gene organisation in exons and introns. Alternative splicing consists on a deviation from the constitutive splicing in the sense that certain exons are skipped, resulting in various forms of mature mRNA, usually called splicing variants, encoding different protein isoforms [197]. In fact, numerous studies have unveiled the critical and fundamental role of alternative splicing across biological systems [81], dismissing the old notion that assumed that a single gene produced a single mRNA molecule, thus, producing a single protein. Nowadays it is believed that alternative splicing plays an important role in almost every aspect of the cell's functional repertoire, including nucleic acids or membranes, binding between proteins and ligands, and finally, protein localisation and enzymatic properties [81]. The different known alternative splicing events are depicted in Figure 2.2.

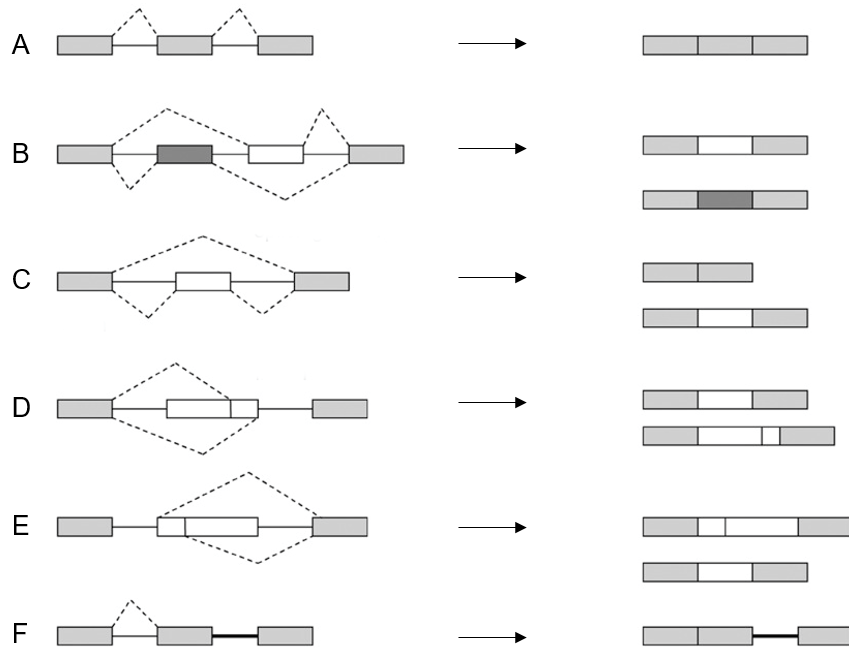


Figure 2.2: Depiction of the five main known types of alternative splicing events. Adapted from [197]. In the left, the pre-processed mRNA transcripts corresponding organisation in exons (represented by boxes) and introns (represented by continuous lines) is schematised. Intron cleavage is represented by dashed lines. In the right, the exonic constitution regarding each splice variant obtained through alternative splicing (with the exception of (A)) is depicted. Description of the splicing events: (A) Constitutive splicing; (B) Mutually exclusive exons splicing event; (C) Cassette alternative exons, also known as exon skipping splicing event (the most prevalent pattern in vertebrates and invertebrates); (D) Splicing taking into account an alternative 3' splice site within an exon's sequence; (E) Splicing taking into account an alternative 5' splice site within an exon's sequence; (F) Intron retention event (commonly prevalent in lower metazoans) [197].

Importantly, before advancing into the next sections of this thesis, here we should make a statement. The coding sequence (CDS) of a gene, depending on the literature, can be seen as a particular case of an open reading frame (ORF), the latter defined as a sequence of codons, each specifying amino acids and flanked by a start and a stop codon. However, this definition is not consensual across the scientific community. Since several authors use the terms CDS and ORF interchangeably when referring to a gene's CDS [58, 174] (including gene loss studies), we are also going to follow this paradigm throughout this thesis. Moreover, for avoiding any misunderstands regarding gene associated terminology, every time we are referring to genes, unless specifically mentioned, we are referring to nuclear eukaryotic protein-coding genes.

2.2 Pseudogenes

In contrast to functional protein-coding genes, pseudogenes have long been recognised as nonfunctional genomic sequences. Historically, the word 'pseudogene' was first mentioned in 1977 when Jacq et. al declared the existence of a group of untranscribed genomic sequences homologous to the 5S DNA in *Xenopus laevis* (African clawed frog) [82, 101]. Shortly after that, pseudogenes were found to be pervasive across multiple lineages including prokaryotes and eukaryotes [48, 61, 62, 100, 101, 106, 155]. The traditional view of pseudogenes assumed that these were considered relics of ancient genes, yet, not containing real functions [155, 211]. Even though this paradigm corresponds to reality for several identified pseudogenes, during the last two decades with the aid of the broad applications of NGS technologies, several pseudogenes revealed to be of particular importance by acquiring diverse functions in regulating development and diseases [89, 94, 101, 151, 152, 184].

Today, pseudogenes are defined as nongenic DNA elements with a high degree of similarity to functional protein-coding genes, but containing disruptive effects impeding them from being either successfully transcribed, translated into a functional protein, or both [213]. The number and constancy of pseudogenes is variable amongst species: for example, in humans, it is estimated that the number of pseudogenes ranges from 10 000 to 20 000 [209, 216] whereas in *D. melanogaster* (the fruit fly), only 110 pseudogenes have been identified [63].

According to their arising mechanisms, pseudogenes can be classified into three different classes (Figure 2.3). These include duplicated pseudogenes, processed pseudogenes and unitary pseudogenes [101, 213]. Duplicated pseudogenes originate from mutated or uncompleted gene duplication, in which process the duplicated gene becomes a pseudogene, whereas the parental gene retains its original function. These preserve their intron-exon like genomic structures and are capable of even displaying the upstream regulatory sequences of their parents [101, 155, 213].

On the other hand, processed pseudogenes result from the retrotransposition mediated by a retrotransposase of processed messenger RNAs (mRNAs) back to the genome. For this reason, they do not retain either intronic structures or upstream regulatory sequences, exhibiting only the exonic sequences that underlie their composition, and, usually, polyadenylation features at

their 3' end [186]. Importantly, unlike duplicated pseudogenes which are often located near their parental genes, processed pseudogenes are more likely to be detected in distant genomic regions or even in different chromosomes relative to them. Moreover, depending on their insertion site, they might either span an existent gene (the host of the processed pseudogene) or penetrate intergenic regions [191]. Finally, unitary pseudogenes result from previously functional protein-coding genes after accumulating *in situ* spontaneous mutations during evolution, therefore losing their primary functions. In this sense, and in contrast to processed and duplicated pseudogenes, unitary pseudogenes do not present any functional counterparts in the same genome, but may have orthologs in relative species. [11, 128, 213].

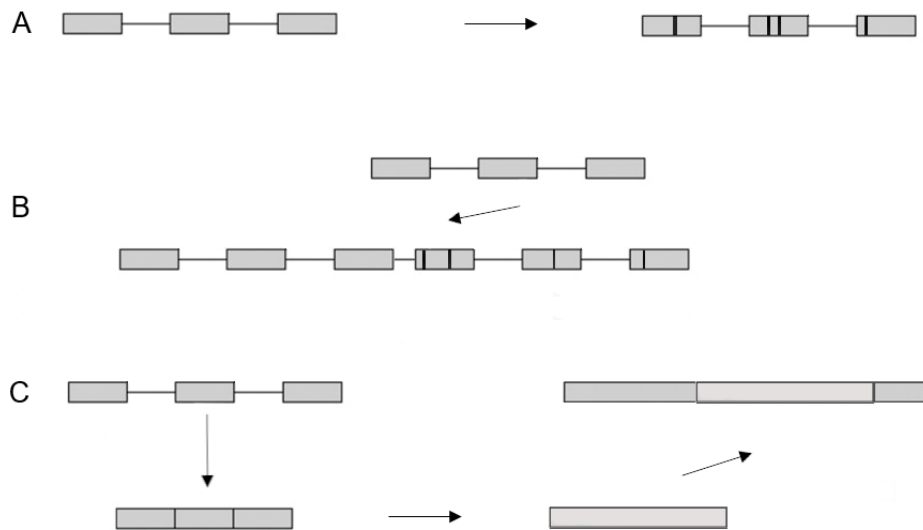


Figure 2.3: The origin and classification of pseudogenes. Adapted from [202]. Coding exons are represented by grey boxes and gene inactivating mutations are illustrated by black bars. (A) *In situ* gene inactivating mutations arising in a long-established protein-coding gene turn it into a unitary pseudogene. (B) Duplicated pseudogene (at the bottom) arising from incomplete or unfaithful gene duplication of a parental gene (at the top). As a consequence, the parent gene remains fully functional, whereas the duplicated copy turns into a duplicate pseudogene. (C) Processed pseudogene arising from the retrotransposition of a processed messenger RNA (mRNA) (on the left), further integrated back into the genome (represented by the longest bar, on the right).

2.3 Unitary pseudogenes

Unlike the abundant pseudogenes that arise from duplication or retrotransposition back into the genome, unitary pseudogenes comprehend the only class of pseudogenes that truly represent gene loss, thus, being especially relevant for this thesis. They represent relics from a previously established functional gene that lost the capability of encoding a functional protein over evolution

[172]. In contrast to the other classes of pseudogenes that underlie redundant loss, unitary gene inactivation directly affects the functional part of the genetic repertoire of a given organism [213].

Similarly, however, to the other two classes of pseudogenes, unitary pseudogenes present a set of inactivating mutations that impede them from being transcribed and/or translated into a functional protein. Concerning the latter case, these include deleterious mutations that will affect its coding sequence, that, in the case of a successful transcription, will culminate into the production of a shorter and/or abnormal protein product [11, 128, 213]. The three classes of inactivating mutations that usually comprise the coding sequence of unitary pseudogenes are below described.

- Nonsense mutation: consists of a point mutation in the coding sequence of a given gene that conducts to the appearance of an in-frame premature stop codon in the transcribed mRNA. As a consequence, this yields a premature end of the translation process, conducting to the synthesis of a truncated, incomplete and likely non-functional protein product [58] (Figure 2.4 B).

- Frameshift mutation: consists in the insertion or deletion (indel) of a number of DNA base pairs (bp) not multiple of three within the coding sequence of a gene. Given the triplet reading nature of gene translation, these indels will shift the reading frame of the original gene's coding sequence starting from the location of the indel and extending through the region encoding the carboxy-terminal of the protein. Therefore, this causes the entire amino acid sequence translationally downstream of the mutation site to bear no relation to the original amino acid sequence, and, in some occasions, might even result into the introduction of premature stop codons. As a final result, frameshift mutations typically lead to the complete loss of normal protein structure and function [58] (Figure 2.4 C).

- Splice site-disrupting mutations: correspond to mutational events (insertions, deletions, substitutions) that disrupt the consensus 5', 3' or both splice sites within an intron, thus, affecting the process of mRNA splicing. Usually, these mutations lead to the abnormal production of processed mRNA molecules including: 1) intron retention, in the cases when a given donor splice site is mutated, hence the intron is not excised since the first is not recognised by the spliceosome; 2) loss of exons, in the cases when the splice sites adjacent to a given exon are both mutated, thus, this being skipped during the process of splicing; 3) shifts in the open reading frame (ORF) of the molecule resulting from the specific combination of exons/exons plus intron(s) arising in the abnormally processed mRNA molecule. Similarly to the above-mentioned mutations, splice site-disrupting mutations also usually conduct to the synthesis of an abnormal protein product [12, 172]. Yet, splice site mutations do not affect the coding sequence of a gene in the cases where the latter only comprises a single exon (single-exon gene).

Finally, in addition to the mentioned gene inactivation mutations, unitary pseudogenes can also present greater levels of erosion such that their exons are completely removed in a given genome, or, simply, being no longer recognisable due to the large accumulation of mutations [172].

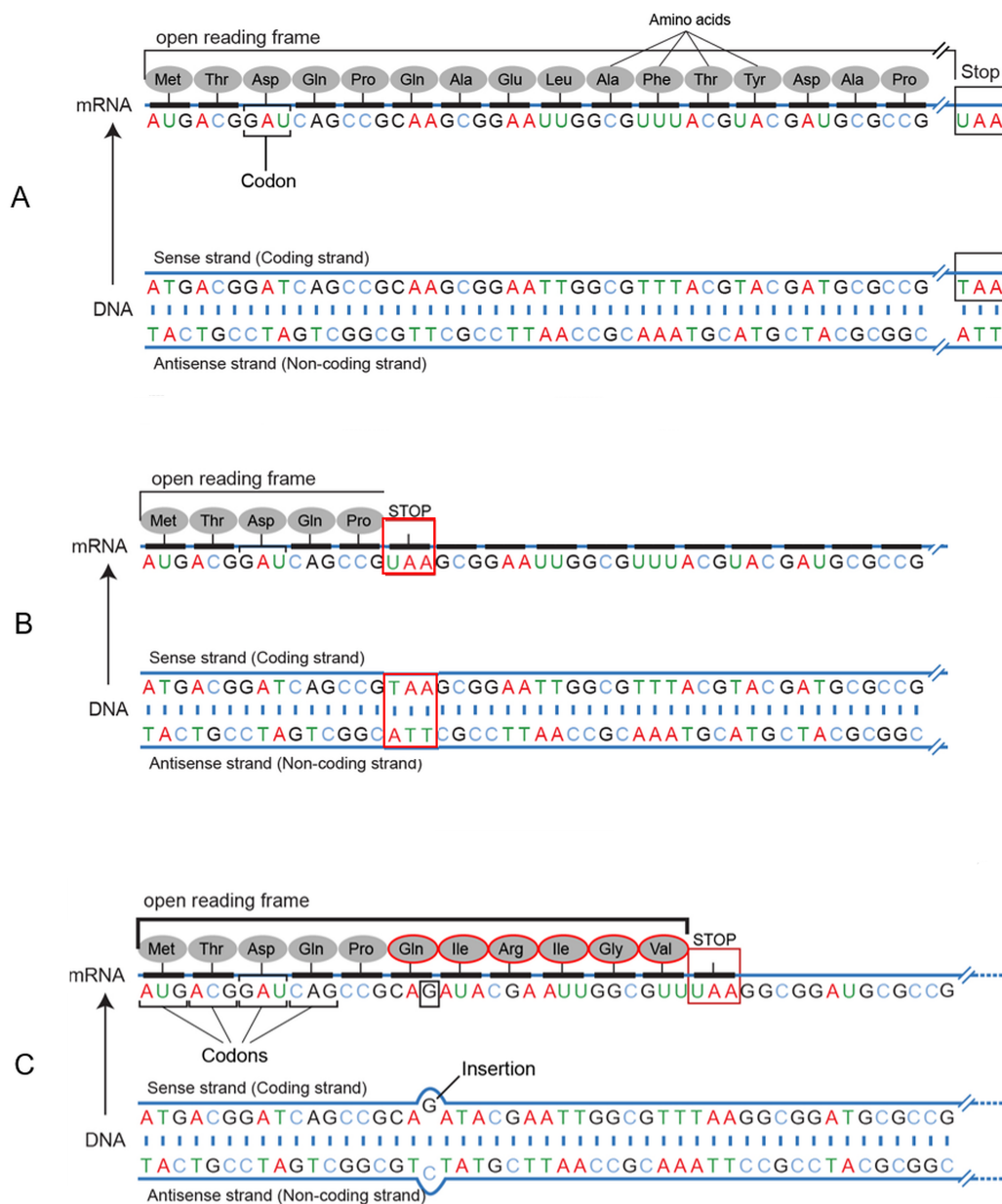


Figure 2.4: Schematic representation of the typical ORF-abolishing mutations found at unitary pseudogenes. Adapted from [1]. (A) Non mutated open reading frame/coding sequence of a protein-coding gene. The original DNA sequence, including the coding and non-coding strand, is represented below. Above, the corresponding mRNA molecule after transcription and processing is represented, as well as the corresponding amino acid encoded by each represented codon. (B) A point mutation leads to the appearance of a premature stop codon within the original coding sequence (CDS) of a gene (nonsense mutation), leading to the synthesis of a truncated, likely nonfunctional protein. (C) Frameshift insertion, leading to the shift of the reading frame underlying the gene's coding sequence, resulting in a shorter and abnormal protein amino acid constitution translationally downstream of the mutation site.

Chapter 3

Related work

3.1 First approaches towards the identification of pseudogenes

Bioinformatics and computational biology approaches have aided the understanding of the mechanisms underlying gene regulation and gene expression [162]. Genome annotation, the process of automatically identifying the functional elements within a given genome has been put in practice since the mid-90s as a result of the advent of the first published genomes [7]. Less explored was the development of automatised and semi-automatised approaches capable of identifying pseudogenes in the context of whole-genomes.

The prevalence of pseudogenes in mammalian genomes [128] impairs accurate genome annotation and might introduce artefacts in molecular experiments especially targeted at functional genes [216]. Due to their abundance, the identification of pseudogenes should comprehend an integral and significant part of whole-genome annotation, being, for this reason, critical to obtain a comprehensive and accurate catalogue of the functional and structural components lying on it [216]. One of the issues underlying pseudogene annotation, is that, given the similarity of processed and duplicated pseudogenes to their functional counterparts within the same genome, as well as due to the high degree of similarity that unitary pseudogenes exhibit relatively to their ancestral counterparts, these are often misincorporated into functional gene collections, with researchers then propagating these errors to subsequent analyses and interpretations [35, 204, 216].

In the attempt of overcoming these issues, several pipelines were proposed in the mid-2000s aiming for accurately annotating and identifying the repertoire of pseudogenes across several mammalian species. Furthermore, online databases such as pseudogene.org¹ [91] and Hoppsigen database² [4] were purposely created for storing and merging the diverse sets of annotated pseudogenes incoming from the pseudogene identification tools available at the time. Historically, the majority of the bioinformatics approaches capable of recognising pseudogenes via whole-genome screening were primarily designed for human and aimed to concretely identify either

¹www.pseudogene.org

²www.pbil.univ-lyon1.fr/databases/hoppsigen.html

processed pseudogenes, duplicated pseudogenes or pseudogenes of both classes [162].

Based on their computational approaches, these could be categorised into two different sets: homology-based approaches with posterior identification of deleterious mutational events, being frameshift and nonsense mutations [4, 210, 215], and approaches based on the relative quantities of synonymous to non-synonymous mutations [187]. Concerning the latter case, and according to the definition of pseudogenes, it is expected that their DNA sequences should be highly homologous to protein-coding genes, however, not under evolutionary constraints [101]. In this sense, some pipelines were proposed to identify processed pseudogenes by determining highly similar sequences to functional genes, yet, with values concerning their ratio of nucleotide nonsynonymous to synonymous substitutions (known as d_N/d_S or K_a/K_s ratio) relatively to their parents within the same genome closer to 1 [187]. Following this criterion, thousands of pseudogenes, including 8000 processed pseudogenes could be identified in humans [101].

Relatively to the homology-based approaches, three main bioinformatics tools were released: PPFINDER [190] and RetroFinder [10], that aimed to identify processed pseudogenes, and PseudoPipe [212], capable of identifying both processed and duplicated pseudogenes. PPFINDER (Processed Pseudogene Finder) [190] was introduced in 2006 and constituted the first standalone tool capable of identifying processed pseudogenes in mammalian genomes. It was purposely created for addressing the issues underlying the genome annotators of the epoch that frequently mistaken processed pseudogenes for real genes or exons, conducting to biologically irrelevant gene predictions. The purpose of this software was not to capture all processed pseudogenes within a given genome but to improve the accuracy of existent genome annotations by identifying processed pseudogenes over annotated (functional) protein-coding genes. PPFINDER starts by establishing a gene model (this is, a representation of an mRNA transcript of a protein-coding gene containing information about its features such as exon-intron boundaries, splice sites and UTRs) for each annotated gene, and further uses this information for aligning the pre-established gene models to a complementary DNA (cDNA) database of the same testing organism using the BLASTn algorithm. Then, the transcripts of the same database corresponding to the best scoring BLAST hits are aligned against the tested genome, and, if a given transcript hits more than in one location, the transcript is marked as pseudogene-derived. Otherwise, the transcript is not marked as pseudogene-derived.

Retrofinder [10], released in 2008, on the other hand, comprehended a unique among the available human processed pseudogene prediction methods by using mRNA sequence alignments for their identification. To accomplish this purpose, Retrofinder uses a sensitive pairwise DNA sequence aligner for aligning human mRNA and RefSeq sequences from the National Center for Biotechnology Information (NCBI) against the human genome. If one of these transcripts aligns more than once at the target genome, and if one of the alignments corresponds to a known gene *locus*, then the remaining alignments are scored on a number of features indicative of retrotransposition such as multiple contiguous exons relatively to the corresponding parent gene, absence of introns, presence of a polyadenine tail in the 3' end, lack of conserved splice sites, and the absence of conserved syntenic with mouse and dog genomes using the syntenic net

alignments from the University of California, Santa Cruz Genome Browser³ [50].

Finally, PseudoPipe [212] is capable of identifying both processed and duplicated pseudogenes within mammalian genomes. For achieving this, Pseudopipe requires a target mammalian genome, the set of non-redundant protein sequences underlying the same genome and the chromosomal coordinates of functional genes. Pseudopipe then searches for homology of all known protein sequences within the genome by making use of a six-frame translational BLAST search (tBLASTn). The BLAST hits are further analysed, redundancies are removed, overlapping hits are merged and those corresponding to known functional genes are ignored. The remaining hits are analysed, and, for each hit, the corresponding functional parent gene is identified. Then, a refined alignment between the target genomic region corresponding to each hit and the protein query corresponding to the identified parental gene is executed. The alignment is inspected and deleterious mutations are determined. Based on their genomic structures, the pseudogenic fragments are further categorised into different classes: duplicated pseudogenes, if displaying intron-exon like structures, processed pseudogenes, if lacking introns and displaying traces of a 3' polyadenine tail, and ambiguous when corresponding to pseudogenic fragments that are too decayed for reliably assessing whether these correspond to processed or duplicated pseudogenes.

3.2 Identification of gene loss

The obvious issue of the above-mentioned bioinformatics approaches lies in their inability to capture real gene loss, being only capable of detecting mere loss concerning either duplicated or retrotransposed copies underlying a parent gene.

As preluded in Chapter 1, formally, for a gene to be classified as lost, it is required that a lineage, which descends from an ancestor presenting an intact protein-coding gene, displays for the corresponding orthologous genomic *locus* either its full absence or gene inactivating mutations that likely result in the synthesis of a non-functional protein (establishing a unitary pseudogene) [172, 213]. Since the identification of gene loss implies a phylogenetic contextualisation, the major conceptual difference between the detection of unitary pseudogenes and the other classes of pseudogenes is the mandatory need of a comparative analysis between at least two species: a reference species, and a target species where the loss of a given gene is being inferred [172, 213, 214].

After a speciation event occurs, the increasing divergence between two resultant species reflects the decrease in their genetic orthology since gains and losses of genes gradually accumulate in each of them. Hence, the presence of unique genes in one species relative to another indicates either gene gains in one or gene losses in the other. However, based on the principle of parsimony, if orthologous genes are present across several species but absent in just one, a single gene loss event is much more plausible than multiple simultaneous gene gain events [213]. In practice, to identify either unitary pseudogenes or the full absence of a set of genes in a given species, a reference gene set from another species is needed. This does not represent a mere operational

³www.genome.ucsc.edu

convenience or necessity, likewise functional orthologs, unitary pseudogenes in particular and species' genetic repertoire in general, are conceptually comparative entities since all species are related to one another through speciation events [213].

For instance, to identify human gene loss, mouse or chimpanzee gene sets might be used as references. Particularly, if the human gene loss occurred after the human-chimp divergence and if the mouse and the chimp orthologs are both functional and conserved, either chimp or mouse orthologs can be used as a reference for inferring it (Figure 3.1 A). However, if the human gene loss has occurred between the human-mouse, before the human-chimp evolutionary divergences and the mouse ortholog is conserved, the human gene loss is only meaningfully identifiable when the mouse gene set is used for the comparison (Figure 3.1 B) [213].

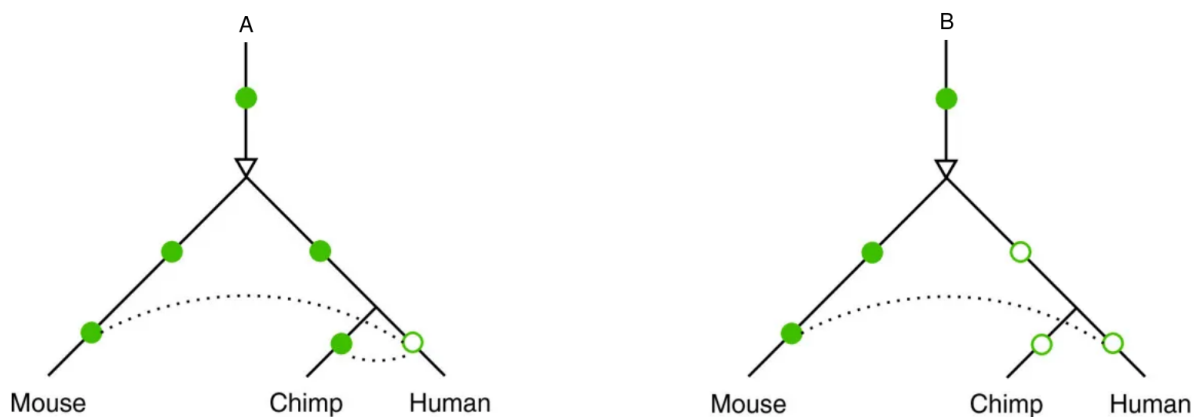


Figure 3.1: Relativity of gene loss identification. Adapted from [213]. Given the above-represented phylogeny of human, chimpanzee, and mouse, human gene loss can arise in different evolutionary moments, namely: (A) in the human lineage after the human-chimpanzee divergence; (B) at the human-ancestral lineage after the human-mouse divergence and before the human-chimp divergence. Functional orthologs are represented by filled green circles, lost genes are represented by green circumferences. The last common ancestor of the represented lineages is represented by an inverted triangle. For each of the two represented cases, and to perform gene loss inference, the biologically meaningful relationship between functional orthologs and the lost counterparts is represented by a dashed line.

In truth, the very first approaches underlying mammalian gene loss identification were indeed focused on human-specific gene losses. Motivated by understanding the genomic signatures underlying the phenotypic divergences between humans and chimpanzees, several gene loss studies were conducted in the early 2000s [60, 196]. Particularly, in 2006, 76 human-specific gene loss events were reported through the comparison of human's already established repertoire of putatively unitary pseudogenes and the draft chimpanzee genome sequence available at the time. Just a few months later, comparative genomics analyses were also conducted for identifying gene losses associated with the human Y-chromosome. By aligning both Y-chromosome sequences from human and chimpanzees, the authors were able to identify ORF-abolishing mutations in diverse human Y-chromosome associated genes [148].

Nonetheless, only a few years ago the expansion of gene loss studies to non-human mammals started to emerge. With the aid of the recent mammalian genomic data, as well as with the development of comparative genomics-based gene loss inference methodologies, the study of gene loss underlying this *taxon* has been increasing over the last decade [172]. Several gene loss identification techniques are nowadays recurrently used, and based on their approach, these can be classified into three major categories.

- Absence of orthologs approach: isolated this approach is infrequently used. It aims to identify gene loss by, at a given set of species, infer the absence of the orthologous counterparts of a gene set or protein set of interest. To accomplish this, the gene set or protein set of interest from a well-established reference species' genome is retrieved and is further heuristically aligned against the target species' genome(s). Finally, each query gene (or gene encoding each protein of interest), is, per target species, classified as absent (lost) or coding according to predetermined defined criteria concerning the corresponding alignment hit quality [87, 213].

- Gene annotation-based approach: in contrast to the previous approach, this has been being widely applied in gene loss studies [108–111, 141] and implies the manual annotation of a given gene or set of genes of interest at a given target species or set of target species corresponding genome(s), with each annotated product being further classified into coding or pseudogenised. Since this methodology is often used for inferring the coding status of specific genes, it can be seen as a targeted analysis-based approach. It can be applied not only in situations where a given gene of interest is not annotated at the target species' genome (either due to the unannotated nature of the genome itself or given the great erosion of its orthologous counterpart such that automatic genome annotators do not even recognise its presence on it) but also in situations where, derived from automatic whole-genome annotations, previous annotation of the ortholog of interest already exists. Concerning the latter case, the necessity of re-annotating the candidate gene over its corresponding orthologous existent annotation is based on the assumption that unitary pseudogenes are often erroneously automatically annotated as functional protein-coding genes, with whole-genome annotators forcing ORF-deleterious mutations such as any shift to the coding frame or premature stop codons to be automatically corrected, foreseeing that such mutations arise from sequencing or assembly artefacts.

This approach is fundamentally separated into two main components: for each target species and candidate gene, the orthologous genomic sequence to the candidate gene is retrieved, and next, the candidate gene annotation is performed over the previously retrieved genomic sequence. In detail, concerning the first component of the approach, if the candidate gene's annotation is already available at the target species' genome, the corresponding genomic sequence is directly collected. For the cases where the target species' genome is annotated, but not the candidate gene, the syntenic conservation of the candidate gene *locus* across the reference and in-study species is first assessed, and, if confirmed, the target species' genomic region flanked by the orthologous neighbouring genes of the candidate gene at the reference species is collected. Finally, regarding the cases for which the genome of the species of interest is not annotated at all, a heuristic alignment between the candidate gene's coding sequence at the reference species and

the genome itself is conducted. Then, the genomic scaffold or genomic contig corresponding to the best alignment hit is further retrieved.

During the gene annotation step (the second component of the approach) the DNA sequences of the individualised coding exons constituting the gene of interest (5' and 3' UTR flanked) are aligned against the collected target species' genomic sequence, and each predicted exon in the target species is screened for ORF-abolishing mutations including the presence of in-frame premature stop codons, frameshift mutations, as well as splice-site disrupting mutations (for genes with the corresponding CDS spanning at least 2 exons). For ensuring that the identified deleterious mutations do not result from sequencing or assembly artefacts, commonly, at least one identified mutation is validated with unassembled genomic sequencing reads from independent genomic sequencing projects from the target species (when available).

- Whole-genome alignment-based approach: this approach was much more recently proposed (2018) [172] and involves conducting multiple pairwise whole-genome alignments between a reference species with high quality, well-annotated genome (usually human) and a set of either annotated or not annotated mammalian genomes (the informant species). Likewise the previous mentioned approach, this approach is based on the principle of protein-coding gene's conservation across evolution (further explored in the next chapter), and, for each annotated gene at the reference species, it aims to project the annotation of each constituting coding exon over the different target genomes, thus, culminating in a *de novo* whole-genome annotation-based in the protein-coding gene's repertoire of the reference species. Then, for each gene of interest, the corresponding ortholog at each target species is inspected, and similarly to the previous approach, each coding exon is screened for ORF-disrupting mutations, with at least a single gene inactivating mutation being validated using unassembled genomic sequencing reads from the same species (when available). The percentage of intact open reading frame (ORF) is further calculated, and a threshold that defines the minimum percentage of intact ORF needed for considering a gene as coding is established. Finally, for each analysed target species, following the defined criteria, each gene of interest is then classified as functional or lost. Likewise the previous approach, this has also been successfully applied in several mammalian gene loss studies [67, 84, 170, 173].

Importantly, each referred approach has its pros and cons. The selected methodology should be carefully selected according to the biological questions to be addressed, the required level of information on mutational evidence, and the computational resources available for that effect. For instance, since it uses heuristics in every step of the analysis, the absence of orthologs approach displays an advantage over the two remaining ones concerning analysis execution time [177]. On the other hand, this approach is not informative on mutational evidence and might even fail for the cases wherein a given species, a given gene is lost but the orthologous reference counterpart used for inferring it, still positively aligns, although in a non-orthologous genomic *locus*, leading to the false inference of gene functionality, underestimating the number of lost genes at the same species.

In contrast, the gene annotation-based approach is way more informative, it tends to accurately identify deleterious mutations, lost exons, and, indirectly, is able to assess the conservation of the identified mutations across the tested species. However, since this approach relies on extensive manual curation, depending on the tested candidate gene (as the number of coding exons varies across genes [165]), as well as in the number of the tested target species (in the sense that, for each one, the gene annotation process needs to be replicated), it might culminate in a labour-intensive and time-consuming process.

Relatively to the whole-genome alignment-based approach, this has the advantage of, once the pairwise genome alignments between the reference and informant species genomes are completed, being easy to infer the coding status of any protein-coding gene (as long as annotated at the reference species genome) at any of the predetermined set of target species. Nevertheless, this methodology displays some serious drawbacks: first, whole-genome alignments require the download of each intervenor genome, that, depending on the genome quality and nature of the corresponding species might occupy several gigabytes of digital space; second, whole-genome alignments, even using algorithmic heuristics are computationally demanding and may fail when confronted by chromosomal rearrangements such as insertions, deletions, or translocational events [7]; third, following each pairwise genome alignment, a process of quality masking and a series of filtering processes must be conducted for reducing the artefacts underlying the aligned product; fourth, since whole-genome alignments are by nature not aware of exon reading frame and splicing sites, these often display at the predicted coding regions spurious frameshift gaps that require the usage of additional software such as CESAR [171] for accurately realigning each coding exon considering the reading frame and splicing sites; and finally, if at a certain point in the analysis the researcher decides to include a new target species on it, it might not be very convenient to run all the previously mentioned steps, required for its inclusion in the study.

Notably, all of these approaches have one aspect in common, they are not straightforward, intuitive or easy to use, therefore preventing their large-scale operation by non-experienced users and impeding their expedite scalability to the quickly emerging mammalian genomic data. Moreover, their lack of true automation (with some of them even comprising fully manual procedures), the high computational resources that some, particularly the whole-genome alignment-based approach requires and the absence of an intuitive user interface constitutes a major disadvantage.

Chapter 4

Pseudo*Checker*

4.1 Introduction

In response to the weaker points of the previously mentioned gene loss inference methodologies and taking into account the predetermined goals (see Section 1.5), we developed Pseudo*Checker*, an integrated online platform for automatised gene loss inference in mammals. In contrast to the existent approaches, Pseudo*Checker* aims to, in a systematic, highly accurate and fully automatic manner, classify the coding status of any protein-coding gene of interest as intact or lost (with different levels of pseudogenisation), in single or multiple mammalian target species by incorporating available corresponding genomic data.

The idea underlying Pseudo*Checker*'s development arose from the necessity of automating a particular existent mammalian gene loss inference methodology, already capable of accurately identifying lost candidate genes, the respective lineages and the underlying gene inactivating mutations, yet, lacking the necessary intuitiveness and speed required not only for accompanying the quick genomic data emergence, but also to allow its large-scale applicability by general, non-experienced users. Particularly, our novel comparative genomics-based approach is the result of a strong inspiration based on the fully manual gene annotation-based approach implemented by Lopes-Marques in several mammalian gene loss studies (overviewed in the last subsection of the previous chapter) [108–111], now, embedded as part of Pseudo*Checker*'s logical core.

On its basis, Pseudo*Checker* constitutes a bioinformatics web service comprising a fully three-step integrated pipeline that makes use of minimalist user input and a set of established parameters necessary for running a given analysis. However, and as we will further explore, our approach is not simply an automatised pipeline aiming to recognise gene loss. Since Pseudo*Checker* displays the ultimate goal of endorsing the overall convenience of gene loss inference, our approach was specifically designed to constitute a general-purpose, study-independent standalone bioinformatics tool, capable of, not only accurately unveiling the genomic signatures left by gene loss, but also displaying a set of unique features required for levelling up the scalability, effectiveness and quickness of gene loss prediction.

Explicitly, *PseudoChecker*, currently in Version 0.8, displays the ability of:

- Automatically, in a user convenient way and comprising an intuitive graphical user interface (GUI), identifying gene loss and the respective mutational evidence for a set of unlimited target species with available genomic data;
- Unveiling ancestral gene loss events by accurately recognising and displaying conserved mutations across closely related *taxa* within a given analysis;
- Objectively measuring the erosion of the gene of interest (the candidate gene) at a given target species by assigning it an index of pseudogenisation, the PseudoIndex;
- Including external functional gene datasets into the analyses;
- Exporting the produced data throughout an analysis, useful for performing downstream complementary tasks including phylogenetic reconstructions and selection analyses;
- Executing a complete analysis in a short amount of time, from less than a second to a few minutes (see Chapter 5).

Indeed, our approach displays several stronger points over the existing methods reviewed in the previous chapter. Table 4.1 highlights the strongest innovations of our method, depicting a comparison between the different features and requirements of each available gene loss inference approach, including ours.

Table 4.1: Comparison of the different features and requirements regarding each available gene loss inference approach, including *PseudoChecker*. (i) Reveals mutational evidence. (ii) Makes use of specialised algorithms capable of aligning pseudogenic and viable coding sequences. (iii) Includes an index of pseudogenisation. (iv) Does not require whole-genomes. (v) Fully automatic. (vi) Works in the cloud. (vii) Includes a user-friendly graphical user interface (GUI). (NES) Not every study.

Approach	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)
Absence of orthologs approach							
Gene annotation-based approach	✓	NES		✓			
Whole-genome alignment-based approach	✓						
<i>PseudoChecker</i>	✓	✓	✓	✓	✓	✓	✓

Throughout the next subsections, we will start by exploring every detail of our approach, from *PseudoChecker*'s pipeline overview to each algorithmic detail underlying it, culminating in Section 4.5, where we will illustrate three different examples of analyses ran with *PseudoChecker*: one simulated test case that will serve for exploring the tool as a whole, and two analyses underlying real test cases, one of which comprising a gene that is functional in every tested target species, and the other regarding a lost gene in several mammalian species.

4.2 Overview of PseudoChecker's structure and pipeline

For annotating protein-coding genes in a genome of a newly sequenced species, a set of different approaches can be used. First, on the experimental side, genes can be annotated by sequencing partial or full-length mRNA transcripts, followed by the mapping of the obtained reads or assembled transcripts back to the genome, though, with the disadvantage that only genes with a reasonable high expression in the sampled tissues or cell types can be identified [19, 171]. As a complementary approach, *ab initio* methods tend to predict genes by taking into account statistical sequence patterns while screening a given genome [150]. Yet, since genes' coding sequences are reasonably conserved across species, a much more convenient procedure is to make use of already annotated genes in a given (reference) species for performing gene annotation in a set of related species corresponding genomes [171].

This principle is therefore exploited in homology-based approaches, with these including the ones that comprise whole-genome annotators, the vast majority of gene loss inference methodologies (gene annotation-based and whole-genome alignment-based approaches), and, now, PseudoChecker. These comparative exon and gene finding methodologies make use of cross-species gene conservation and rely on the fact that reading-frame preserving insertions, deletions (indels) and synonymous changes are more frequent in coding exons rather than in other genomic regions [171].

Following this rationale, PseudoChecker, in a, yet, simplistic view, per analysis, infers the coding status of a given candidate gene in a given target species by annotating its orthologous counterpart, followed by the screening of the annotated sequence for gene inactivation mutations. Therefore, our approach is aimed to be applied in two different circumstances: 1) *de novo* candidate gene annotation in yet, unannotated mammalian genomes when the inference of its the coding status is for some reason relevant to know; 2) re-annotation of candidate genes in previously annotated genomes, however suspiciously erroneously annotated as functional protein-coding genes.

For accomplishing this, hence, running a PseudoChecker analysis, we simply require two main user inputs:

1. The candidate gene's nucleotide coding sequence (CDS) and underlying coding exon(s)' nucleotide sequence(s), both annotated and retrieved from a given (reference) species (FASTA format) (If more than one gene isoform is annotated, for instance, if different splice variants are annotated, the user must choose one of these);
2. Per target species, the genomic sequence at where the candidate gene's orthologous counterpart will be annotated/predicted (FASTA format). The user is responsible for ensuring that the inserted sequence is orthologous to the in-study gene of interest. This can, for instance, include a genomic sequence underlying the in-study gene's annotation, derived from automatic whole-genome annotations, or in general, to a genomic sequence for which the user believes that the putative orthologous gene or pseudogene of the gene of interest

is physically located. As target sequences, *PseudoChecker* supports either partial/full genomic contigs, scaffolds and long sequencing reads.

Optionally, and even though the following feature does not constitute the fundamental part of the purposes of *PseudoChecker*, we also support the assignment of predetermined (functional) coding sequences, obtained for instance, from biological databases or experimental analyses (FASTA format). This optional inclusion is particularly useful for the cases when the user pretends to compare predicted functional or pseudogenised sequence(s) by *PseudoChecker* with already existent ones.

Finally, once the input data is correctly assigned and the parameters underlying the different components of the three-step integrated pipeline are selected, the latter is executed:

1. Coding sequence prediction: For each target species, *PseudoChecker* annotates the orthologous coding exons and, consequently, the viable (or not) orthologous coding sequence of the in-study gene by performing a deterministic pairwise alignment of each reference species' coding exon against the corresponding inputted genomic sequence;
2. Alignment by MACSE [159]: Once the first step is concluded, *PseudoChecker* runs MACSE, a specialised standalone software capable of producing a deterministic (pairwise or multiple) alignment between both predicted (coding or pseudogenised) sequence(s), predetermined coding sequences (if exist), and the reference species' (functional) coding sequence, taking into account the eventual presence of frameshift and premature stop codons;
3. Report of the mutational evidence and PseudoIndex computation: For each target species, and according to the previous alignment produced by MACSE, eventual cross-species conserved and non-conserved gene deleterious mutations, including frameshift mutations and in-frame premature stop codons relative to the reference species' CDS, splice site disrupting mutations, full gene loss or exon loss are revealed. Finally, and again, per target species, taking into account the underlying absence, presence and level of mutational evidence, a pseudogenisation score, the PseudoIndex, is assigned to it.

Structurally, and since *PseudoChecker* constitutes an online platform, it is compartmentalised into two main components: the front-end system and the back-end system.

- Front-end system: comprises the *PseudoChecker*'s graphical user interface (GUI), therefore, the first layer of interaction between the user and *PseudoChecker*'s logical core and integrates the web pages that precede and are dynamically produced throughout a given analysis. This includes the main web page required for biological data input and parameter selection, the main results web page that provides, in a conveniently structured interface, the analysis' results and, finally, per target species, a web page dynamically produced underlying the coding sequence prediction step of the pipeline, accessible through the main results page;
- Back-end system: comprises all the pieces of code and additional software (MACSE) executed during a *PseudoChecker* analysis.

The front-end system was entirely implemented in HyperText Markup Language (HTML) + Cascading Style Sheets (CSS) and Javascript, whereas the back-end system, except for MACSE (the unique standalone tool used in the pipeline, written in Java) was developed in Python 3. The establishment of the connection between the *PseudoChecker*'s front-end and back-end systems was developed resorting of PHP.

At this stage, *PseudoChecker* is not yet online, but it is completely fully functional within the machine used for its development. The entire system was developed in a local machine comprising a central processing unit (CPU) with 6 physical cores and 16 gigabytes of random access memory (RAM), by making use of a locally hosted web server managed by PHP 7.2 Development Server. *PseudoChecker*'s entire conceptualisation is summarised in the figure below.

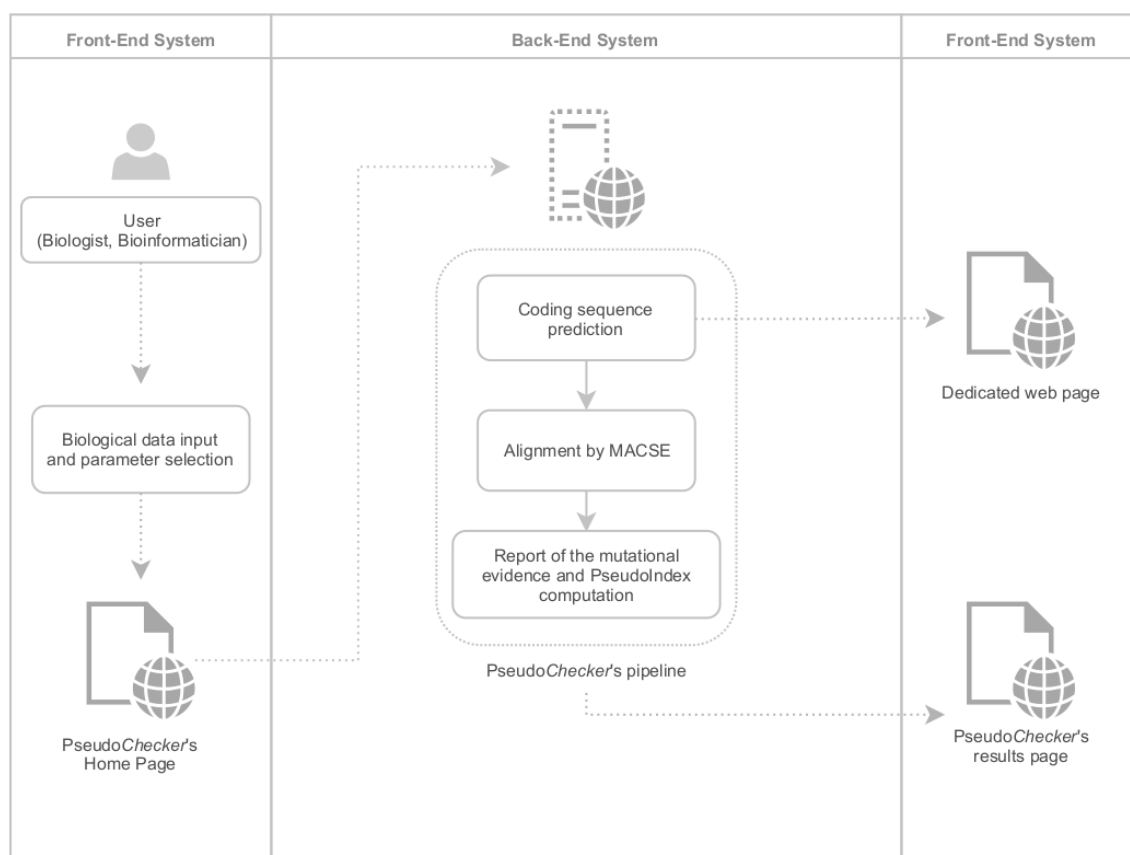


Figure 4.1: *PseudoChecker*'s conceptualisation. From left to right and top to bottom: The user inputs the biological data at *PseudoChecker*'s home page, selects the adequate parameters underlying the different steps of the pipeline and starts the analysis. A PHP request is sent to the server and *PseudoChecker*'s pipeline runs internally as a single process: For each target species, the orthologous sequence to the reference species' candidate gene coding sequence (CDS) is annotated and a dedicated HTML page for the visualisation of the outcome of this process is produced; MACSE performs the alignment between the annotated sequence(s), predetermined coding sequences (if existing) and the reference species' (functional) coding sequence; Finally, for each target species, gene inactivating mutations are revealed, and its PseudoIndex computed. The results are revealed in a user-friendly webpage, the *PseudoChecker*'s main results page.

4.3 *PseudoChecker*'s pipeline in detail

4.3.1 Coding sequence prediction

4.3.1.1 Global approach

As mentioned in the previous subsection, coding sequence prediction corresponds to the first step of *PseudoChecker*'s three-step integrated pipeline. This component is responsible for annotating, at each target species' genomic sequence, the orthologous sequence to the reference species' candidate gene's CDS.

DNA pairwise sequence alignment constitutes the most fundamental problem in genomics. The alignment problem aims to establish the mapping between the characters of two sequences that approximates some relation underlying the user's interest. By definition, an alignment of two strings is formed by inserting spaces in arbitrary locations along each of them so that they end up with the same length and without spaces at the same position of the two augmented sequences. Notably, in comparative genomics, the ultimate goal of DNA alignment lies not only in understanding the true history of change between sequences, but also to identify coding regions, the transcriptional control regions lying beyond genes' coding sequences, to determine evolutionary distances, aiding genome assembly and finding common regions between two genomes [7, 75].

Three main alignment problems are currently proposed: global alignment, local alignment and semi-global alignment. The global alignment problem is aimed to compare two strings s and t for cases where it is believed that both sequences are related along their entire length. Therefore, this alignment problem is usually applied in comparisons of DNA sequences that naturally present roughly the same length, being genes' coding sequences, orthologous gene fragments or polymerase chain reaction products obtained from the same molecule. On the other hand, local pairwise alignments aim to detect similarities between pairs of small fragments of both sequences, this is, pairs of substrings of s and t , useful for detecting conserved motifs within two sequences that, globally, might be quite dissimilar. Finally, the semi-global alignment problem, a hybrid method between the first two problems, is recurrently used in cases where it is believed that s and t are related along the entire length of the region at where they overlap [13, 14].

This hybrid method constitutes a small variation of the global alignment problem, and while an optimal global alignment seeks the best, full-length alignment of two sequences, in other words, algorithmically, the best way to match up two sequences along with their entire match, for some applications, however, within an alignment, it is desirable to relax this requirement by not penalising gaps at the beginning of one of the sequences and/or at the end of one of the sequences [182].

The semi-global alignment problem (also known as global alignment with free end gaps) is then typically applied for specific bioinformatics tasks that comprise, for instance: genome assembly, where we seek overlap between sequencing reads, this is, to align the end of a read with the beginning of another, hence, expecting some gaps at the beginning and end of alignment; alignment of intronless genes against an entire genomic region, that, since they will be completely covered by the genomic DNA, gaps at both beginning and end of each gene are expected; and finally, in eukaryotic coding exon alignment against genomic sequences, in the sense that is expected that a coding exon matching an entire gene will be flanked by either non-genic regions, untranslated regions (UTRs) or intronic regions, and, for this reason, adjacent gaps to the exon should not be penalised [182].

A comparative example between the optimal pairwise alignment for two simulated DNA sequences produced by the three different alignment problems and following the same maximisation criterion is depicted in Figure 4.2.

```

ATCCGAACATCCAATCGAAGC
A---G--CATGCAAT-----
A

CATCCAAT
CATGCAAT
B

ATCCGAA-CATCCAATCGAAGC
-----AGCATGCAAT-----
C
```

Figure 4.2: Optimal pairwise DNA sequence alignment for the simulated sequences A: ATCCGAACATCCAATCGAAGC and B: AGCATGCAAT following three different alignment problems and the same alignment maximisation criterion. (A) Global alignment: gaps are penalised throughout the entire alignment, end to end; (B) Local alignment: the best matching subsequences of each sequence are aligned; (C) Semi-global alignment: gaps are not penalised, in this case, in both ends of the bottom aligned sequence.

Given the above explanation, in this component of Pseudo*Checker*'s pipeline, we conduct semi-global alignments for aligning each reference species' candidate gene's coding exon against each target species' genomic sequence. Only using this approach is possible to, algorithmically speaking, achieve the best pairwise alignment that fulfils our needs. Likewise in global- and local-alignment

problems, it is possible to compute, following a given maximisation criterion, the optimal semi-global alignment of two sequences using dynamic programming algorithms. In our approach, we are using the variation of the classical global alignment Needleman-Wunsch algorithm (1970) [140] implemented in the module Pairwise2 from Biopython [31], a Python package that contains freely available tools for computational molecular biology and bioinformatics.

In detail, when aligning each coding exon of the reference species' candidate gene, we do not penalise the corresponding upstream and downstream end gaps. In other words, we are not assigning any cost to gaps that appear before the first and after the last character of the reference species' exonic sequence. As a result, we obtain, for each exon alignment, the reference species' coding exon aligned sequence and the target species' genomic aligned sequence as represented in Figure 4.3 A.

However, for obtaining a more intuitive and relevant alignment, *PseudoChecker* deliberately trims the non-relevant part of it by ignoring the target species' nucleotide bases overlapping the upstream and downstream end gaps of the reference species' coding exon aligned sequence, and removes the same end gaps found at the latter sequence (Figure 4.3 B). Then, for a given target species, the predicted exonic sequence orthologous to the (aligned) reference species' coding exon corresponds to the previously trimmed target species' genomic aligned sequence with its internal gaps collapsed (Figure 4.3 C).

```

      ATAGGAA---CATCCAATGTAAGC
      ----GAAAAACATGCAAT-----
              A

      GAA---CATCCAAT
      GAAAAACATGCAAT
              B

      GAACATCCAAT
              C

```

Figure 4.3: Illustrative example of an optimal semi-global exon alignment produced by *PseudoChecker*. (A) Original semi-global alignment between a simulated reference species' coding exon (sequence at the bottom) and simulated genomic sequence of a given target species (sequence at the top). (B) Trimmed alignment that represents only the relevant part of the original alignment. (C) Target species' predicted orthologous exonic sequence to the reference species' one.

Therefore, per target species within a given analysis, by aligning, from the 5' to the 3' end of the reference species' gene, each composing coding exon against the test species inputted genomic sequence, the test species' predicted orthologous sequence to the reference species' CDS constitutes no more than the result of a progressive concatenation of each predicted coding exon's sequence underlying each conducted alignment (following the same order).

For performing accurate gene's (viable or not) coding sequence annotation, however, the reference species' 5' and/or 3' coding exon must not display any UTR fragment. This implies that the first codon of the reference species' 5' coding exon must always correspond to 'ATG' and the last codon of the reference species' 3' coding exon should always correspond to a stop codon ('TGA', 'TAA', or 'TAG') (the same rationale is applied for Pseudo*Checker* analyses comprising intronless genes). For the cases where at least one reference species' coding exon contains an UTR fragment on its sequence, Pseudo*Checker* automatically trims the UTR fragment from it (as long as the user indicates its presence in Pseudo*Checker*'s home page by selecting the option 'Coding exon(s) contain(s) UTR(s)' in the UTR trimming parameter selection box).

Moreover, for this component of its pipeline, Pseudo*Checker* also requires the selection of another parameter: the minimum percentage of exon alignment identity for an alignment to be considered as a positive alignment. This parameter is of utmost importance since it will define if a given exon alignment corresponds to a real, biologically meaningful alignment, rather than a non-specific alignment. By default, Pseudo*Checker* sets this value to 65%, with this being the one we found to be adequate during Pseudo*Checker*'s development. Nevertheless, it can still be manipulated by the user at Pseudo*Checker*'s home page in order to fulfil its necessities. For instance, if the reference species is highly divergent from the test species, the user might prefer choosing a lower value for this parameter, whereas, for the opposite case, a higher value might be the right option.

Each predicted exon for which the corresponding alignment identity is inferior to this value will be considered as lost at the corresponding target species (or so eroded that any similarity is destroyed), therefore, being excluded from the final annotated/predicted sequence. Sequences under this condition are considered partial coding sequences. In contrast, sequences that comprise all predicted orthologous exonic sequences to the reference species' ones are considered full coding sequences.

After each positive exon alignment at a given target species' genomic sequence, Pseudo*Checker*'s will automatically force any upcoming exon alignment to occur at a target species' downstream genomic region relative to the one underlying the previous exon alignment. This is required since naturally, 3' coding exons must always occur after 5' coding exons. As a consequence, since Pseudo*Checker* will progressively shorten the genomic sequence of the species in question, this will also progressively decrease the time of alignment execution.

For increasing even further the quality of coding sequence prediction, as an additional, and entirely optional feature, in this component of the integrated pipeline, we provide the user with the opportunity of automatically extending the reference species' 3' coding exon (or single-exon)

alignment for searching a hypothetical downstream in-frame final stop codon, when this missing in the original alignment. This feature is particularly useful for the cases where the C-terminus of the protein encoded by the in-study gene is slightly divergent in size across the tested mammalian lineages, and, as a consequence, the original alignment is not capable of detecting the more distant test species' final stop codon relative to the reference species' one (Figure 4.4).

```

AATGAACCCAAATTAACCCAG
AATGTACCGAACTTAACCTAG

A

AATGAACCCAAATTAACCCAGAATTTCTGA
AATGTACCGAACTTAACCTAG#####

B

```

Figure 4.4: Illustrative example of the extension of an exon alignment concerning a simulated reference species' 3' exon. The reference species' aligned exonic sequence is represented as the bottom sequence, and the target species' simulated genomic aligned sequence is represented as the sequence at the top. (A) Original alignment. (B) Extension of the original alignment for finding a downstream in-frame stop codon (represented by # at the reference species' 3' exon aligned sequence.)

Yet, for avoiding an exaggerated, likely non biologically meaningful extension of the alignment, *PseudoChecker* restricts the alignment extension to a maximum range of 5 amino acids (15 nucleotides). For activating this feature, the option 'Automatically extend reference species' 3' coding exon (or single coding exon) alignment to find a missing (in the original alignment) downstream final stop codon:' must be selected at *PseudoChecker*'s home page, prior to the execution of a given analysis.

Finally, in this step of the pipeline *PseudoChecker* categorises the predicted coding sequences into two different categories:

- Reliable sequences: those not presenting frameshift gaps and/or in-frame premature stop codons underlying each composing coding exon alignment;
- Less reliable sequences: those presenting frameshift gaps underlying at least a composing exon alignment and/or displaying an in-frame premature stop codon within the predicted sequence.

As we will further see, this categorisation of the predicted coding sequence(s) will be important for increasing the quality of the second step of *PseudoChecker*'s pipeline, the alignment by MACSE (see Section 4.3.2).

As the output of the present component of the pipeline, Pseudo*Checker* automatically redirects to the second component of it (Alignment by MACSE) the predicted coding sequences (each per target species, being a reliable and/or less reliable, partial and/or full coding sequence), and, again, per target species, for display purposes at Pseudo*Checker*'s home page, as well as, in some part, due to its necessity at the final step of the pipeline, Pseudo*Checker* saves in memory the following additional information:

- Exon annotation related information, that includes: the sequence of each predicted exon, the length of it, the underlying reference species' coding exon alignment identity (%), the alignment coordinates at the corresponding genomic sequence, the position that each predicted exon occupies within the predicted coding sequence, adjacent splice sites (viable or not) and each exon considered as lost;
- General information concerning the predicted coding sequence, that includes: the sequence itself, the type of the sequence (partial, full, reliable or less reliable); when possible, for the viable cases, also computes and saves: the average, maximum and minimum intron size (bp); the largest and smallest predicted exon (and respective size in bp); average exon size (bp) and average alignment identity (%); splice site integrity underlying the composing predicted exons (number of actually valid splice sites/number of possible valid splice sites); coding sequence's percentage of guanine-cytosine content (%GC) and its length (in base pairs, bp).

A high level overview of the basic approach underlying this component of Pseudo*Checker*'s pipeline is summarised at Algorithm 1.

Algorithm 1: High level overview of the basic approach underlying the coding sequence prediction component of *PseudoChecker*'s pipeline

Input: Target species' genomic sequence(s), Reference species' coding exon(s) and coding sequence of the in-study gene

Output: Predicted coding sequence per target species, coding sequence prediction statistics and exon annotation related information

```

for species in target species do
    predicted coding sequence  $\leftarrow$  empty string;
    for exon in reference species' coding exon(s) do
        align exon at the target species' genomic sequence;
        if alignment identity < minimum alignment identity then
            set predicted exon as lost and save the assignment;
            set predicted coding sequence as a partial coding sequence;
        else
            predicted coding sequence += predicted exon's sequence;
            save alignment identity;
            save predicted exon's sequence;
            save adjacent (viable or not) splice site(s);
            save alignment coordinates within target species' genomic sequence;
            save the position of the predicted exon within the predicted coding sequence;
            force any upcoming alignment to be executed in a target species' downstream
            region relative to the target species' genomic sequence region concerning the
            present alignment;
            if alignment displays frameshift gaps and/or in-frame premature stop codons
            then
                set predicted coding sequence as less reliable;
            else
                set predicted coding sequence as reliable;
    if all reference species' coding exons (or single coding exon for single-exon genes)
    positively aligned at the target species' genomic sequence then
        set predicted coding sequence as a full coding sequence;
    save predicted coding sequence;
    compute and save predicted coding sequence associated general information;

```

4.3.1.2 Available similarity scoring schemes

Preceding the coding sequence prediction step, however, the similarity scoring scheme used for computing each optimal exon alignment must be selected. Since the dynamic programming algorithm that we employ will follow the determined scoring scheme for finding the alignment (or the multiple alignments) that maximise(s) its score, this will directly define how it will proceed to capture the relationship between both aligned sequences.

In this sense, since any erroneous coding exon prediction of the in-study gene might culminate in the false inference of gene inactivation mutations in the subsequent step of Pseudo*Checker*'s pipeline, the accurate selection of this parameter is crucial for the success of coding sequence annotation, and thus, gene loss prediction. Importantly, in the cases where, for a given alignment, multiple optimal solutions exist, we only recover one (the first retrieved by the Pairwise 2 module from Biopython). This is explained by the fact that: 1) we are not interested in identifying all possible changes between two strings, but rather to annotate orthologous exons to the reference species' ones, independently of ambiguous gap or character positions that may arise between alignments following the same similarity scoring scheme; 2) since we are aligning reasonably conserved DNA sequences, it is likely, and given our experience from the early tests, that only a single optimal solution exists; 3) as we will see below, we prefer to spend computation time in testing multiple different scoring schemes for aligning the same sequences, likely inducing greater levels of differences, than when compared to multiple optimal solutions underlying the same given similarity scoring scheme.

Here, Pseudo*Checker* provides two fixed (this is, applied to all executed exon alignments) similarity scoring schemes for punctuating matches, mismatches and gaps within an alignment. These were established during our early tests and can be defined as follows: a similarity scoring scheme optimised for sequences underlying very closely related species (for instance, to be used in cases where the reference species corresponds to human and the target species belongs to non-human primates), and a similarity scoring scheme optimised for sequences underlying slightly more divergent species. Both scoring schemes differ specifically in the mismatch penalty applied during optimal alignment computation, and, in this sense, one will strongly penalise mismatches, thus, forcing alignments to present higher levels of alignment identity, being suitable for aligning DNA sequences of closely related species, and the other will be more tolerant to their presence within an alignment, hence, resulting in alignments with larger amounts of mismatches, a more adequate approach for aligning slightly more divergent DNA sequences (Table 4.2).

However, Pseudo*Checker* does not make use of a simple match/mismatch/linear gap penalty. It has been proven that the first nucleotide (or amino acid) inserted/deleted (defined as gap opening) during alignment operations is, from a biological point of view, more significant than the subsequent ones (called gap extension), making the so-called affine gap model a viable solution underlying biomolecules alignment [23]. For this reason, in our approach, for both mentioned similarity scoring schemes, we also employ gap opening and gap extension costs, with these corresponding to -8 and -1, respectively.

As a result, in practice, this will tend to capture alignments that will favour contiguous gaps, instead of multiple adjacent single gaps, allowing us, therefore, to reduce the likeliness of within a produced alignment, encountering unrealistic multiple isolated gaps, producing, biologically speaking, more meaningful alignments.

Table 4.2: Fixed similarity scoring schemes provided by *PseudoChecker* underlying the coding sequence prediction component of its pipeline.

Similarity Scoring Scheme	Match	Mismatch	Gap opening	Gap extension
Optimised for closely related species	+1	-3	-8	-1
Optimised for slightly divergent species	+1	-1	-8	-1

Nonetheless, even though *PseudoChecker* displays these two optimised (for the above-mentioned circumstances) fixed similarity scoring schemes, it is currently known that, from a biological point of view, to achieve an interesting alignment between two sequences, different scoring schemes should be tested. Indeed, since the accurate selection of a similarity scoring scheme for the production of a meaningful alignment will not only depend on the evolutionary relationship of the species underlying the sequences to be aligned, but also in the degree of divergence underlying the tested gene [23, 182], from this, it is apprehended that both above-mentioned similarity scoring schemes are only useful to use when there is a clear idea about the evolutionary divergence between both reference and test species, the candidate gene’s conservation degree across them, or both.

Therefore, the selection of one of these scoring schemes might not be very convenient for the user if there is uncertainty concerning this information. Given this, *PseudoChecker* provides a third, however, slightly more time consuming, dynamic similarity scoring scheme. This is known as the best-fit similarity scoring scheme, and, in truth, this does not consist on a single (fixed for all executed alignments) similarity scoring scheme, but rather constitutes a set of multiple different similarity scoring schemes that will be dynamically tested per each exon alignment, followed by the selection of the unique alignment (amongst all) that better suits predetermined criteria. This constitutes a splice site- and reading-frame aware similarity scoring scheme, and, amongst the available three, is by far the most recommended similarity scoring scheme to be chosen prior to a *PseudoChecker* analysis.

In contrast to both fixed similarity scoring schemes, this dynamic scoring scheme intends to make *PseudoChecker*’s produced exon alignments more resistant to the evolutionary divergence of the in-study species and/or gene itself by testing different combinations of match/mismatch scores, with the ultimate goal of finding alignments that yield predicted intact exons, being the ones presenting conserved adjacent splice sites without underlying reading frame-disrupting indels.

Following the above explanation, once this scoring scheme is selected by the user, for each test species, and per reference species' coding exon, following each pair underlying a given match reward (M) and mismatch penalty (m) (see Table 4.3), an alignment of the same reference species' coding exon against the test species' genomic sequence is executed. The match score varies from 1 to 8, and the mismatch score varies from -1 to -5. In this sense, (as we will further see, in the worst-case scenario) $8 * 5$ match/mismatch different combinations will be explored, with each underlying a different alignment execution (Table 4.3). The applied gap opening and extension penalties are fixed and constitute the same applied in both previously-mentioned similarity scoring schemes (-8 and -1, respectively).

Table 4.3: Tested combinations of match/mismatch scores concerning the best-fit similarity scoring scheme underlying the first step of Pseudo*Checker*'s integrated pipeline. From left to right, top to bottom, and following this order, in the worst-case scenario, per reference species' coding exon and per target species' genomic sequence, an alignment between the two sequences, and following the corresponding match/mismatch scoring pair (M, m), is executed. Gap associated penalties are fixed, being -8 for gap opening and -1 for gap extension.

Match/Mismatch	-1	-2	-3	-4	-5
1	(1,-1)	(1,-2)	(1,-3)	(1,-4)	(1,-5)
2	(2,-1)	(2,-2)	(2,-3)	(2,-4)	(2,-5)
3	(3,-1)	(3,-2)	(3,-3)	(3,-4)	(3,-5)
4	(4,-1)	(4,-2)	(4,-3)	(4,-4)	(4,-5)
5	(5,-1)	(5,-2)	(5,-3)	(5,-4)	(5,-5)
6	(6,-1)	(6,-2)	(6,-3)	(6,-4)	(6,-5)
7	(7,-1)	(7,-2)	(7,-3)	(7,-4)	(7,-5)
8	(8,-1)	(8,-2)	(8,-3)	(8,-4)	(8,-5)

If, in fact, 40 alignments are executed (constituting the worst-case scenario), each is then automatically screened and according to the following criteria, the best alignment is retrieved:

1. First, Pseudo*Checker* screens all produced alignments for the conservation of adjacent splice sites within the resulting predicted exons (in-study intronless genes are not considered in this step). If conserved (GT/GC-AG splice site pair for alignments involving coding exons located between a 5' and a 3' coding exon, GT/GC for alignments involving a 5' coding exon, and AG for alignments involving a 3' coding exon), then, all alignments following this condition are filtered and the others excluded from the selection. If not a single alignment follows this condition, none is excluded from the selection and all alignments pass to the next filtering step. The present step is of utmost importance in the sense that, as we have seen in Chapter 2, since the dinucleotides underlying splice sites are highly conserved, a good indication of a correct exon boundary prediction corresponds to their conserved presence immediately adjacent to the predicted exonic sequence [172];

2. Next, *PseudoChecker* screens the alignments obtained from the previous step and filters the ones that do not present frame-disrupting indels. Alignments that, even when displaying frameshift gaps, if these correspond to compensatory ones (this is, frameshift pairs in which the first element of the pair disrupts the reading frame, but it is further downstream compensated by the second element of it, restoring the original reading frame) are also considered as valid, in the sense that these usually arise from artefacts resulting from inadequate combinations of mismatch/gap associated penalties, therefore not constituting real biological mutations. Within this component of the selection, intending to reduce the false positive inference of gene loss, we aim to certify (or not) that there is at least an alternative alignment that allows for the reading frame to be intact, or to be not disrupted at the 3' end of the alignment. If none alignment reunites one of these conditions, all alignments pass to the next and final step of the selection;
3. Finally, our approach screens all obtained alignments from the previous step and returns the one containing fewest gaps. We apply this last filter for avoiding alignments displaying an unrealistic number of gaps and gap length. At this stage, if at least two alignments comprise the same number of gaps, only one is (randomly) returned.

However, since testing all 40 scoring schemes per each exon alignment implies slowing down the execution time of a given *PseudoChecker* analysis for approximately 40x when comparing to the same analysis ran using one of the two fixed similarity scoring schemes (constituting an issue specially when dealing with large exons), for this motive, here we start by following a greedy approach turned to find alignments that are more prone to underline gene functionality:

1. For in-study genes composed by multiple coding exons, during each exon alignment following the best-fit similarity scoring scheme, the first obtained alignment that yields a predicted exonic sequence displaying adjacent consensus splice sites and presents no gaps, frame-preserving gaps (contiguous gaps with length multiple of three) or compensatory frameshift gaps, this is automatically chosen as the best and final alignment;
2. On the other hand, for analyses involving single-exon genes, during alignments' execution, the first alignment to display no gaps, frame-preserving gaps or compensatory frameshift gaps is automatically selected as the best alignment.

Importantly, notice that if not a single alignment, for both conditions (multi-exon and single-exon genes), displays these requirements (which means that the predicted exon will likely afterwards induce deleterious mutations when included into the corresponding predicted coding sequence), then all 40 alignments were necessarily executed, with the selection of the best alignment being made according to 3 above-mentioned filtering steps. This is, therefore, the key factor that determines whether, following the best-fit similarity scoring scheme, 40 alignments per each reference species' coding exon are executed or not.

4.3.1.3 Heuristic optimisations

As already stated, the variation of the Needleman-Wunsch algorithm that we are employing in this component of the pipeline uses dynamic programming techniques for finding the optimal semi-global alignment between a pair of DNA sequences. These methods ensure the optimal alignment by dividing the large alignment problem into a series of smaller problems, further using the solutions to the smaller problems to reconstruct the solution to the larger problem. Nonetheless, computationally speaking, deterministic alignments are not suitable for aligning very large sequences since, in these conditions, they demand large memory and time consumptions [75, 137].

Particularly, the Needleman-Wunsch algorithm's implementation by the Pairwise2 module from Biopython requires $\mathcal{O}(nm)$ time and space, where n and m are the lengths of the sequences to be aligned. This means that, with an increase of the input m and n , the time and space outcomes will quadratically increase. Plus, since we are applying the affine gap model in our alignments, the computation time can increase by a factor of about 3 when compared to the linear gap model [75]. As a result, when dealing with very long sequences, the problem becomes too expensive to solve in practise [7].

Since Pseudo*Checker* accepts as target species' genomic sequences entire genomic contigs, scaffolds and long sequencing reads, a heuristic method necessarily needs to be implemented. Here, we are using an anchoring strategy for determining the ideal, reduced fraction of the test species' genomic sequence to conduct semi-global alignments.

In detail, preceding each exon alignment against a given target species' genomic sequence: 1) we divide the sequence of the corresponding reference species' coding exon into a set of $L - k + 1$ k -mers, this is, all subsequences of length k , where L is the length of the complete exon, and k is the minimum value between the exon's length and 28; 2) for each produced set of k -mers, a match of each k -mer with tolerance up to 4 mismatches is searched at the target species' genomic sequence. If found, this process immediately stops and the produced alignment anchor will be defined as the target species' genomic portion flanked by the upstream and downstream limits of the match underlying the lowest number of tolerated mismatches and more closely located to its 5' end, extended in both sides by 5 times the length of the reference species' coding exon to be aligned. If not, this process is repeated for a k value decreased by a unit, until the alignment anchor is defined; 3) Finally, once the previous step is concluded, the deterministic semi-global alignment of the respective coding exon will be executed over the predetermined target species' alignment anchor.

Importantly, such as all heuristic alignment-based methods, this does not guarantee to find the optimal alignment between the two sequences [7], however from our experience this constitutes a nearly-optimal alignment strategy that is much more efficient to use for the cases where aligning very large sequences is a necessity, intending to offer the best trade-off between alignment quality, time and memory space requirements.

Yet, in every possible circumstance, this heuristic method is not applied. Since, again, the objective of this component of *PseudoChecker* lies in accurately annotate the orthologous sequence(s) to the reference species' gene's CDS, we only apply this method for the cases where the length of the reference species' coding exon to be aligned adding 1 unit, multiplied by the length of the given target species' genomic sequence adding 1 unit surpasses 200 000. This constitutes a reasonable value that reflects the maximum number of cells underlying the initialisation matrices required for performing deterministic alignments, for which, according to the characteristics of the used development machine, an alignment is easily computable.

A high level overview of the present heuristic method is depicted in the Algorithm 2.

Algorithm 2: High level overview of the heuristic method applied at the first component of *PseudoChecker*'s pipeline

Input: Reference species' coding exon to be aligned, Target species' genomic sequence

Output: Flanked portion of the target species' genomic sequence for conducting the deterministic pairwise alignment

$el \leftarrow$ length of the reference species' coding exon to be aligned at the target species' genomic sequence;

$gl \leftarrow$ length of the target species' genomic sequence;

if $el + 1 * gl + 1 > 200000$ **then**

for $k \leftarrow \min(el, 28)$ **to** 1 **do**

 divide reference species' coding exon into a set of $el - k + 1$ -mers;

 search match of each produced k -mer at the target species' genomic sequence with tolerance up to 4 mismatches;

if *at least one exists* **then**

 break;

 retrieve the one underlying the fewest number of mismatches and more closely located to the 5' end of the target species' genomic sequence;

 return the coordinates of the target species' genomic sequence flanked by the upstream and downstream limits of the detected match, extended in both sides by 5 times the length of the reference species' coding exon;

 send the instruction to execute the deterministic alignment over the predetermined alignment anchor;

else

 reject the heuristic method;

 send the instruction to execute the deterministic alignment over the entire search space (the entire target species' genomic sequence);

4.3.2 Alignment by MACSE

4.3.2.1 Introduction to MACSE

A prerequisite for evolutionary studies lies on the multiple comparison of orthologous genes' coding sequences. As previously outlined, genes' coding sequences can be considered either at the nucleotide or amino acid level. Numerous tools nowadays exist to align multiple DNA sequences such as CLUSTAL [70], DIALIGN [134], or MUSCLE [45]. Yet, when dealing with protein-coding sequences, these display a major handicap: they do not take into account the corresponding amino acid translations. As a consequence: 1) as nucleotide sequences are less conserved, similarities that could be detected at the amino acid level might be obscured at the nucleotide level, hence, difficulting an accurate alignment; 2) since their underlying maximisation criterion do not penalise insertions and deletions that induce frameshift gaps, for instance, a protein-coding sequence displaying an insertion of two nucleotides (or base pairs) followed by a downstream insertion of 7 nucleotides will have the same gap-related penalties as a more biologically meaningful scenario of an insertion of three nucleotides followed by an adjacent insertion of 6 nucleotides [159].

For overcoming these problems, a common strategy consists of using a three-step approach: first, all coding nucleotide sequences are translated into their amino acid composition, then, these amino acid sequences are aligned, and finally, the obtained protein alignment is used for deriving the nucleotide one [159]. Some algorithms and respective implementations such as revTrans [199], TranslatorX [2] or transAlign [15] were specifically developed to automate this straightforward alignment strategy. Nonetheless, all of these display the inability to handle unexpected frameshift gaps. As a consequence, the underlying amino acid translation that follows such events is no longer the correct one, either resulting in the prevention of alignment execution or resulting in a translated amino acid sequence that will look like a highly divergent, orphan sequence, yielding a partly aberrant DNA alignment [159].

MACSE (Multiple Alignment of Coding SEquences), with its first version being released in 2011 [159], was the first and a still widely used solution able to multiple align coding sequences based on their amino acid translations while accounting for frameshift mutations and premature stop codons. This MACSE's unique capability of aligning gene datasets containing both non-functional (pseudogenised) and functional sequences while preserving the underlying codon structure, makes its respective integration into Pseudo*Checker* the perfect solution for, not only, accurately identifying conserved and non-conserved frameshift mutations and/or premature stop codons across the in-analysis lineages' corresponding sequences, but also, by exporting the alignment, to make it a viable option for directly performing downstream phylogenetic and selection analyses (the estimation of d_N/d_S ratios) with methods based on codon models of sequence evolution [159].

When used as a standalone tool and a given multiple (or pairwise) alignment is ran, as output, MACSE provides two equivalent alignments in FASTA format: one at the nucleotide level and one at the amino acid level. At the amino acid level, MACSE represents a stop codon by the often-used symbol '*' and, in its turn, a codon containing underlying a frameshift mutation is represented by an extra symbol, the '!'. At the nucleotide level, codons underlying frameshifts make use of the symbol '!' to maintain the structure of the reading frame and no special representation is used for depicting stop codons [159] (Figure 4.5).

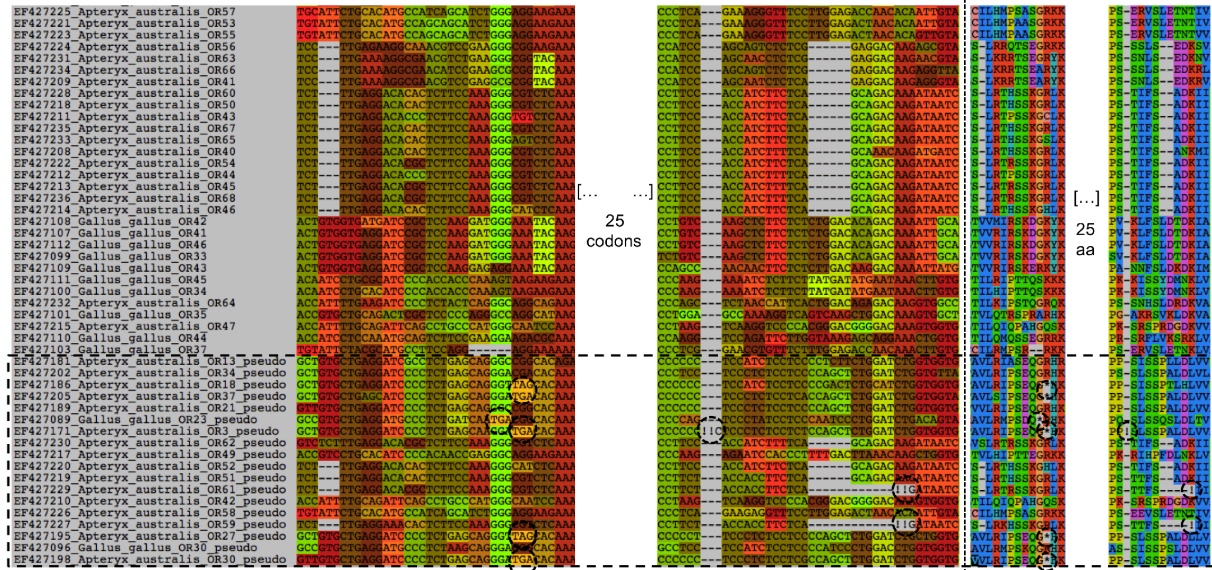


Figure 4.5: Illustrative example of a multiple sequence alignment executed by MACSE. Adapted from [159]. 93 functional and 18 pseudogenic sequences from *Apteryx australis* (brown kiwi) and *Gallus gallus* (domestic chicken) olfactory repertoires were multiply aligned by MACSE and visualised with SEAVIEW [52]. The equivalent alignment region is displayed at the nucleotide (on the left) and amino acid (on the right) levels. The 18 pseudogenic sequences are represented within a dashed box. Stop codons (represented by an * at the amino acid level) and frameshifts (represented by exclamation marks) inferred by MACSE are circled. Importantly, setting the presence of '!' within codons underlying frameshift mutations allows MACSE to maintain the reading frame, thus permitting the accurate and biologically meaningful comparison of orthologous pseudogenic and viable coding sequences.

Algorithmically speaking, MACSE solution is an improved version of the Needleman-Wunsch algorithm [140] that, while displaying the original asymptotic complexity ($\mathcal{O}(nm)$ time and space), for computing each optimal pairwise alignment, adds alignment costs associated to frameshifts mutations and stop codons. For the cases where more than 2 sequences comprise the inputted sequence dataset, MACSE extends the pairwise to a multiple sequence alignment by using a progressive alignment strategy to obtain an initial draft of the multiple sequence alignment that is subsequently refined.

In detail, first, all pairwise sequence similarities within the inputted dataset are estimated based on the frequencies of their nucleotide k -mers, this is, subsequences of k nucleotides. Next, those similarities are used to produce a dichotomic rooted guide tree that makes use of the UPGMA distance method, here not used for building a phylogeny of the sequences, but rather to build a guide tree that will group similar sequences which must be aligned first. Here, the leaves of the produced tree are associated with the sequences to be aligned, whereas its internal nodes are associated with the multiple sequence alignment of the sequences comprising the corresponding clade. Therefore, the internal nodes are processed bottom-up, and the alignment of a node of the tree is obtained by aligning the corresponding two descendants.

Finally, the resulting multiple sequence alignment at the root node is then used as the initial draft for the final alignment. MACSE uses a 2-cut refinement strategy that partitions the draft solution into two sub-alignments that are subsequently re-aligned. Nevertheless, the resulting multiple sequence alignment only replaces the draft one, if the sum-of-pair score, this is, the sum of the scores of the pairwise alignments that the multiple sequence alignment induced, improves throughout the refinement [159].

Given its characteristics, for computing an alignment, MACSE requires the selection of a more complex similarity scoring scheme, when compared to other pairwise solutions (particularly, when compared to the one underlying the previous component of Pseudo*Checker*'s pipeline) or multiple alignment solutions. Indeed, the final score of a given MACSE optimal alignment will depend on the bonus/cost of each elementary event that it encompassed [159]. The MACSE underlying complex scoring scheme can be summarised as follows:

1. First, MACSE will define the scoring outcome of nucleotide alignments based on the amino acid translation of each codon, therefore, the bonus/cost of one amino acid facing another is defined by the amino acid substitution matrix that it employs;
2. In second, the cost of a stretch of n consecutive frame-preserving gaps (gaps that surge in a number multiple of three) is defined by a gap stretch opening cost, and a gap extension cost proportional to its length n ;
3. Finally, the cost of a frameshift and a stop codon will depend, first, if they are located within the interior or at the extremities of the sequence, and second, in the type of the sequence they will appear (less reliable or reliable sequences).

In a greater level of detail, MACSE displays a set of manipulable parameters, each underlying the bonus/cost of a given elementary event. Table 4.4 provides an in-depth description of each of these.

Table 4.4: In-depth description of the alignment elementary bonus/costs associated with the scoring function employed by MACSE. The elementary costs for which the corresponding value can differ between reliable and less reliable sequences are marked with *. Since the occurrence of a stop codon in the last codon of a given coding sequence should not be penalised at all, the terminal stop codon elementary cost represents the unique parameter that cannot be manipulable by the user (equal to 0, tagged with a **).

MACSE elementary bonus/cost	Description
Scoring matrix	Amino acid substitution matrix used for punctuating matches/mismatches between amino acids
Gap opening	Cost of internal frame-preserving gap opening (those between the first and last codon)
Terminal gap opening	Cost of terminal frame-preserving gap opening (those before the first nucleotide and after the last one)
Gap extension	Cost of internal frame-preserving gap extension (between the first and last codon)
Terminal gap extension	Cost of terminal frame-preserving gap extension (those before the first nucleotide and after the last one)
Frameshift*	Cost of an internal frameshift (those between the first and last codon)
Terminal frameshift*	Cost of a terminal frameshift (those in the first and last codon)
Stop codon*	Cost of a stop codon arising between the first and the last codon
Terminal stop codon**	Cost of a stop codon arising at the last codon

As a consequence of such sophisticated alignment scoring function, the trickiest part underlying a MACSE alignment, is, and according to the authors, in fact, to choose the final scoring scheme to compute the optimal alignment that satisfies the user's needs. Particularly, one of the problems underlying common multiple sequence alignment (by common we are referring to multiple alignment of functional coding sequences) is the selection of the relative costs of gap opening and gap extensions with respect to amino acid substitution. Slightly changing these costs will strongly impact the final result, and no efficient strategies until now have been found to define an ideal cost for these parameters [200]. Furthermore, given that MACSE additionally requires the selection of frameshift and stop codons associated costs, the final complexity level is increased [159].

Fortunately, however, the authors provide a default scoring scheme, that, from their experience, has proven to be suitable for the vast majority of cases [159]. MACSE allows for applying different alignment elementary costs in two different types of sequences: reliable or less reliable sequences. Reliable sequences are most likely functional, thus, should not present any premature stop codon or frameshift mutation throughout the aligned sequence, therefore, by default, these present higher costs for those events to occur. In contrast, less reliable sequences should comprise the ones more prone to display frameshifts and/or premature stop codons across the alignment, and, for this reason, these display associated lower elementary costs for the same events (Table 4.5).

These differences in alignment elementary costs proposed between these two types of sequences will only affect parameters related with frameshift and stop codon elementary costs, whereas the rewards and costs associated to amino acid alignment, costs related with gap opening and gap extension are equal for both types of sequences. MACSE uses BLOSUM62 as the default substitution matrix. The remaining alignment costs underlying each elementary event are summarised at the Table 4.5.

Table 4.5: Default values for MACSE alignment elementary costs for both types of sequences (reliable and less reliable), exclusive to reliable and exclusive to less reliable sequences.

Common costs	Gap opening	Terminal gap opening	Gap extension	Terminal gap extension
—	7	6.3	1	0.9
Reliable sequences	Frameshift	Terminal frameshift	Stop codon	Terminal stop codon
—	30	10	50	0
Less reliable sequences	Frameshift	Terminal frameshift	Stop codon	Terminal stop codon
—	17	10	10	0

Nonetheless, in some cases, using the MACSE default costs is not sufficient to achieve the pretended alignment. Therefore, for some situations, it is useful for the user to adjust determined elementary costs to turn an unexpected alignment into a more biologically meaningful one. For instance, if a given alignment displays too many small (frame-preserving) gaps, the user might want to try increasing the cost underlying the gap opening event (this would favour the alignment with fewer but longer gaps). On the other hand, if the aligned sequences display unexpected frameshifts ripping apart a more meaningful premature stop codon, the user might try to either decrease the stop codon associated penalty or increase the frameshift associated costs.

4.3.2.2 Exploiting MACSE in Pseudo*Checker*

The latest release of MACSE program, version 2.0, is distributed as an open source, Java executable file [160]. For running a MACSE analysis, this only requires a set of arguments and sequence input, being these, in detail, the scoring values attributed to all alignment elementary events (followed during the optimal alignment computation, that, in case of omission, the default values are applied), a FASTA file containing the reliable sequences and/or a file comprising less reliable sequences in the same format. In this component of the pipeline, Pseudo*Checker* exploits MACSE by running it for performing a deterministic alignment between the predicted (viable or not) coding sequence(s) at the previous step of the pipeline, the reference species' candidate gene's CDS, and the optional predetermined inputted functional coding sequence(s) (if existing).

Prior to a given Pseudo*Checker* analysis (and similarly to MACSE when used as a standalone tool), Pseudo*Checker* allows to manipulate each alignment elementary cost at its home page (by default, it displays the default MACSE elementary alignment costs, strongly recommended by us, since they have proven to be suitable for the majority of our early tests). Here, however Pseudo*Checker* fixes the genetic code (also manipulable in MACSE) to the standard genetic code, the unique genetic code that is suitable for analysing nuclear mammalian protein-coding genes' coding sequences, also fixing the amino acid substitution matrix to the default one, BLOSUM62, for which the (again, recommended by us to use) default elementary alignment costs are optimised [159]. When the coding sequence prediction component of the pipeline finishes executing, Pseudo*Checker* automatically runs MACSE providing it as input the defined MACSE similarity scoring scheme at Pseudo*Checker*'s home page, the reliable and/or less reliable predicted coding sequences arising from the first step of the pipeline, the predetermined functional coding sequence(s) (when existing, automatically assigned as (a) reliable sequence(s)), and the

reference species' gene's coding sequence (also automatically assigned as a reliable sequence).

Importantly, this prior knowledge and compartmentalisation of the predicted coding sequences in either reliable and less reliable sequences (see Section 4.3.1.1) have demonstrated us to aid and improve, biologically speaking, the alignments produced by MACSE. However, this only makes sense when (and as corroborated by MACSE default costs), for reliable sequences, the gap and stop codon associated penalties are superior to the ones applied for the less reliable sequences concerning the same events. For this reason, even though *PseudoChecker* does not impede the opposite case from occurring, we strongly discourage applying lower gap and stop codon associated penalties to reliable sequences when compared to less reliable sequences.

One should notice, however, that less reliable sequences might not end up constituting pseudogenic sequences. First, because the frameshift gaps that arose during the corresponding coding exon annotation at the previous component of the pipeline might collapse, now when realigned with MACSE, that, as we have seen, takes into account the amino acid translation while aligning nucleotide sequences, and second, because, in the case of the predicted coding sequence, indeed, when aligned with MACSE, displaying frameshifts and/or in-frame premature stop codons, if these are closely located to the 3' end of the sequence, it might turn out not to be problematic (see Section 4.4). In contrast, from our experience, reliable sequences are more likely to, when introduced into MACSE alignment, in fact, to constitute viable coding sequences (not pseudogenised).

Finally, *PseudoChecker* will internally redirect the output from MACSE, this is, as previously mentioned, the resulting alignment in two different FASTA files, one at the nucleotide level, and the other at the amino acid level, directly to the third, and final part of the pipeline: reporting of the mutational evidence and PseudoIndex computation (Section 4.3.3). The summarised approach in respect to this component of *PseudoChecker*'s pipeline is depicted in Algorithm 3.

Algorithm 3: High level overview of the basic approach underlying the Alignment by MACSE component of *PseudoChecker*'s pipeline

Input: Reference species' candidate gene's coding sequence, Reliable and/or less reliable predicted coding sequence(s) obtained from the first component of *PseudoChecker*'s pipeline, predetermined coding sequence(s) (if existing), MACSE defined similarity scoring scheme at *PseudoChecker*'s home page

Output: MACSE produced alignment

if *at least one predetermined coding sequence exists* **then**

foreach *predetermined coding sequence in the entire available set* **do**
assign it as a MACSE reliable sequence;

assign reference species' CDS as a MACSE reliable sequence;

`Execute_MACSE_Alignment` (*Reliable sequence(s)*, *Less reliable sequence(s)*, *MACSE selected alignment costs*);

4.3.3 Report of the mutational evidence and PseudoIndex computation

In this component of its integrated pipeline, PseudoChecker directly analyses the output from the two previous components of it for reporting detected mutational evidence. Taking into account the information produced during the coding sequence prediction step, PseudoChecker reports, for each target species, each lost exon and splice site disruptive mutations. As a disrupted 5' (donor) splice site, we consider any deviation from the consensus GT/GC dinucleotide, and in its turn, any alteration of the consensus AG dinucleotide for the acceptor splice site (3' splice site) is assumed as its disruption.

Next, the output from the MACSE alignment is scrutinised and, again, for each target species, within the corresponding aligned predicted coding sequence, detected deleterious mutations that include in-frame premature stop codons and frameshift mutations relative to the reference species' aligned CDS are recognised. Particularly, per target species underlying a predicted full coding sequence, besides each detected mutation being annotated, and its position in the alignment being reported, the corresponding predicted coding exon for which where it occurs is also disclosed. Species comprising partial coding sequences are excluded from this feature (see Limitations, Section 4.6). Finally, taking into account the presence and level of mutational evidence, a PseudoIndex is computed and attributed at the same species.

At the main results page, PseudoChecker visually reports the produced MACSE alignment by combining both two MACSE outputted alignments (one at the nucleotide level, and the other at the amino acid level) in a single, and more intuitive alignment (Figure 4.6).



Figure 4.6: Illustrative example of a MACSE alignment displayed by PseudoChecker comprising 5 simulated sequences, each underlying a simulated species.

For a convenient visualisation, the reference species' CDS aligned sequence is always shown at the top, and the alignment is colour graded according to the codon structure represented on it. Therefore, each set of three nucleotides adjacent to one another is represented with different background colours. At the nucleotide level, (frame-preserving) gaps are represented by a codon '-' '-' underlying a white background. At the amino acid level, in contrast, no special representation is applied. Regarding frameshift mutations, at the nucleotide level, these are represented by a

partial codon underlying one or two exclamation marks, each highlighted in orange, and at the amino acid level, no representation is used. For in-frame stop codons, relatively to the nucleotide level, these are represented in red font and white background colour and, at the amino acid level, these are represented by a *, also in red font. Finally, if a given amino acid is different relative to the reference species' one at the same alignment position, the first is represented with a grey background colour (Figure 4.6).

As we have mentioned when introducing MACSE, exclamation marks arise within (partial) codons that underlie frameshift mutations for preserving the structure of the reading frame. These partial codons arise in different forms, being '!!N', '!N!' or 'N!!', with 'N' corresponding to any of the 4 DNA nucleotide bases. Importantly, when considered isolated, each will pinpoint different interpretations. For instance, a partial codon represented by a '!!N', '!N!' or '!!N' can either represent the deletion of two nucleotides or a single nucleotide insertion. However, since frameshift mutations are being inferred relative to the reference species' candidate gene's CDS, the mutational interpretation underlying each partial codon must be made according to the reference species' codon observed at the same alignment site.

First, if at the reference species' candidate gene's coding sequence, a given codon is represented by 'NNN', where N represents any of the 4 DNA nucleotide bases, and at a given target species corresponding sequence the observed codon at the same alignment site is represented by a partial codon containing at least one '!', this partial codon should be understood as underlying frameshift deletions, wherein each exclamation mark represents a single nucleotide deletion. In contrast, if a given codon at the reference species' sequence corresponds to a set of three gaps '- - -', and the corresponding codon of a given species aligning at the same site is, represented by a '!!N', '!NN', 'N!!', 'NN!' or '!N!', this will represent an insertion of the 'N' nucleotide base, for which the N corresponds to any of the 4 DNA nucleotide bases.

Indeed, this rationale underlying partial codons interpretation must be taken into consideration by the user when observing the alignment, being directly incorporated into *PseudoChecker*'s logical core, as it will be further assumed during *PseudoIndex* computation (see Section 4.4).

Importantly, not every MACSE produced alignment is viable for running a *PseudoChecker* analysis. Even though, from our experience, such event is unlikely to occur, since the produced alignment depends on the selected MACSE similarity scoring scheme at *PseudoChecker*'s home page, inadequate choices regarding the same scoring scheme for the in-study sequence dataset might lead to the appearance of frameshift mutations and/or premature stop codons at the reference species' CDS and/or at inputted predetermined coding sequences (if existing). As these sequences are supposed to be functional, therefore, not presenting any frameshift or premature stop codon arising within their aligned sequences, *PseudoChecker* automatically rejects any analysis concealing these erroneous situations, consequently interrupting the analysis from that point (Figure 4.7).



Figure 4.7: Illustrative example of a MACSE alignment not accepted by PseudoChecker. Partial codons arise within the simulated reference species' aligned sequence. PseudoChecker automatically rejects the analysis from this point.

A high level overview of the basic approach underlying this third and final component of PseudoChecker's pipeline is depicted in Algorithm 4.

Algorithm 4: High level overview of the basic approach underlying the third and final component of PseudoChecker's pipeline

Input: Information saved in memory regarding the first component of PseudoChecker's pipeline, MACSE alignment produced during the previous component of PseudoChecker's pipeline

Output: Report of mutational evidence and PseudoIndex assignment per target species

foreach *aligned sequence in MACSE alignment* **do**

if *aligned sequence corresponds to the reference species' CDS or to a predetermined coding sequence* **then**

if *premature stop codon or frameshift mutation arises in it* **then**

 interrupt the analysis;

else

if *aligned sequence corresponds a full predicted coding sequence* **then**

if *premature stop codon or frameshift mutation arises in it* **then**

 report for the corresponding species the position of the mutation within the MACSE alignment and the corresponding exon for which where it occurs;

foreach *species in target species* **do**

if *displays lost exons and/or disrupted splice sites* **then**

 report them;

 compute PseudoIndex;

4.4 PseudoIndex

Accurately measuring the level of pseudogenisation of a given poses several challenges. For instance, issues related to alignments, evolutionary changes in the exon-intron structures of conserved genes, including splice site shifts over evolution, lineage-specific exons and precise intron deletions, all mimic inactivating mutations in genes that, in fact, are functional. Additionally, even real mutations might not indicate gene loss, for example, when a given frameshift indel arises but is further downstream compensated by another one that restores the original reading frame, or when such frameshifts and/or premature stop codons arise close to the sequence region encoding the C-terminus of the resulting protein, which is under less evolutionary constraint [172].

Taking into account all these factors while manually screening a given predicted gene might turn out to constitute a confusing and consuming task. Therefore, for overcoming this challenge, we have built into *PseudoChecker* the PseudoIndex, a user assistant metric that intends to measure at a glance the erosion state of a given gene at a given species by inspecting the presence and levels of its mutational evidence. Importantly, since the user will have the ultimate interpretation of the level of mutational evidence, therefore, the classification of the in-study gene in a given species as coding or pseudogenised, this should not be seen as the ultimate verdict underlying its coding status, but rather as a metric that takes into account multivariate factors aiding for the accurate classification.

Explicitly, for each target species, PseudoIndex takes into account three different components: the absent exonic content component, that takes into account the percentage of exonic content present at the reference species' gene that does not positively align at its genomic sequence; the shifted codons component, that takes into account the percentage of codons that are read out of the original reading frame; and the truncated sequence component, that measures the percentage of the target species' gene's sequence that is not translated into protein, preceded by the presence of a premature stop codon.

Here, however, since these shift during evolution, and even though we report their disruption at *PseudoChecker*'s main results page, we do not take into account splice site abolishing mutations. Such splice sites shifts likely happen when a new (and alternative) splice site arises within an exon and the previous one is mutated. This implies that, in these cases, although a given reference species' exon positively aligns at a given target species' genomic sequence, and the adjacent dinucleotides to the resulting predicted coding exon are indeed orthologous to the reference species' ones, these do not function as splice sites anymore [171].

From this, the main issue associated with this factor lies in distinguishing real splice site mutations from splice site shifts over evolution [171]. Given this motive, and even though we make all efforts for finding exon alignments yielding predicted exons displaying adjacent conserved splice sites (namely within the best-fit scoring scheme from the first step of the pipeline), intending to maintain misclassification rate at low values, we decided to not penalise this event

at all. Additionally, since, and as we have already seen in Chapter 2, that real lost genes usually present other mutational events besides splice site disruptive mutations, we consider this to be a reasonable and non-problematic decision. As a consequence of this outright non-penalisation of (hypothetical) splice site abolishing mutations, the real loss of splice site pairs (GT/GC-AG) that results from precise intron deletions (common in rodents [36]) is also automatically not penalised.

The PseudoIndex attributed value for each in-study target species varies on a discrete scale from 0 to 5, with a PseudoIndex of 0 suggesting the full functionality of the candidate analysed gene and a PseudoIndex of 5 indicating its full erosion. Concretely, for a given target species, a PseudoIndex of 0 indicates that it presents an intact, or almost intact version of the in-study gene, and a PseudoIndex of 1 and 2 indicates that, although the predicted gene has shown some mutational evidence, this likely does not affect the final protein product. A PseudoIndex of 3, importantly, indicates a doubtful case, for which the coding status of the corresponding gene should be manually inspected. Finally, a PseudoIndex equal to 4 or 5 suggests the loss of the in-study gene.

Each of the three above-mentioned components of PseudoIndex will yield a sub-PseudoIndex, this is, a representation of the mutational evidence underlying each analysed component. Then, the final PseudoIndex value attributed to each target species within an analysis will correspond to the highest value amongst those computed sub-PseudoIndex.

In fact, setting the final PseudoIndex value as the highest value amongst the three computed sub-PseudoIndex is the factor that makes PseudoIndex taking into account all the mentioned factors. For instance, if a given predicted gene does not display any premature stop codon or frameshift mutations, yet presents over 90% of missing exonic content, the corresponding species' attributed PseudoIndex will always be 5, therefore, suggesting its loss in the same species. On the other hand, for a given species, if the corresponding predicted gene's coding sequence is composed by the totality of the reference species' orthologous coding exons, does not display any premature stop codon, but displays single or multiple frameshifts mutations that lead to a massive percentage value of codons read out of frame, its assigned PseudoIndex value will equally going to be 5.

Indeed, due to this approach, more complex cases are capable of accurately being addressed, for instance: 1) wherein a given target species, the corresponding predicted coding sequence does not display any frameshift mutations or premature stop codons, but a given coding exon is lost, even though for which, isolated, its weight over the total in-study gene is not sufficient to be considered as strong gene loss evidence, but since its absence leads to the disruption of the reading frame, shifting the frame over a large content of the predicted sequence, a high PseudoIndex value is attributed to the same species; 2) or, for the opposite case, when there are no premature stop codons or frameshift mutations arising in a given predicted coding sequence, and given coding exon is missing, but since its absence does not lead to the disruption of the reading frame, and since it displays a minimum weight over the entire in-study gene, this combination of mutational events likely yields a low assigned PseudoIndex value to the affected species.

4.4.1 Absent exonic content component

In this component of PseudoIndex, *PseudoChecker* measures the harmful impact that the absence of single or multiple coding exons displays on the annotated gene at a given target species. For accomplishing this, *PseudoChecker* starts by computing the weight that each coding exon has at the reference species' gene by dividing its nucleotide length over the entire reference species' coding sequence length. Then, the percentage of absent exonic content is the result of the sum of this computed ratio for each absent exon, multiplied by 100.

Different obtained values will yield different sub-PseudoIndex values. Table 4.6 establishes the correspondence between each percentage value of absent exonic content and the returned sub-PseudoIndex for this component of PseudoIndex.

Table 4.6: Relation of the percentage value of absent exonic content with the attributed sub-PseudoIndex for the absent exonic content component of PseudoIndex.

Absent exonic content (%)	sub-PseudoIndex
≤ 10	0
> 10 and ≤ 15	1
> 15 and ≤ 20	2
> 20 and ≤ 25	3
> 25 and ≤ 30	4
> 30	5

4.4.2 Shifted codons component

Within this component of PseudoIndex, *PseudoChecker* measures the impact that frameshift mutations have on the in-study predicted gene's coding sequence at a given target species. Here, our approach takes into account isolated frameshifts, in other words, a single frameshift that occurs within a given sequence, multiple frameshifts that do not compensate each other, compensatory frameshifts, that, as we have previously mentioned, restore the original status of the reading frame by compensating each other, and reading frame disruptive effects caused by the absence of a single or multiple exons.

To accomplish this, this PseudoIndex's component initially calculates the percentage of shifted codons that a given sequence displays. *PseudoChecker* starts by counting the codons that are read out of the original reading frame (gapped codons, read as '- -' are not considered) from the 5' end to the 3' end of the sequence, then, divides the obtained number by the number of total codons within the sequence, and further multiplies it by 100 (gapped codons are not, once again, considered). However, this value is only calculated within the sequence portion that initiates at the first observed in-frame start codon and ends in the last available codon. Different values obtained for this ratio will, similarly to the previous component of PseudoIndex, result in different computed preliminary sub-PseudoIndex values (Table 4.7).

Table 4.7: Relation underlying the percentage value of shifted codons with the attributed preliminary sub-PseudoIndex for the shifted codons component of PseudoIndex.

Shifted codons (%)	Preliminary sub-PseudoIndex
≤ 10	0
> 10 and ≤ 15	1
> 15 and ≤ 20	2
> 20 and ≤ 25	3
> 25 and ≤ 30	4
> 30	5

This rationale lies in the possibility of frameshift mutations arising upstream of the first start codon not corresponding to real mutations. In common cases, the first codon of a predicted coding sequence will correspond to a start codon, however, in the eventuality of such event do not occur, since the first existent downstream start codon should not be dismissed as the hypothetical real start codon at the sequence of the species in question, for these cases, frameshifts that occur upstream of it should be less penalised than frameshifts arising downstream of the first observed start codon. Yet, if at least one frameshift mutation arises upstream of the first observable start codon within a given species' predicted coding sequence, its minimum value underlying the computed sub-PseudoIndex for this component will always be equal to 3. Then, the final sub-PseudoIndex value obtained for this component of PseudoIndex will constitute the highest value between 3 and the preliminary sub-PseudoIndex that resulted from the previously computed percentage of shifted codons (Table 4.7).

In contrast, in the absence of frameshift mutations arising upstream of the first in-frame start codon, the computed sub-PseudoIndex value will solely be dependent on the value computed for the preliminary sub-PseudoIndex (Table 4.7). Finally, for the cases where none start codons are detected within a sequence of a given species, therefore, not being possible to distinguish between likely real frameshifts, from less likely real frameshifts, its attributed sub-PseudoIndex for this component of PseudoIndex corresponds to 3. Figure 4.8 displays several illustrative examples of frameshift mutational events that will produce different computed sub-PseudoIndex values regarding the present component of PseudoIndex.



Figure 4.8: Illustrative example of several different frameshift mutational events observable within a MACSE alignment regarding a simulated PseudoChecker analysis.

According to the mutational evidence unveiled in Figure 4.8, Species 1 displays, for this component of PseudoIndex, a sub-PseudoIndex equal to 3, as the unique 2-nucleotide frameshift insertion that the corresponding simulated sequence displays, arises upstream to the first (maybe alternative and real) observable start codon. Opposing, Species 3 displays a sub-PseudoIndex equal to 5 since the unique frameshift deletion that the corresponding predicted coding sequence displays leads to a large percentage value of shifted codons. In its turn, Species 4, even though displaying two compensatory frameshifts underlying its simulated sequence (a single nucleotide insertion followed by a downstream deletion of 1 nucleotide), also displays a sub-PseudoIndex equal to 5 since the percentage value of shifted codons is still greater than 30%. Finally, Species 2 displays a sub-PseudoIndex of 0 since its simulated sequence does not display any frameshift mutational event.

Since all these simulated species and corresponding simulated sequences present no other mutational evidence besides frameshift mutations (excepting for Species 2 that does not present any mutational evidence at all), the attributed (general) PseudoIndex value will always correspond to the same value returned when computing the sub-PseudoIndex one for underlying the shifted codons component of PseudoIndex.

4.4.3 Truncated sequence component

Finally, this last component of PseudoIndex takes into account the percentage of truncated sequence that each species corresponding sequence displays. This is defined as the number of non-gapped codons that are not translated into protein (following either an in-frame premature stop codon, or an out-of-frame premature stop codon biologically translated as a real stop codon, as a consequence of an upstream disruption of the reading frame) further divided by the number of codons within the sequence, multiplied by 100.

Similarly to the rationale underlying the previously explored component of PseudoIndex, however, this ratio is only calculated within the sequence portion that initiates at the first observable in-frame start codon, ending in the last available codon. Different values obtained for this metric will also yield different preliminary sub-PseudoIndex values (Table 4.8).

Table 4.8: Relation underlying the percentage value of truncated sequence with the attributed preliminary sub-PseudoIndex for the truncated sequence component of PseudoIndex.

Truncated sequence (%)	Preliminary sub-PseudoIndex
≤ 10	0
> 10 and ≤ 15	1
> 15 and ≤ 20	2
> 20 and ≤ 25	3
> 25 and ≤ 30	4
> 30	5

This is imposed since we should not exclude the hypothesis that the first downstream observable start codon relative to the first codon within a given species' predicted coding sequence, constitutes, in fact, the real start codon in it. Therefore, if this, in fact, biologically constitutes the real start codon in the sequence in question, any premature stop codon that arises upstream of it should not lead to deleterious effects at the protein level.

Nevertheless, if an in-frame or out-of-frame premature start codon (that is translated as an effective stop codon due to the upstream disruption of the original reading frame) arises prior to the first observed start codon of the in-analysis sequence, the minimum corresponding species assigned sub-PseudoIndex value for this component of PseudoIndex will be equal to 3. As a consequence, the final attributed sub-PseudoIndex to the affected species will be defined by the maximum value between 3 and the one defined by the preliminary sub-PseudoIndex, that relates the interval of where the percentage of truncated sequence lies with the assigned value (Table 4.8).

In contrast, the sub-PseudoIndex value attributed to a given species will only depend on the percentage of truncated sequence, if its sequence displays no premature stop codons arising upstream of the first observed start codon. Finally, for the cases where no start codons are found at a given species' sequence, therefore, being difficult to distinguishing between likely real premature stop codons from less likely real ones, for the same species, the assigned sub-PseudoIndex concerning this PseudoIndex's component will be equal to 3.

Figure 4.9 displays several illustrative examples of different mutational events concerning the appearance of premature stop codons, respectively yielding different sub-PseudoIndex outcomes for the present component of PseudoIndex.

	2	5	8	11	14	17	20	23	26	29	32	35	38	41
Reference Species	ATGCCCGGACTAAAATTGCCATGGAAATTTAACCCAATT	TGA												
	M	P	G	L	K	L	P	W	K	F	N	P	I	*
Species 1	ATGTCCGGACTAAGATTGCCATGGAAATTTAACCCAATT	TGA												
	M	S	G	L	R	L	P	W	K	F	N	P	I	*
Species 2	ATACCC	TGA	ATGAAATTGCCATGGAAATTTAACCCAATT	TGA										
	I	P	*	M	K	L	P	W	K	F	N	P	I	*
Species 3	ATGCCCGGACTAAAAT	TGA	CCATGGAAATTTAACCCAATT	TGA										
	M	P	G	L	K	*	P	W	K	F	N	P	I	*

Figure 4.9: Illustrative example of several mutational events concerning the presence of premature stop codons observable within a MACSE alignment underlying a simulated *PseudoChecker* analysis.

Taking into account the mutational evidence depicted in Figure 4.9, Species 1 displays a sub-PseudoIndex equal to 0 since there aren't present any premature stop codons at the corresponding simulated sequence. Species 2, in its turn, displays a sub-PseudoIndex of 3 since the immediately downstream adjacent start codon to the observable stop codon likely constitutes the real start codon at its sequence. Finally, Species 3 displays a sub-PseudoIndex equal to 5 since the

percentage of truncated sequence surpasses 30%.

Similarly to the previously analysed component of *PseudoIndex*, since no other mutational evidence besides the presence of premature stop codons are present in the simulated sequences of these simulated species (except for Species 1 that doesn't display mutational evidence at all), the assigned (global) *PseudoIndex* value for each of them is equal to the attributed one concerning the sub-*PseudoIndex* value for the truncated sequence component of *PseudoIndex*.

4.5 Illustrative examples of analyses ran with *PseudoChecker*

As proof of concept, throughout this section, we will now explore *PseudoChecker* as a whole, daily use, standalone tool. In a first approach, we will run a simulated analysis to introduce *PseudoChecker*'s front-end system and next, we will put *PseudoChecker*'s to practise by analysing two different real test cases, each involving the inference of the coding status of a selected gene in a set of target mammalian species.

4.5.1 Running a *PseudoChecker* analysis

As already reported in Section 4.2, *PseudoChecker*'s front-end system includes a home page required for inputting biological data into analysis, as well as to choose the different available parameters concerning the different components of the pipeline (Figure 4.10). In detail, this home page is divided into 4 different sections:

- On the top left, one dedicated for inputting the sequences underlying the reference species' candidate gene's coding exon(s), coding sequence (CDS), each target species' genomic sequence and predetermined coding sequences (if existing);
- Other for briefly resuming *PseudoChecker*'s pipeline and for providing the instructions required for correctly conducting an analysis (on the top right);
- A specialised section for the selection of the parameters related to the name to be attributed to the analysis, the coding prediction step of the pipeline, the similarity scoring schemes to be applied during the same process, the activation of automatic trimming of untranslated regions (UTRs) and selection of the minimum exon alignment identity percentage value (bottom right);
- Finally, a, by default, hidden section, responsible for providing more complex parameters, being these the manipulable MACSE alignment elementary costs, and the optional extension of the reference species' 3' coding exon or single coding exon alignment for searching a downstream (if missing in the original alignment) final stop codon (bottom left).

Figure 4.10: Overview of *PseudoChecker*'s home page. The home page displays 4 sections: biological data input (top left), selection of the name of the analysis and coding sequence prediction step associated parameters (bottom right), a brief explanation of the pipeline and instructions for running a *PseudoChecker*'s analysis (top right), and finally, a (hidden by default, here activated) section underlying the selection of advanced parameters (bottom left section), for which, this is accessible by clicking at the button 'More options'.

For this demonstration purpose, and as depicted in Figure 4.10, we simulated a candidate gene, the corresponding reference species' exonic sequences and its complete coding sequence (CDS). As for the simulated 9 target species, we simulated, for each of these, the corresponding genomic sequence, for which the header corresponded exactly to the name of the simulated species itself (Species_1 to Species_9). Note that, even though we do not impose any restriction regarding the header name of each target species' genomic sequence, for turning the analysis more intuitive, its name should correspond to the actual name of the species. Underlying the coding sequence prediction step of the analysis, no UTR trimming was performed, the best-fit similarity scoring scheme was selected, the minimum exon alignment identity percentage value, as well as the MACSE alignment elementary costs were left with the default values and we dismissed the optional 3' coding exon alignment extension for searching a (if missing in the original alignment) downstream final stop codon. Finally, as job title, we named the analysis as Test_case _1, and no predetermined coding sequences were included in it (Figure 4.10).

Once the biological data is inputted and parameter selection is completed, a *PseudoChecker* analysis initiates by clicking at the button 'Run analysis' (Figure 4.10). Once it ends, *PseudoChecker* automatically redirects the user to the *PseudoChecker*'s main results page (Figure 4.11), with this constituting a structured page, purposely divided into a set of HTML frames, each responsible for providing specific information produced during a given analysis:

- Alignment visualisation: This HTML frame is responsible for providing the MACSE produced alignment during the second component of *PseudoChecker*'s pipeline (top left);
- Mutational evidence: This, per target species underlying a predicted full coding sequence, and per corresponding predicted coding exon, reports each eventual recognised mutational event retrieved from the MACSE alignment (frameshift mutations and in-frame premature stop codons, top right HTML frame);
- PseudoIndex report: This HTML frame, in turn, displays, for each target species, the corresponding assigned PseudoIndex during the analysis. For the cases where multiple species display the same PseudoIndex value, these are grouped (centre right);
- Coding sequence prediction statistics: This frame presents, per target species, the statistics computed in the first step of *PseudoChecker*'s pipeline for the corresponding predicted coding sequence, including the splice integrity level (bottom right HTML frame);
- Additional information: Responsible for reporting lost exons per target species, information related with the MACSE alignment, the predicted coding sequences that were assigned as reliable or less reliable, the selected similarity scoring scheme concerning the coding sequence prediction step of the analysis, the existence (or not) of included predetermined coding sequences, and, finally, displays the time elapsed during analysis' execution (centre left);
- Exporting data and MACSE alignment costs: this HTML frame provides a series of buttons able to exporting the produced data during analysis (MACSE alignment, predicted coding

sequences at the nucleotide and amino acid level, log of the analysis) and to display (in an hidden by default HTML frame) the selected MACSE alignment elementary costs prior to the analysis' execution (bottom left).

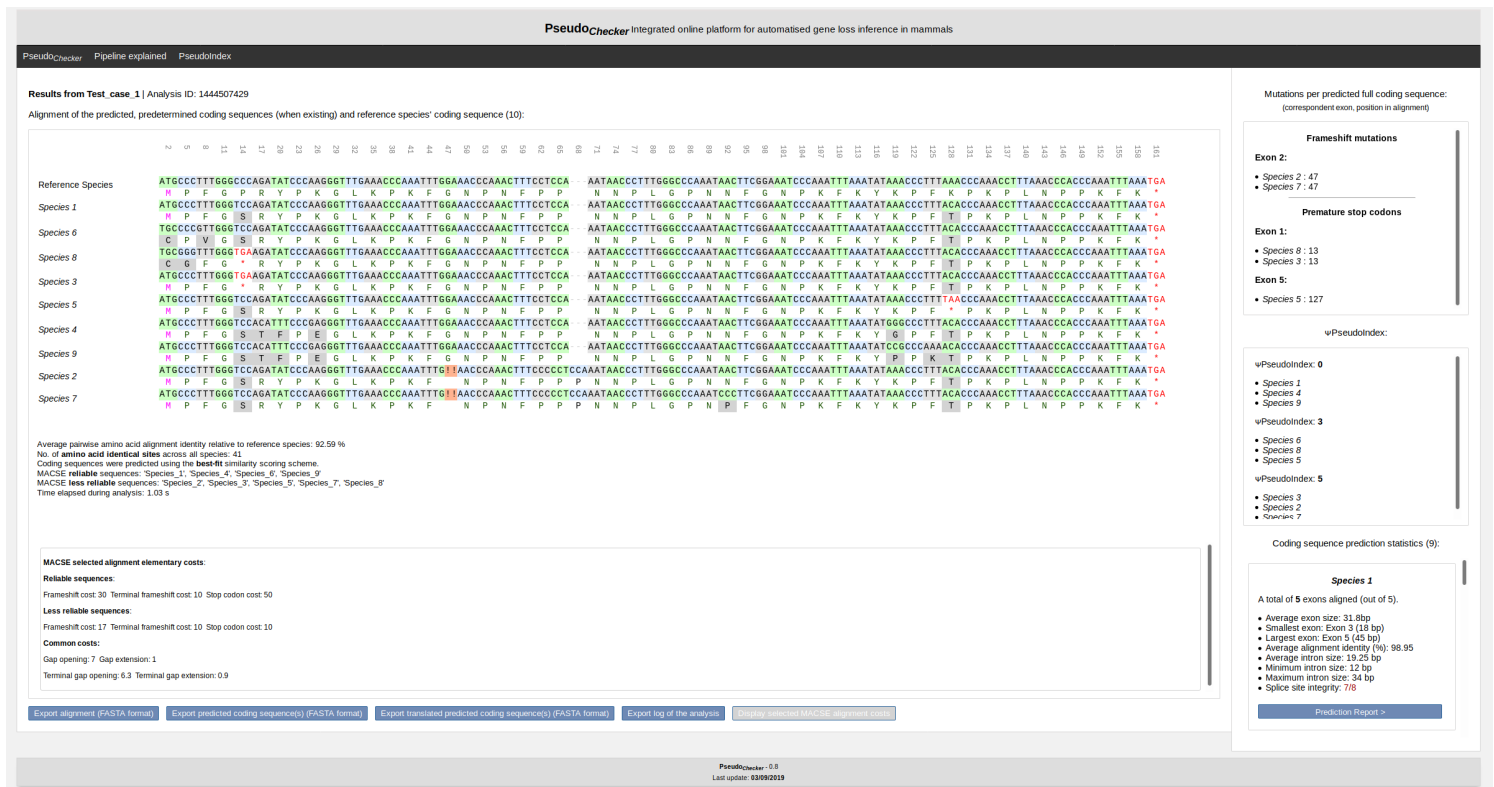


Figure 4.11: *PseudoChecker*'s main results page underlying the demonstrative simulated analysis.

According to the executed analysis (Figure 4.11), *PseudoChecker* inferred the presence of a conserved in-frame premature stop codon between the predicted coding sequences of Species 3 and 8, as well as a non-conserved in-frame premature stop codon arising in the predicted coding sequence of Species 5 (according to the MACSE alignment, top left HTML frame). Relatively to Species 2 and 7, a conserved two-nucleotide deletion (position 47 within the alignment) was also detected. All of these detected mutations are reported into the mutational evidence frame, and the corresponding affected coding exon is also provided (top right).

At PseudoIndex report's frame, since no gene inactivating mutations were found for the Species 1,4 and 9, these are the ones presenting a totally viable coding sequence for the tested simulated gene, hence, revealing a PseudoIndex of 0. Per target species, the coding sequence prediction statistics are depicted in a scrollable HTML frame (bottom right), and the full coding sequence prediction report for each of them is available by clicking at the button 'Prediction Report' (this is only available for in-study genes comprising more than a single coding exon) (Figure 4.11).

The additional information section unveils a high conservation level of the aligned coding sequences relative to the reference species' one (92.59% of amino acid alignment identity), displays the number of conserved amino acid identical sites across species within the alignment (41), the

selected similarity scoring scheme for the coding sequence prediction step, the corresponding headers of the inputted target species' simulated genomic sequences assigned as reliable or less reliable, and finally, the time elapsed during analysis (1.03 seconds). Below, 4 buttons are available for exporting the produced data during the analysis, and an additional one reveals in a (hidden by default) HTML frame, the MACSE alignment elementary costs selected prior to the analysis execution, for which, in this case, this is activated (Figure 4.11).

The accessible prediction report page via *PseudoChecker*'s main results page underlying each target species' coding sequence, visually displays the result of each reference species' coding exon alignment against the corresponding target species' genomic sequence occurred during the first step of the pipeline. The adjacent (viable or not) splice sites of each predicted exon are also represented, differentially coloured whether conserved or not (Figure 4.12).

The alignment identity of each exon alignment, the position of the target species' genomic region at where the alignment starts and ends is also displayed. At the right, the same overall coding sequence prediction statistics displayed in *PseudoChecker*'s main results page, however, with two additional metrics, namely the predicted coding sequence's percentage of guanine-cytosine content (GC-Content) and length (in base pairs) are presented. Finally, two available buttons are also provided, one for exporting the predicted coding sequence at the nucleotide level and the other at the amino acid level (FASTA format) (Figure 4.12).

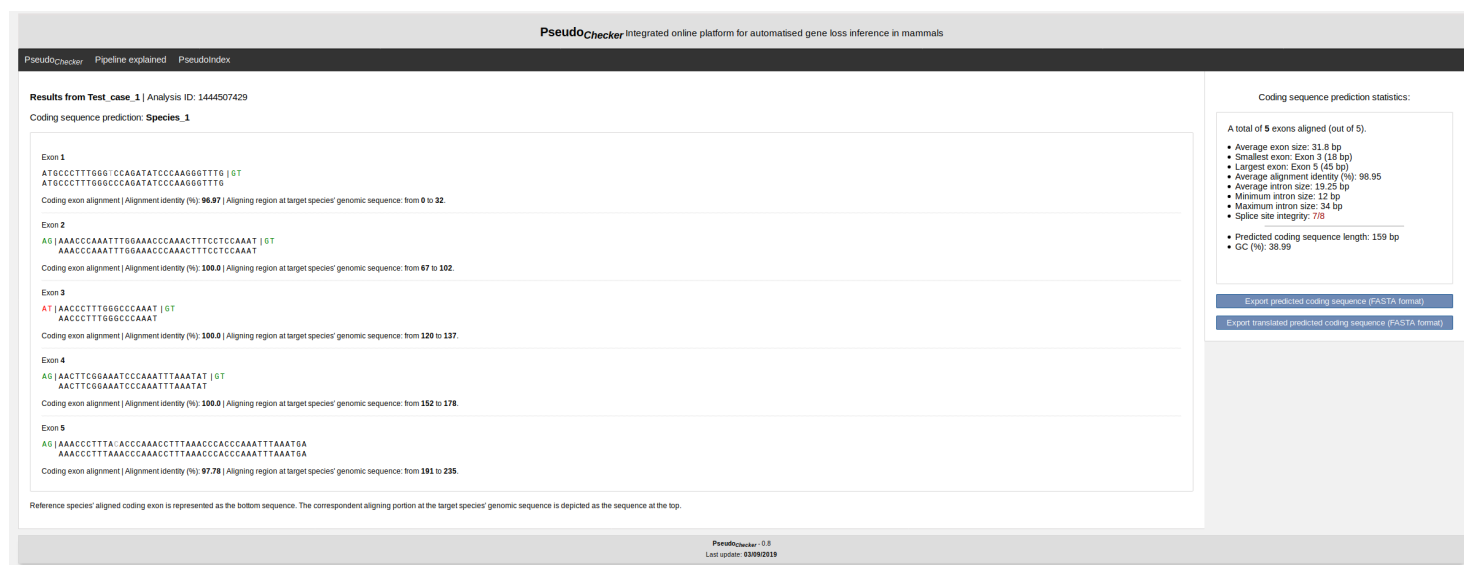
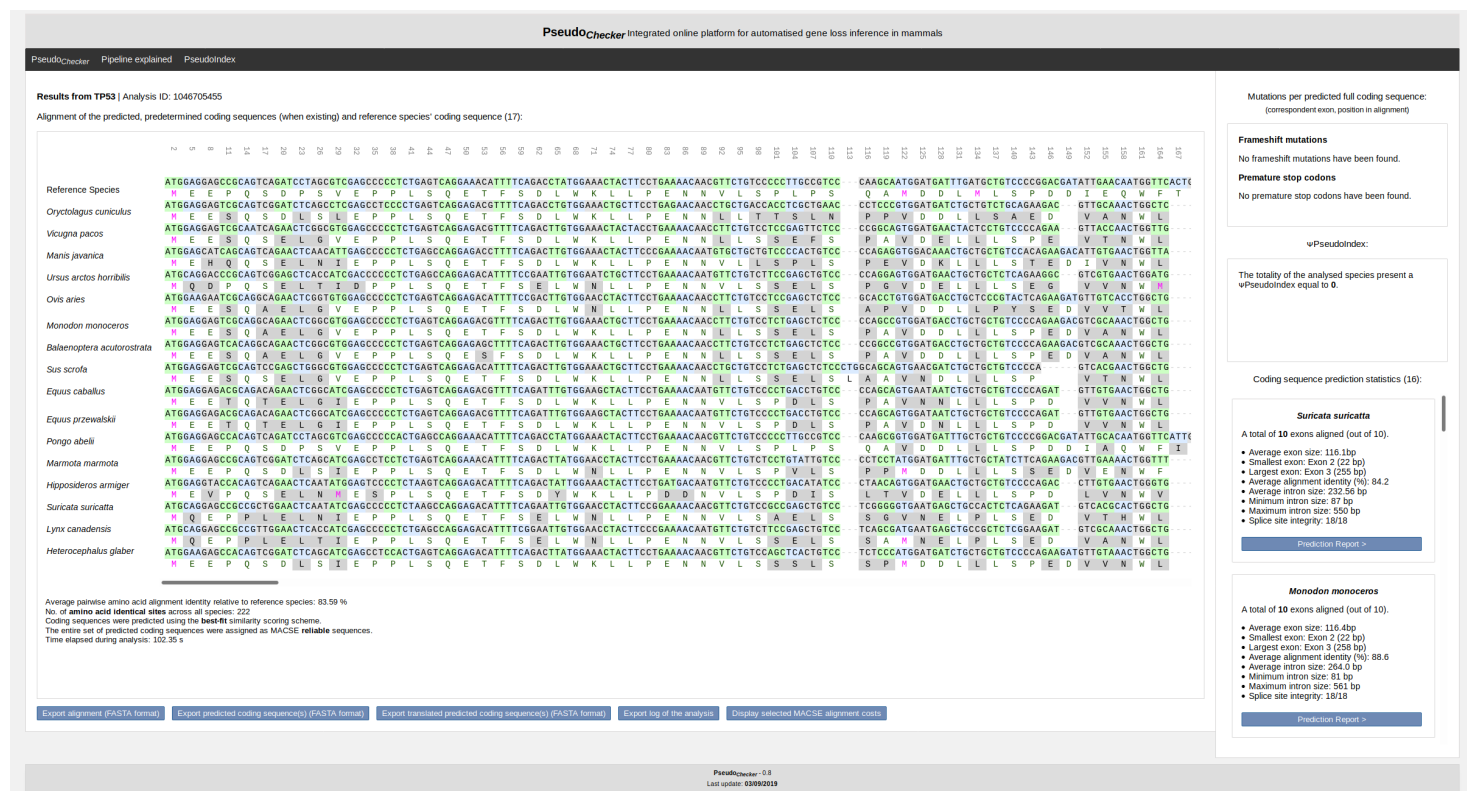


Figure 4.12: *PseudoChecker*'s coding sequence prediction report for simulated Species 1. Each exon alignment displays, at the bottom, the aligned sequence of the reference species' coding exon, and, at the top, the aligning portion of the corresponding target species' genomic sequence. Adjacent splice sites are separated from the exonic predicted sequence by a '|', and below, the alignment identity and alignment coordinates at the target species' genomic sequence are represented. General coding sequence prediction statistics are represented at the right, and the buttons below allow for exporting the predicted coding sequence at nucleotide and amino acid levels (button at the top, and bottom, respectively).

4.5.2 PseudoChecker confirms the functionality of *TP53* gene across mammals

The *TP53* gene (also known as *P53* gene) is widely studied by cancer biologists since it first emerged in the late 1970s as a character in cancer discourse. Currently, it is known that it plays a pivotal role as a tumour suppression agent, and displays over 80 000 related publications to date, being believed to be highly pervasive across the mammalian phylogeny [102, 130]. Given its importance for organisms as a constraint against tumour growth, and since there are no reports of *TP53* loss over any species of this group, this constitutes an interesting test case regarding gene functionality to be addressed by *PseudoChecker*.

Here, we applied *PseudoChecker* for confirming the functionality of this gene in several mammals. We set human as the reference species and downloaded from the NCBI Gene database the corresponding *TP53* longest isoform’s coding exons and coding sequence. As target species, a total of 16 randomly selected mammalian species and the corresponding genomic sequences underlying an already existent *TP53* gene annotation, again, available at NCBI Gene database, were directly collected and imported into *PseudoChecker*. Finally, for the coding sequence prediction step, we opted for the best-fit similarity scoring scheme and, since we, as in-study target species, incorporated somewhat divergent species from human, we set the minimum exon alignment identity to 50%. All the remaining parameters were left with the default values.



As expected, concerning the analysis outcome, and although there are some less conserved blocks across the alignment, *PseudoChecker* didn't recognise any gene inactivating mutation in any target species, therefore assigning, for each of them, a PseudoIndex equal to 0, confirming the functionality of the *TP53* gene in all the tested lineages (Figure 4.13).

4.5.3 *PseudoChecker* corroborates the convergent loss of *PON1* gene in several marine mammals

An interesting case of convergent gene loss concerns the Paraoxonase 1 (*PON1*) gene. This gene is responsible for encoding a bloodstream enzyme that reduces the oxidative damage over lipids in low- and high-density lipoprotein particle content, that, as a consequence, culminates in preventing atherosclerotic plaque formation. Notably, this gene presents genetic lesions in cetaceans, in pinniped and sirenian lineages, being, notwithstanding, intact in all terrestrial mammals. This convergent loss event is likely explained by parallel mechanistic shifts in marine mammals ancestors' lipid metabolism and/or bloodstream oxidative environment that discarded the necessity of keeping a viable *PON1* gene [127].

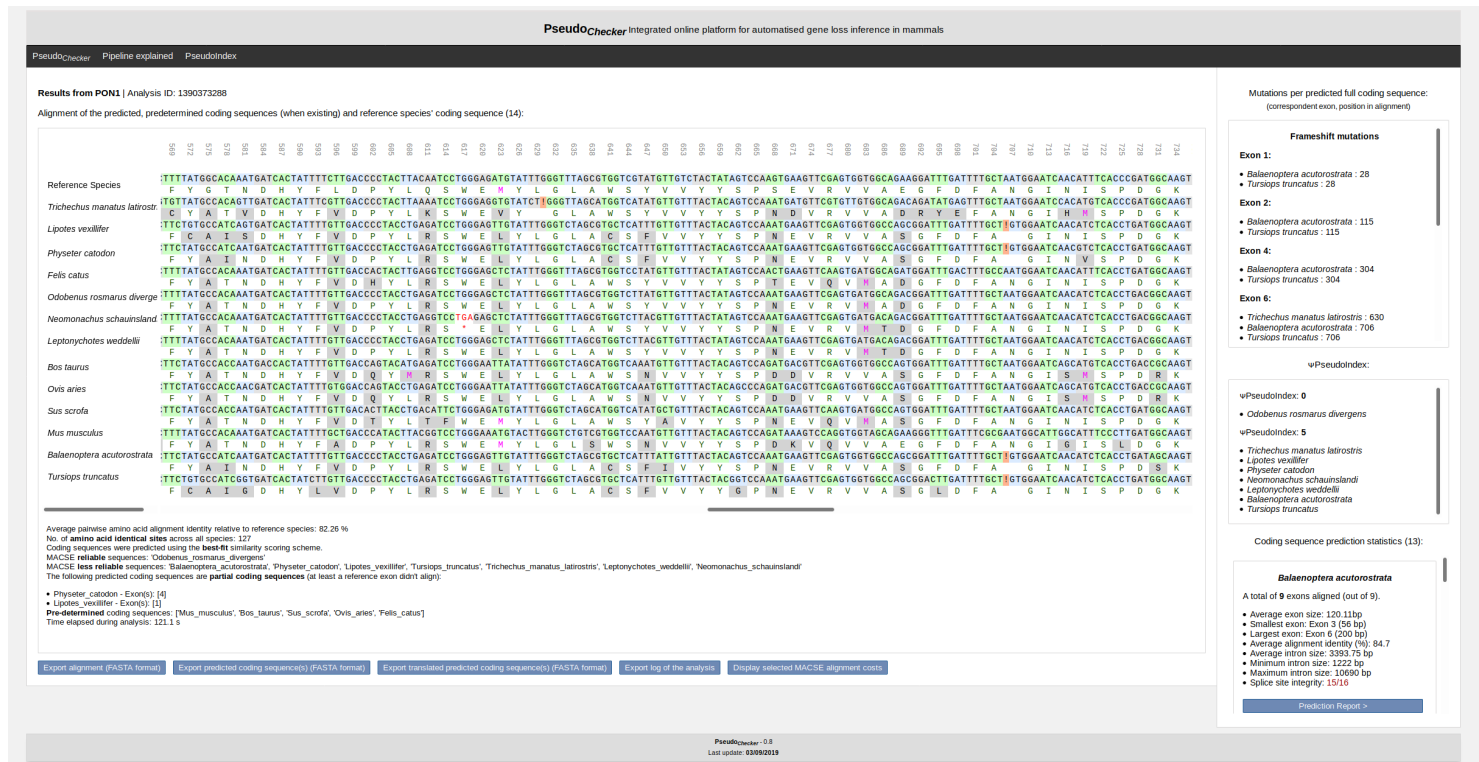
Interestingly, however, the loss of *PON1* original function led to the inactivation of a self-defence mechanism within the affected lineages when confronted by specific man-made organophosphorus compounds, implying potential risks for those species habiting in modern environments [127].

From this, in this illustrative example of *PseudoChecker* in action, we aimed to re-confirm the iconic loss of *PON1* gene in the scrutinised marine mammals at the original published study [127]. To accomplish this, we inspected all affected lineages with *PON1* gene annotation available at the Gene database from NCBI and incorporated into *PseudoChecker* the underlying *PON1* annotated genomic sequences. As reference species, and similarly to the previous illustrative case, we used human, and, from the same database, we retrieved the corresponding coding exons and coding sequence of the longest annotated *PON1* gene isoform.

Furthermore, for depicting the capability of *PseudoChecker* distinguishing functional from lost genes, we also included the unique marine mammal with *PON1* gene annotation at NCBI that was proven to not contain any disrupting mutation for the same gene, *Odobenus rosmarus divergens*, the Pacific walrus.

Finally, for demonstrative purposes of the ability of *PseudoChecker* including external functional gene datasets into a given analysis, we also inputted as predetermined coding sequences, the longest annotated isoform of *PON1* gene of 5 different randomly selected terrestrial mammals: *Bos taurus* (cattle), *Sus scrofa* (pig), *M. musculus* (house mouse), *Ovis aries* (sheep) and *Felis catus* (domestic cat).

The best-fit similarity scoring scheme was selected, the minimum exon alignment identity was set to 50%, the MACSE alignment elementary costs were left with the default values and no optional automatic extension of the 3' coding exon was requested. The conducted *PseudoChecker* analysis's results are depicted in Figure 4.14.

Figure 4.14: PseudoChecker's main results page underlying the *PON1* gene test case.

As expected, looking at the results (Figure 4.14), the *Pacific walrus* is the only species displaying a PseudoIndex of 0, while all the others, since display an inactivated version of *PON1* gene, display a PseudoIndex equal to 5.

Remarkably, the trans-species conserved mutations observed in the alignment suggest that an ancestral *PON1* inactivation event occurred in their common ancestor, particularly, the conserved single nucleotide deletion across *Lipotes vexillifer* (Yangtze River dolphin), *Physeter catodon* (Sperm whale), *Balaenoptera acutorostrata* (Minke whale) and *Tursiops truncatus* (common bottlenose dolphin) suggests that the *PON1* gene inactivation occurred prior to the cetacean radiation.

In contrast, *PON1* loss across non-cetacean closely related marine mammals is supported in our analysis by the non-conserved loss-of-function mutations detected within the predicted sequences of *Trichechus manatus latirostris*, the Florida manatee, *Leptonychotes weddellii* (Weddell seal), and *Neomonachus schauinslandi* (Hawaiian monk seal).

Finally, the inputted predetermined (functional) coding sequences (mentioned in the additional information section), automatically included in the alignment, as one would expect, do not present any deleterious mutation (Figure 4.14).

4.6 Limitations

Even though *PseudoChecker* displays several advantages over the few existing methods, it still presents two pitfalls. First, *PseudoChecker* is only able to declare identified frameshift mutations and premature stop codons for predicted full coding sequences, being predicted partial coding sequences excluded from this feature (those presenting the loss/absence of at least one in-study gene's reference species' coding exon).

This happens since, as we have mentioned, MACSE will automatically impose exclamation marks (!) in the most appropriate alignment location for maintaining the original structure of the reading frame (following a given similarity scoring scheme). However, when such single or multiple exons are missing, if their absence results into the disruption of the reading frame within a given partial coding sequence, it is, from our experience, very likely that, aiming to compensate the same reading frame's disruption, exclamation marks will arise adjacently to the neighbouring exons of the missing ones (Figure 4.15).

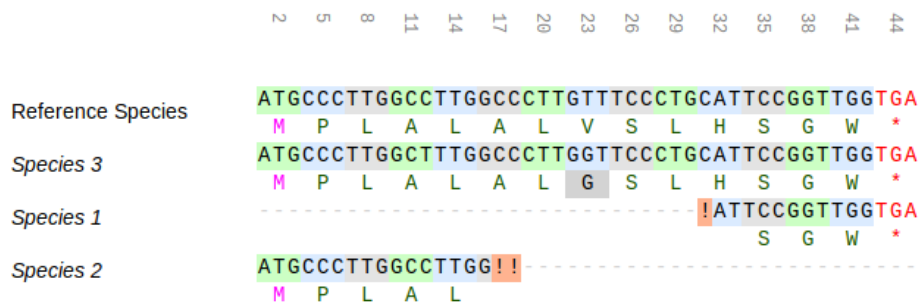


Figure 4.15: Illustrative example of a *PseudoChecker* simulated analysis underlying a MACSE alignment comprising sequences with their reading frame disrupted by the absence of exons. Notice that the exclamation marks arising within Species 1 and 2 corresponding aligned sequences do not correspond to biological nucleotide deletions, but are rather the result of MACSE adjustments for maintaining the original structure of their reading frame, after being disrupted by the respective simulated loss of (adjacent) coding exons.

Importantly, this constitutes an issue since it might be difficult for the user to distinguish between real biological mutations from alignment adjustments produced by MACSE aiming to preserve the structure of the reading frame.

Moreover, a greater problem occurs when dealing with high gene erosion in a given target species. For these cases, during the MACSE alignment step, it is possible that nucleotide bases that should be aligning at the orthologous ones, erroneously align at non-orthologous bases, leading to a non-accurate inference of the location underlying both recognised frameshift mutations and stop codons (Figure 4.16). Given these both factors, we decided to exclude predicted partial coding sequences from the feature underlying the reporting of mutational evidence.

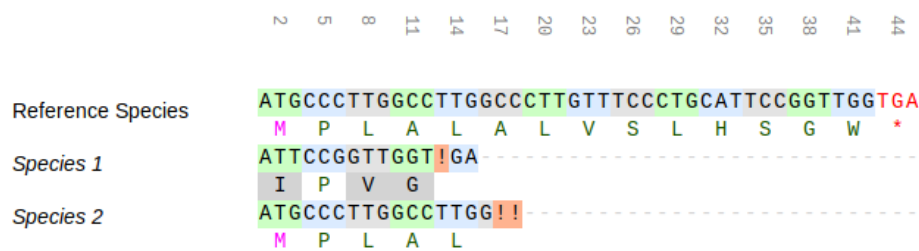


Figure 4.16: Illustrative example of a *PseudoChecker* simulated analysis underlying a MACSE alignment presenting erroneously aligning regions mediated by the aggressive loss of coding exons. Even though the predicted coding sequence underlying Species 2 is, in fact, accurately aligning at the orthologous region of the reference species' one, however, the predicted coding sequence for Species 1 is erroneously aligning at the 5' end of the reference species' one, whereas it should be aligning at the 3' end region of it, the truly orthologous sequence region to one another.

Finally, the second limitation of our approach, in its turn, lies in the inability of, similarly to the gene annotation-based and whole-genome alignment-based gene loss inference approaches (see Section 3.2), validating annotated mutations with unassembled sequencing reads obtained from either genomic and transcriptomics projects from the respective affected species.

However, this is not problematic since the real power of *PseudoChecker* comes when analysing inactivation occurrences of a given gene in multiple closely related target species. By doing so, frameshift mutations and/or premature stop codons conserved in at least two closely related lineages automatically reduce the likeliness of the observed mutational evidence being the result of sequencing and/or genome assembly artefacts, supporting instead, that these are the result of actual real biological mutations [172].

Chapter 5

Pseudo*Checker*'s performance evaluation

5.1 Experimental design

For testing the classification performance of our approach, we applied it directly to both ground truth cases of published lost genes in mammals and presumably functional genes across Mammalia, independently. For the first case, we scrutinised recently published studies of mammalian gene loss occurrences, the afflicted lineages, and confronted Pseudo*Checker* with a total of 30 lost genes. In contrast, for the latter case, Pseudo*Checker* was applied to a set of 30 presumably functional genes across the mammalian lineage, with these being determined according to the following pre-established criteria.

In the assumption that highly expressed protein-coding genes are less prone to suffer deleterious mutations [5], therefore being lost by mutational inactivation, first, we scrutinised The Human Protein Atlas¹ database for recovering 30 highly expressed genes: 15 retrieved from 15 randomly selected tissues with available expression data, the tissue-specific atlas, and other 15 collected from the cell-specific atlas, a sub-database that contains expression information for different human cellular compartments. To concretely establish the functional protein-coding gene repertoire applied to Pseudo*Checker*, we imposed two filters to ensure the orthology and viability of each selected gene across Mammalia. For each of the selected tissue or cellular compartment samples, the most expressed gene: 1) was assessed to be annotated in at least one belonging species per each 19 out of 20 orders of mammals (according to the latest mammalian classification [21]) with available and annotated genomes at NCBI; 2) and should not be published as lost in any mammalian lineage. If both of these conditions were verified, then the correspondent gene was included in the repertoire of tested functional genes. If not, the subsequent most expressed genes from the same inspected sample were screened until being found a given one obeying to both conditions, therefore, being included into the same final catalogue.

¹www.proteinatlas.org

For each of the 60 tested genes, and for each of the three similarity scoring schemes provided by *PseudoChecker* underlying the first component of the pipeline, a *PseudoChecker* analysis has been run. In this sense, a total of 180 analyses were conducted, with the similarity scoring scheme variation being relevant for measuring the impact that different schemes may have in the classification outcome. For all analyses, we fixed human as the reference species, not only for standardising the reference lineage in all of them, but also given the fact that the corresponding genome is by far the most firmly established genome amongst mammals, not only displaying a comprehensive and vastly complete gene annotation, but also including manual curation for the vast majority of annotated genes [172].

Per analysed gene, human-curated annotated sequences were preferred, the longest isoform of the annotated ortholog was retrieved from NCBI's Gene database and inputted into the analysis. No extension of the 3' (or single-exon) alignment for searching a missing (in the original alignment) final stop codon was requested, and the MACSE elementary alignment costs were left with the default values, not only since it would be impracticable to test all possible combinations of the respective manipulable alignment elementary costs, but also given the fact that, as we have indicated before, the default alignment elementary costs are indeed reliable for running the vast majority of analyses (Section 4.3.2).

Relatively to the target lineages to be included in each analysis, concerning the ones involving lost genes, the respective affected species with annotation available at NCBI for the in-study gene were inspected, and if existent and displaying no assembly gaps at the corresponding annotated genomic sequence (represented by contiguous N's, that could negatively influence the performance analysis' outcome), these were directly collected and inputted into the analysis. If not, the corresponding lineages were excluded from the analysis.

For the analyses actuating over presumable functional genes, for each in-study gene, and for ensuring the maximum representability of mammals within the conducted analyses, the same previously scrutinised 20 mammalian orders were inspected for the presence of at least one belonging species presenting annotation of the in-study gene at NCBI. This should not display assembly gaps (N's) at the corresponding genomic sequence and should not present the low-quality protein tag in at least one annotated gene isoform, an NCBI RefSeq tag indicating that the annotated sequence was modified for correcting deleterious indels and stop codons (arising either from assembly artefacts or real biological mutations that, being a doubtful case could influence our analysis). Per mammalian order, if a given lineage violated at least one of these conditions, the other representative individuals of the same order (if existent) were scrutinised for a viable gene annotation of the in-study gene. If not, the corresponding order was excluded for the analysis.

Finally, since we employed a great diversity of species within each analysis and used human as the reference species, the minimum exon alignment identity was fixed to 50% for all analyses. All 180 analyses were conducted within *PseudoChecker's* development local machine and all of these individually ran on a single processor. For each of them, we recovered the assigned PseudoIndex per tested target species and the time of analysis' execution, in seconds.

The entire produced raw data associated with the three sets of 30 analysis each underlying functional genes within the respective target lineages (each set corresponding to a different tested similarity scoring scheme) is included at Table A.1. This presents, per tested gene and corresponding isoform, the included target species and the PseudoIndex attributed to each of them per tested similarity scoring scheme. Similarly, the entire raw data associated to the three sets of 30 analysis each underlying lost genes within the tested species is included at Table B.1. This includes, again, per tested gene and corresponding isoform, the included target species and the PseudoIndex assigned to each of them per tested similarity scoring scheme. For all the executed analyses, the elapsed time in seconds is also presented.

To each of the total 6 sets of 30 analysis each, corresponds a set of p observations (or cases), for which p corresponds to the sum of the number of target species included in each composing analysis. To each observation (in other words, each species) corresponds a given PseudoIndex that will be used for computing the further explored evaluation metrics. Concretely, regarding lost genes at the corresponding tested species, each of the three sets displays 155 observations (a total of 155 species) and each set underlying presumable functional genes over the corresponding target species contains a total of 479 observations (479 species).

5.2 Evaluation metrics

Per tested similarity scoring scheme underlying *PseudoChecker*'s coding sequence prediction step of its pipeline, we are evaluating the percentage of well classified gene functionality events, as well as the percentage of well classified gene loss events. We define a gene loss event as the loss of a given gene in a given species, and a gene functionality event as the (presumable) presence of a given viable gene in a given species.

Taking advantage of each computed PseudoIndex per species (or observation), then, for each set of 30 analysis involving functional genes at the tested target species (each underlying a different similarity scoring scheme), we calculated the fraction of predicted gene functionality events by *PseudoChecker* over the actual tested gene functionality events, in other words, the total number of gene functionality occurrences that were correctly detected by our approach over the total number of tested gene functionality occurrences.

The percentage of well classified gene functionality events (WCGFE) is then given by the formula:

$$WCGFE = \frac{PGFE}{AGFE} * 100 \quad (5.1)$$

where PGFE constitutes the number of predicted gene functionality events by *PseudoChecker* and AGFE corresponds to the number of actual tested gene functionality events.

Regarding each set of the 30 analysis involving lost genes at the respective inputted target species, similarly, we computed the fraction of the total number of predicted gene loss events by *PseudoChecker* over the total number of actual tested gene loss events. This is given as the percentage of well classified gene loss events (WCGLE), expressed by the formula:

$$WCGLE = \frac{PGLE}{AGLE} * 100 \quad (5.2)$$

where PGLE constitutes the number of predicted gene loss events by *PseudoChecker* and AGLE corresponds to the number of actual tested gene loss events.

For computing both of these ratios, however, we first need to turn our approach into a binary classifier by converting the PseudoIndex scale into two different categories. For accomplishing this, and following the PseudoIndex biological significance for each composing value of its scale (see Section 4.4), tested species underlying values of PseudoIndex between 0 and 2 yield a predicted gene functionality event, whereas values of PseudoIndex equal to 4 or 5, in its turn, will correspond to a predicted gene inactivation event.

For each observation (or species), depending on its PseudoIndex assigned value, therefore, this will either result in a predicted gene functionality event or a predicted gene loss event. Finally, since these represent doubtful cases that must be manually inspected, species presenting PseudoIndex values equal to 3 should not be considered within *PseudoChecker's* classification performance evaluation.

5.3 Results

Before diving into the final results concerning both mentioned classification evaluation metrics, we first should assess the percentage of discarded observations per tested similarity scoring scheme for both the 30 analysis involving lost genes and the 30 remaining ones involving functional ones (Table 5.1).

Table 5.1: Percentage value of discarded observations per tested similarity scoring scheme and type of analysis (involving lost genes or functional ones). (CR): Optimised for closely related species; (SD): Optimised for slightly divergent species; (BF): Best-fit similarity scoring scheme.

	Lost Genes			Functional Genes		
Similarity scoring scheme	CR	SD	BF	CR	SD	BF
Discarded observations	6	2	3	7	11	5
Total observations	155	155	155	479	479	479
Discarded observations (%)	3.87	1.29	1.93	1.46	2.29	1.04

As we can see, for all sets of 30 analysis each, we discarded a very small percentage of cases for further classification quality evaluation, which, in some way, means not only that *PseudoChecker* is indeed capable of producing a low-level rate of doubtful predictions, but also that this slight discardment process is less likely to influence our classification quality evaluation outcome.

Removed the doubtful observations from each set, per tested similarity scoring scheme, the percentage of well classified gene functionality events (Table 5.2), as well as the percentage of well classified gene loss events (Table 5.3) can be now calculated.

Table 5.2: Percentage value of well classified gene functionality events per tested similarity scoring scheme including the obtained average value for all of these. (CR): Optimised for closely related species; (SD): Optimised for slightly divergent species; (BF): Best-fit similarity scoring scheme.

Similarity scoring scheme	CR	SD	BF
PGFE	433	456	465
AGFE	472	468	474
WCGFE (%)	91.73	97.43	98.10
Average WCGFE (%)	95.75		

First, looking at the classification quality of *PseudoChecker* in identifying functional genes as such, we indeed obtained satisfactory results. As we have stated when introducing the first component of *PseudoChecker*'s pipeline (section 4.3.1), since it will condition the prediction of each target species' coding sequence, the accurate selection of the similarity scoring scheme used for its execution is of utmost importance for the success of gene functionality inference, as any bad prediction evolving the composing exons will, in the worst-case scenario, ultimately lead to the introduction of false mutations, further being misclassified as real ones.

Therefore, given its splice-site and reading-frame aware nature, testing, in the worst-case scenario, a variety of alignments for each reference species' coding exon, as we were somewhat expecting, the best-fit similarity scoring scheme was the best performer in this field, displaying the highest percentage value of well classified gene functionality events amongst the remaining available similarity scoring schemes. Indeed, it was capable of correctly predicting 465 gene functionality events over a total of known 474 ones, therefore reinforcing our recommendation for its usage within the majority of *PseudoChecker* analyses.

In contrast, looking at the similarity scoring scheme optimised for closely related species, this, even though achieving very reasonable results, with over 90% of correct classifications underlying gene functionality, it performed not so well when compared to the first-mentioned one (Table 5.2). Given the fact that we made all efforts to incorporate a representative individual of each mammalian order into the analyses underlying functional genes, and in the sense that we have used human as reference species for each of them, this is likely explained since this similarity scoring scheme is not, and as its name indicates, highly suitable to act over more distant species, like, in our case, human relative to monotremes represented in our dataset by *Ornithorhynchus*

anatinus, the platypus, or relative to marsupials, represented in our analyses by *Monodelphis domestica* (gray short-tailed opossum). As a matter of fact, for the cases where this similarity scoring scheme failed, a substantial part of these, actually corresponded to the same referred more distant lineages relative to human (see Table A.1).

Relatively to the similarity scoring scheme optimised for slightly divergent species, likely due to its more relaxed mismatch penalty when compared to the one optimised for closely related species, therefore being more tolerant to nucleotide substitutions within the corresponding produced alignments, it failed in fewer cases when compared to the latter one. Yet, since it only tests one similarity scoring scheme per executed alignment, therefore limiting a possible increased diversity of different alignments that is offered by the best-fit similarity scoring scheme, it is also understandable that it performed a little bit worse than best-fit one.

Notably, and globally speaking, however, under this component of our performance evaluation, *PseudoChecker* performed very well in correctly capturing gene functionality, with an average value of well classified gene functionality events across the three different tested similarity scoring schemes of 95.75% (Table 5.2).

Moving to *PseudoChecker's* power in classifying actually lost genes as such, we also obtained fairly good results. Nonetheless, here we should be more careful while analysing the results per tested similarity scoring scheme. Since we, again, know that an erroneous coding sequence prediction of a given in-study gene in a given target species will almost lead, in some way or another, to the accumulation of (false) deleterious mutations, arising, for example, from bad exon boundaries predictions, therefore culminating in the false inference of a gene inactivation event, it is not possible to distinguish if a given gene loss event was classified as such due to the presence of (real) biological mutations that conducted to a *PseudoIndex* value for the affected species superior to 3, the result of a bad coding sequence prediction that led to the same effect, or a combination of both factors.

Therefore, and even though we obtained gratifying values of well classified gene loss events for all tested similarity scoring schemes (Table 5.3), we should reline our analysis specially attending to the best-fit similarity scoring scheme, that, as already stated, aiming to produce the best and most biologically meaningful coding sequence prediction, gives us at the end of the analysis, a higher level of certainty that the classification outcome is the result of real biological identified mutations and not errors concerning the coding sequence prediction step.

Table 5.3: Percentage value of well classified gene loss events per tested similarity scoring scheme including the obtained average value for all of these. (CR): Optimised for closely related species; (SD): Optimised for slightly divergent species; (BF): Best-fit similarity scoring scheme.

Similarity scoring scheme	CR	SD	BF
PGLE	147	147	149
AGLE	149	153	152
WCGLE (%)	98.65	96.07	98.02
Average WCGLE (%)	97.58		

As observable in Table 5.3, the best-fit similarity scoring scheme yielded a very positive outcome, only misclassifying a total of 3 cases within a total of 152 cases. In fact, the failing cases for this component of the evaluation, this is, actual gene loss events misclassified as gene functionality events are the result, either from different criteria applied from the original authors for classifying the respective gene at the respective target species as lost when compared with our PseudoIndex, or, particularly speaking, from its inherent (one more time, in our opinion) reasonable non-penalisation of splice site inactivating mutations.

Since different approaches will yield different results, and since they depend on different assumptions, display different biases and are likely to make use of different criteria for assigning a given genomic fragment as a pseudogene, there is no way to systematically resolve all the differences between such methods [91].

Finally, making an overall balance, taking into account the obtained results for the performed classification evaluation metrics underlying gene loss and gene functionality events, under the tested conditions, we can consider that, not only *PseudoChecker* was highly capable of accurately classifying real lost genes as such, as well as functional genes where its functionality at the involved target species was also presumable, but also that PseudoIndex is actually well calibrated for performing this type of binary classification. Finally, taking into account the temporal information recorded for each ran analysis, and even though this does not constitute essential part of the analysis, we achieved a very reasonable average value of 133.09 seconds of analysis time (approximately 2.22 minutes).

Chapter 6

Conclusions

6.1 Main contribution

Great attention in the past has been paid to gene duplication as the pervasive source of genetic variation amongst biological lineages. With the recent increase in genomic data, however, it is now known that gene loss as a mechanism of evolutionary change is much more prominent than what would be expected a few years ago [5].

The establishment of a comprehensive gene loss catalogue covering the totality of gene inactivation events and the respective affected *taxa* is, indeed, of utmost importance for providing valuable information underlying the molecular mechanisms that accompanied the establishment of new species, and to unveil the links between lost genes and phenotypic traits [5].

Yet, to build a complete gene loss database, it is ultimately required to develop computational approaches capable of reliably identifying and mapping gene losses over the tree of life. A giant breakthrough in this area was precisely established with the development of *PseudoChecker*: the first integrated online platform for automatised gene loss inference in mammals. Our approach is aimed to be repeatedly applied, by any, even non-experienced users, in researches directed to identify and unveil the molecular signatures underlying mammalian gene inactivation occurrences, in a straightforward, convenient, systematic and highly accurate process.

Our comparative genomics-based approach, as opposed to all existing methods, is the unique amongst those displaying full automation, the built-in inclusion of an index of pseudogenisation taking into account multiple carefully selected factors, and finally, but not least important, displaying its integration as a web service, not only allowing us to keep *PseudoChecker* as a multi-platform and always updated bioinformatics tool, but facilitating its expedite and easy usability for quickly identifying gene loss.

6.2 Perspectives and future work

Nonetheless, at this stage, *PseudoChecker* Version 0.8 still constitutes a prototype. Functionally speaking, even though in our development machine, *PseudoChecker* is fully usable, meets all the predetermined goals, and is prepared for being immediately launched on the web, a few improvements at both functional and technical levels can still be made.

As future work, and concerning functional enhancements, we aim to develop novel features that include:

- Support for analyses involving multiple genes at once, useful for analysing groups of genes directly associated with a given phenotype;
- Support for the study of polymorphic loss, this is, gene loss affecting, per given lost gene, only one of the two alleles of the affected species, therefore, being especially relevant for unveiling insights related with ongoing processes of gene loss in natural populations [5];
- Additions to the front-end system for visually displaying the information underlying the multiple tested scoring schemes and the respective alignment outcomes for each exon alignment following the best-fit similarity scoring scheme;
- Support for different alignment colouring schemes either for both MACSE and reference species' coding exons alignments;
- Finally, the support for built-in selection analyses involving the in-study sequences, using, for this effect, widely adequate software such as codeml from PAML [203].

As for the technical level, improvements related to *PseudoChecker*'s computational performance can also be employed. The most relevant one, from our point of view, is, in fact, the implementation of multi-core processing. This would not only increase the speed of analyses since multiple exon alignments could be executed at the same time, but it would also avoid the requirement of using the greedy approach applied in the best-fit similarity scoring scheme. As a consequence, since all 40 similarity scoring schemes would always be tested, this could yield even better results when searching for the most adequate exon alignment (against a given test species' genomic sequence) per each reference species' coding exon.

Additionally, the implementation of the employed variation of the Needleman-Wunsch algorithm for conducting semi-global exon alignments in mid-level, compiled programming languages such as C or C++, instead of Python, would also benefit from great speed improvements, that, once again would culminate in the decrease of the time of analyses' execution.

Finally, a natural evolution of *PseudoChecker* would consist in nothing else but its expandability to non-mammalian animal lineages. This, in fact, constitutes the ultimate future goal for *PseudoChecker*, and since the strong foundation for this effect is already established, we strongly believe that with future small improvements and adequate testing, *PseudoChecker* can easily

be suitably applied in gene loss researches targeted for other animals, therefore, aiding and promoting the study of this new paradigm of evolution in a wider range of lineages.

Appendix A

Raw data associated with Pseudo*Checker*'s performance evaluation over gene functionality events

This appendix contains the raw data associated to the 90 analysis conducted during Pseudo*Checker*'s performance evaluation over gene functionality events. This comprehends a main table that, besides other accessory information, includes, per tested gene, per target species, and per tested similarity scoring scheme underlying the coding sequence prediction step of Pseudo*Checker*'s pipeline, the corresponding assigned PseudoIndex. Finally, the time of execution concerning each of the 90 analysis is also disclosed.

Table A.1: Raw data underlying the 90 analysis conducted during PseudoChecker's performance evaluation over presumable cases of gene functionality. This table comprehends, per analysed gene (2nd column), the correspondent NCBI's accession number of the retrieved human ortholog isoform (3rd column), the correspondent number of coding exons (4th column), the employed target lineages into analysis and the correspondent mammalian order (5th and 6th columns, respectively), the PseudoIndex (PI) assigned for each species according to the three tested similarity scoring schemes (the one optimised for closely related species (CR), the one optimised for slightly divergent species (SD) and the best-fit similarity scoring scheme (BF)), and the corresponding time of analysis' execution (TA), in seconds (last three columns, each per distinct tested similarity scoring scheme).

<i>n</i>	Gene	Isoform	E	Presumably functional in	Order	PI			TA (s)		
						CR	SD	BF	CR	SD	BF
1	GFAP	NM_002055.5	9	<i>Monodelphis domestica</i>	Didelphimorphia	0	0	0	116.32	119.44	115.93
				<i>Phascolarctos cinereus</i>	Diprotodontia	0	0	0			
				<i>Orycteropus afer afer</i>	Tubulidentata	0	0	0			
				<i>Chrysochloris asiatica</i>	Afrosoricida	0	0	0			
				<i>Elephantulus edwardii</i>	Macroscelidea	0	0	0			
				<i>Loxodonta africana</i>	Proboscidea	0	0	0			
				<i>Trichechus manatus latirostris</i>	Sirenia	0	0	0			
				<i>Dasypus novemcinctus</i>	Cingulata	0	0	0			
				<i>Pan troglodytes</i>	Primates	0	0	0			
				<i>Ochotona princeps</i>	Lagomorpha	0	0	0			
				<i>Mus musculus</i>	Rodentia	0	0	0			
				<i>Condylura cristata</i>	Eulipotyphla	0	0	0			
				<i>Phyllostomus discolor</i>	Chiroptera	0	0	0			
				<i>Canis lupus familiaris</i>	Carnivora	0	0	0			
				<i>Manis javanica</i>	Pholidota	0	0	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			
				<i>Bos taurus</i>	Artiodactyla	0	0	0			
2	CHGB	NM_001819.3	5	<i>Ornithorhynchus anatinus</i>	Monotremata	3	3	3	867.21	976.82	1016.21
				<i>Sarcophilus harrisii</i>	Dasyuromorpha	5	0	0			
				<i>Phascolarctos cinereus</i>	Diprotodontia	5	0	0			
				<i>Orycteropus afer afer</i>	Tubulidentata	0	0	0			
				<i>Chrysochloris asiatica</i>	Afrosoricida	0	0	0			
				<i>Pan troglodytes</i>	Primates	0	0	0			
				<i>Ochotona princeps</i>	Lagomorpha	0	0	0			
				<i>Mus musculus</i>	Rodentia	0	0	0			
				<i>Condylura cristata</i>	Eulipotyphla	0	0	0			
				<i>Phyllostomus discolor</i>	Chiroptera	0	0	0			
				<i>Canis lupus familiaris</i>	Carnivora	0	0	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			
				<i>Bos taurus</i>	Artiodactyla	0	0	0			
				<i>Ornithorhynchus anatinus</i>	Monotremata	0	5	0			
				<i>Monodelphis domestica</i>	Didelphimorphia	0	0	0			
				<i>Phascolarctos cinereus</i>	Diprotodontia	1	0	0			
				<i>Orycteropus afer afer</i>	Tubulidentata	0	0	0			
				<i>Chrysochloris asiatica</i>	Afrosoricida	0	0	0			

<i>n</i>	Gene	Isoform	E	Presumably functional in	Order	PI			TA (s)		
						CR	SD	BF	CR	SD	BF
4	NPPA	NM_006172.4	3	<i>Loxodonta africana</i>	Proboscidea	0	0	0	42.46	43.46	47.04
				<i>Trichechus manatus latirostris</i>	Sirenia	0	0	0			
				<i>Pan troglodytes</i>	Primates	0	0	0			
				<i>Ochotona princeps</i>	Lagomorpha	0	0	0			
				<i>Mus musculus</i>	Rodentia	0	0	0			
				<i>Condylura cristata</i>	Eulipotyphla	0	0	0			
				<i>Phyllostomus discolor</i>	Chiroptera	0	0	0			
				<i>Canis lupus familiaris</i>	Carnivora	0	0	0			
				<i>Manis javanica</i>	Pholidota	0	0	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			
				<i>Bos taurus</i>	Artiodactyla	0	0	0			
5	APOA2	NM_001643.2	3	<i>Ornithorhynchus anatinus</i>	Monotremata	0	0	0	42.46	43.46	47.04
				<i>Monodelphis domestica</i>	Didelphimorphia	0	0	0			
				<i>Sarcophilus harrisii</i>	Dasyuromorpha	0	0	0			
				<i>Phascolarctos cinereus</i>	Diprotodontia	0	0	0			
				<i>Chrysochloris asiatica</i>	Afrosoricida	0	0	0			
				<i>Elephantulus edwardii</i>	Macroscelidea	0	0	0			
				<i>Loxodonta africana</i>	Proboscidea	0	0	0			
				<i>Trichechus manatus latirostris</i>	Sirenia	0	0	0			
				<i>Dasyus novemcinctus</i>	Cingulata	0	0	0			
				<i>Tupaia chinensis</i>	Scandentia	0	0	0			
				<i>Pan troglodytes</i>	Primates	0	0	0			
				<i>Ochotona princeps</i>	Lagomorpha	0	0	0			
				<i>Mus musculus</i>	Rodentia	0	0	0			
				<i>Condylura cristata</i>	Eulipotyphla	0	0	0			
				<i>Phyllostomus discolor</i>	Chiroptera	0	0	0			
				<i>Canis lupus familiaris</i>	Carnivora	0	0	0			
				<i>Manis javanica</i>	Pholidota	0	0	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			
				<i>Bos taurus</i>	Artiodactyla	0	0	0			
5	APOA2	NM_001643.2	3	<i>Ornithorhynchus anatinus</i>	Monotremata	5	0	0	25.92	22.36	26.85
				<i>Monodelphis domestica</i>	Didelphimorphia	5	4	4			
				<i>Sarcophilus harrisii</i>	Dasyuromorpha	0	0	0			
				<i>Phascolarctos cinereus</i>	Diprotodontia	5	0	0			
				<i>Orycteropus afer afer</i>	Tubulidentata	0	0	0			
				<i>Chrysochloris asiatica</i>	Afrosoricida	0	0	0			
				<i>Elephantulus edwardii</i>	Macroscelidea	0	0	0			
				<i>Loxodonta africana</i>	Proboscidea	0	0	0			
				<i>Trichechus manatus latirostris</i>	Sirenia	0	0	0			
				<i>Tupaia chinensis</i>	Scandentia	0	0	0			
				<i>Pan troglodytes</i>	Primates	0	0	0			
				<i>Ochotona princeps</i>	Lagomorpha	0	0	0			
				<i>Mus musculus</i>	Rodentia	0	0	0			
				<i>Condylura cristata</i>	Eulipotyphla	0	0	0			
				<i>Phyllostomus discolor</i>	Chiroptera	0	0	0			
				<i>Canis lupus familiaris</i>	Carnivora	0	0	0			
				<i>Manis javanica</i>	Pholidota	0	0	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			
				<i>Bos taurus</i>	Artiodactyla	0	0	0			

<i>n</i>	Gene	Isoform	E	Presumably functional in	Order	PI			TA (s)		
						CR	SD	BF	CR	SD	BF
6	KLHDC8A	NM_001271863.1	4	<i>Ornithorhynchus anatinus</i>	Monotremata	0	0	0	83.55	66.94	127.46
				<i>Monodelphis domestica</i>	Didelphimorphia	2	0	0			
				<i>Phascolarctos cinereus</i>	Diprotodontia	0	0	0			
				<i>Orycteropus afer afer</i>	Tubulidentata	0	0	0			
				<i>Chrysochloris asiatica</i>	Afrosoricida	0	0	0			
				<i>Elephantulus edwardii</i>	Macroscelidea	0	0	0			
				<i>Loxodonta africana</i>	Proboscidea	0	0	0			
				<i>Trichechus manatus latirostris</i>	Sirenia	0	0	0			
				<i>Dasyus novemcinctus</i>	Cingulata	0	0	0			
				<i>Pan troglodytes</i>	Primates	0	0	0			
				<i>Ochotona princeps</i>	Lagomorpha	0	0	0			
				<i>Mus musculus</i>	Rodentia	0	0	0			
				<i>Phyllostomus discolor</i>	Chiroptera	0	0	0			
				<i>Canis lupus familiaris</i>	Carnivora	0	0	0			
				<i>Manis javanica</i>	Pholidota	0	0	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			
				<i>Bos taurus</i>	Artiodactyla	0	0	0			
7	SFTPC	NM_003018.4	5	<i>Ornithorhynchus anatinus</i>	Monotremata	4	0	0	126.46	51.12	65.09
				<i>Sarcophilus harrisii</i>	Dasyuromorpha	5	3	0			
				<i>Phascolarctos cinereus</i>	Diprotodontia	3	3	0			
				<i>Orycteropus afer afer</i>	Tubulidentata	0	0	0			
				<i>Chrysochloris asiatica</i>	Afrosoricida	0	0	0			
				<i>Elephantulus edwardii</i>	Macroscelidea	0	0	0			
				<i>Loxodonta africana</i>	Proboscidea	0	0	0			
				<i>Trichechus manatus latirostris</i>	Sirenia	0	0	0			
				<i>Pan troglodytes</i>	Primates	0	0	0			
				<i>Ochotona princeps</i>	Lagomorpha	4	3	0			
				<i>Mus musculus</i>	Rodentia	0	0	0			
				<i>Condylura cristata</i>	Eulipotyphla	0	0	0			
				<i>Phyllostomus discolor</i>	Chiroptera	0	0	0			
				<i>Canis lupus familiaris</i>	Carnivora	3	3	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			
				<i>Bos taurus</i>	Artiodactyla	0	0	4			
8	FABP1	NM_001443.3	4	<i>Ornithorhynchus anatinus</i>	Monotremata	0	0	0	11.94	10.98	11.48
				<i>Orycteropus afer afer</i>	Tubulidentata	0	0	0			
				<i>Chrysochloris asiatica</i>	Afrosoricida	0	0	0			
				<i>Elephantulus edwardii</i>	Macroscelidea	0	0	0			
				<i>Loxodonta africana</i>	Proboscidea	0	0	0			
				<i>Trichechus manatus latirostris</i>	Sirenia	0	0	0			
				<i>Tupaia chinensis</i>	Scandentia	0	0	0			
				<i>Pan troglodytes</i>	Primates	0	0	0			
				<i>Ochotona princeps</i>	Lagomorpha	0	0	0			
				<i>Mus musculus</i>	Rodentia	0	0	0			
				<i>Condylura cristata</i>	Eulipotyphla	0	0	0			
				<i>Phyllostomus discolor</i>	Chiroptera	0	0	0			
				<i>Canis lupus familiaris</i>	Carnivora	0	0	0			
				<i>Manis javanica</i>	Pholidota	0	0	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			
				<i>Bos taurus</i>	Artiodactyla	0	0	0			

<i>n</i>	Gene	Isoform	E	Presumably functional in	Order	PI			TA (s)		
						CR	SD	BF	CR	SD	BF
9	MS4A1	NM_021950.3	6	<i>Ornithorhynchus anatinus</i>	Monotremata	5	0	0	418.57	278.95	279.32
				<i>Monodelphis domestica</i>	Didelphimorphia	5	2	0			
				<i>Sarcophilus harrisii</i>	Dasyuromorpha	5	5	0			
				<i>Phascolarctos cinereus</i>	Diprotodontia	5	0	0			
				<i>Orycteropus afer afer</i>	Tubulidentata	3	3	3			
				<i>Chrysochloris asiatica</i>	Afrosoricida	0	0	0			
				<i>Elephantulus edwardii</i>	Macroscelidea	0	0	0			
				<i>Loxodonta africana</i>	Proboscidea	0	0	0			
				<i>Trichechus manatus latirostris</i>	Sirenia	0	0	0			
				<i>Pan paniscus</i>	Primates	0	0	0			
				<i>Ochotona princeps</i>	Lagomorpha	0	0	0			
				<i>Mus musculus</i>	Rodentia	3	0	0			
				<i>Condylura cristata</i>	Eulipotyphla	0	0	0			
				<i>Phyllostomus discolor</i>	Chiroptera	0	0	0			
				<i>Canis lupus familiaris</i>	Carnivora	0	0	0			
				<i>Manis javanica</i>	Pholidota	0	0	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			
				<i>Bos taurus</i>	Artiodactyla	0	0	0			
10	KRT13	NM_153490.3	8	<i>Ornithorhynchus anatinus</i>	Monotremata	0	3	0	362.84	521.81	279.88
				<i>Monodelphis domestica</i>	Didelphimorphia	0	0	0			
				<i>Sarcophilus harrisii</i>	Dasyuromorpha	0	0	0			
				<i>Phascolarctos cinereus</i>	Diprotodontia	0	0	0			
				<i>Orycteropus afer afer</i>	Tubulidentata	0	0	0			
				<i>Chrysochloris asiatica</i>	Afrosoricida	0	0	0			
				<i>Elephantulus edwardii</i>	Macroscelidea	0	0	0			
				<i>Loxodonta africana</i>	Proboscidea	0	0	0			
				<i>Trichechus manatus latirostris</i>	Sirenia	0	0	0			
				<i>Dasypus novemcinctus</i>	Cingulata	0	0	0			
				<i>Tupaia chinensis</i>	Scandentia	0	0	0			
				<i>Pan troglodytes</i>	Primates	0	0	0			
				<i>Ochotona princeps</i>	Lagomorpha	0	0	0			
				<i>Mus musculus</i>	Rodentia	0	0	0			
				<i>Erinaceus europaeus</i>	Eulipotyphla	4	0	0			
				<i>Phyllostomus discolor</i>	Chiroptera	5	0	0			
				<i>Canis lupus familiaris</i>	Carnivora	0	0	0			
				<i>Manis javanica</i>	Pholidota	0	0	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			
				<i>Bos taurus</i>	Artiodactyla	0	0	0			
11	THEM5	NM_182578.4	6	<i>Ornithorhynchus anatinus</i>	Monotremata	3	2	2	479.05	439.97	607.44
				<i>Monodelphis domestica</i>	Didelphimorphia	2	2	0			
				<i>Phascolarctos cinereus</i>	Diprotodontia	5	2	2			
				<i>Orycteropus afer afer</i>	Tubulidentata	0	0	0			
				<i>Loxodonta africana</i>	Proboscidea	0	0	0			
				<i>Dasypus novemcinctus</i>	Cingulata	0	0	0			
				<i>Tupaia chinensis</i>	Scandentia	0	0	0			
				<i>Pan troglodytes</i>	Primates	0	0	0			
				<i>Oryctolagus cuniculus</i>	Lagomorpha	0	0	0			
				<i>Mus musculus</i>	Rodentia	0	0	0			
				<i>Condylura cristata</i>	Eulipotyphla	5	0	0			

<i>n</i>	Gene	Isoform	E	Presumably functional in	Order	PI			TA (s)		
						CR	SD	BF	CR	SD	BF
				<i>Phyllostomus discolor</i>	Chiroptera	0	0	0			
				<i>Canis lupus familiaris</i>	Carnivora	0	0	0			
				<i>Manis javanica</i>	Pholidota	0	0	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			
				<i>Bos taurus</i>	Artiodactyla	0	0	0			
12	TMPRSS15	NM_002772.3	25	<i>Ornithorhynchus anatinus</i>	Monotremata	5	5	5	2878.9	1657.0	1535.64
				<i>Monodelphis domestica</i>	Didelphimorphia	0	0	0			
				<i>Phascolarctos cinereus</i>	Diprotodontia	5	5	5			
				<i>Pan troglodytes</i>	Primates	0	0	0			
				<i>Mus musculus</i>	Rodentia	5	0	0			
				<i>Phyllostomus discolor</i>	Chiroptera	5	0	0			
				<i>Canis lupus familiaris</i>	Carnivora	0	0	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			
				<i>Bos taurus</i>	Artiodactyla	0	0	0			
13	PLIN1	NM_001145311.2	8	<i>Ornithorhynchus anatinus</i>	Monotremata	5	5	0	328.14	317.55	271.56
				<i>Monodelphis domestica</i>	Didelphimorphia	0	0	5			
				<i>Phascolarctos cinereus</i>	Diprotodontia	5	0	0			
				<i>Orycteropus afer afer</i>	Tubulidentata	0	0	0			
				<i>Chrysochloris asiatica</i>	Afrosoricida	0	0	0			
				<i>Elephantulus edwardii</i>	Macroscelidea	0	0	0			
				<i>Trichechus manatus latirostris</i>	Sirenia	0	0	0			
				<i>Pan troglodytes</i>	Primates	0	0	0			
				<i>Ochotona princeps</i>	Lagomorpha	0	0	0			
				<i>Mus musculus</i>	Rodentia	0	0	0			
				<i>Condylura cristata</i>	Eulipotyphla	0	0	0			
				<i>Phyllostomus discolor</i>	Chiroptera	0	0	0			
				<i>Canis lupus familiaris</i>	Carnivora	0	0	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			
				<i>Bos taurus</i>	Artiodactyla	0	0	0			
14	AKAP4	NM_003886.3	6	<i>Ornithorhynchus anatinus</i>	Monotremata	0	0	0	383.51	367.11	334.63
				<i>Chrysochloris asiatica</i>	Afrosoricida	0	0	0			
				<i>Trichechus manatus latirostris</i>	Sirenia	0	0	0			
				<i>Tupaia chinensis</i>	Scandentia	0	0	0			
				<i>Pan troglodytes</i>	Primates	0	0	0			
				<i>Mus musculus</i>	Rodentia	0	0	0			
				<i>Condylura cristata</i>	Eulipotyphla	0	0	0			
				<i>Phyllostomus discolor</i>	Chiroptera	0	0	0			
				<i>Canis lupus familiaris</i>	Carnivora	0	0	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			
				<i>Bos taurus</i>	Artiodactyla	0	0	0			
15	UPK2	NM_006760.4	5	<i>Ornithorhynchus anatinus</i>	Monotremata	5	5	0	20.14	20.95	22.04
				<i>Monodelphis domestica</i>	Didelphimorphia	0	3	0			
				<i>Orycteropus afer afer</i>	Tubulidentata	0	0	0			
				<i>Chrysochloris asiatica</i>	Afrosoricida	0	0	0			
				<i>Elephantulus edwardii</i>	Macroscelidea	0	0	0			
				<i>Loxodonta africana</i>	Proboscidea	0	0	0			
				<i>Trichechus manatus latirostris</i>	Sirenia	0	0	0			
				<i>Pan troglodytes</i>	Primates	0	0	0			
				<i>Ochotona princeps</i>	Lagomorpha	0	0	0			

n	Gene	Isoform	E	Presumably functional in	Order	PI			TA (s)		
						CR	SD	BF	CR	SD	BF
				<i>Mus musculus</i>	Rodentia	0	0	0			
				<i>Condylura cristata</i>	Eulipotyphla	0	0	0			
				<i>Phyllostomus discolor</i>	Chiroptera	0	0	0			
				<i>Canis lupus familiaris</i>	Carnivora	0	0	0			
				<i>Manis javanica</i>	Pholidota	0	0	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			
				<i>Bos taurus</i>	Artiodactyla	0	0	0			
16	TMPO	NM_003276.2	4	<i>Ornithorhynchus anatinus</i>	Monotremata	0	0	0	240.81	216.97	344.88
				<i>Monodelphis domestica</i>	Didelphimorphia	0	0	0			
				<i>Chrysochloris asiatica</i>	Afrosoricida	0	0	0			
				<i>Loxodonta africana</i>	Proboscidea	0	0	0			
				<i>Trichechus manatus latirostris</i>	Sirenia	0	0	0			
				<i>Pan troglodytes</i>	Primates	0	0	0			
				<i>Oryctolagus cuniculus</i>	Lagomorpha	5	5	5			
				<i>Mus musculus</i>	Rodentia	0	0	0			
				<i>Phyllostomus discolor</i>	Chiroptera	0	0	0			
				<i>Canis lupus familiaris</i>	Carnivora	0	0	0			
				<i>Manis javanica</i>	Pholidota	5	0	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			
				<i>Bos taurus</i>	Artiodactyla	0	0	0			
17	DDX47	NM_016355.4	12	<i>Ornithorhynchus anatinus</i>	Monotremata	0	0	0	119.59	118.06	152.23
				<i>Monodelphis domestica</i>	Didelphimorphia	0	0	0			
				<i>Phascolarctos cinereus</i>	Diprotodontia	0	0	0			
				<i>Orycteropus afer afer</i>	Tubulidentata	0	0	0			
				<i>Chrysochloris asiatica</i>	Afrosoricida	0	0	0			
				<i>Tupaia chinensis</i>	Scandentia	0	0	0			
				<i>Pan troglodytes</i>	Primates	0	0	0			
				<i>Ochotona princeps</i>	Lagomorpha	0	0	0			
				<i>Mus musculus</i>	Rodentia	0	0	0			
				<i>Condylura cristata</i>	Eulipotyphla	0	0	0			
				<i>Phyllostomus discolor</i>	Chiroptera	0	0	0			
				<i>Canis lupus familiaris</i>	Carnivora	0	0	0			
				<i>Manis javanica</i>	Pholidota	0	0	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			
				<i>Bos taurus</i>	Artiodactyla	0	0	0			
18	RPS19	NM_001321484.2	5	<i>Ornithorhynchus anatinus</i>	Monotremata	0	0	0	8.92	8.76	8.79
				<i>Monodelphis domestica</i>	Didelphimorphia	0	0	0			
				<i>Phascolarctos cinereus</i>	Diprotodontia	0	0	0			
				<i>Orycteropus afer afer</i>	Tubulidentata	0	0	0			
				<i>Chrysochloris asiatica</i>	Afrosoricida	0	0	0			
				<i>Elephantulus edwardii</i>	Macroscelidea	0	0	0			
				<i>Loxodonta africana</i>	Proboscidea	0	0	0			
				<i>Trichechus manatus latirostris</i>	Sirenia	0	0	0			
				<i>Dasypus novemcinctus</i>	Cingulata	0	0	0			
				<i>Tupaia chinensis</i>	Scandentia	0	0	0			
				<i>Pan troglodytes</i>	Primates	0	0	0			
				<i>Mus musculus</i>	Rodentia	0	0	0			
				<i>Sorex araneus</i>	Eulipotyphla	0	0	0			
				<i>Phyllostomus discolor</i>	Chiroptera	0	0	0			

<i>n</i>	Gene	Isoform	E	Presumably functional in	Order	PI			TA (s)		
						CR	SD	BF	CR	SD	BF
				<i>Canis lupus familiaris</i>	Carnivora	0	0	0			
				<i>Manis javanica</i>	Pholidota	0	0	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			
				<i>Bos taurus</i>	Artiodactyla	0	0	0			
19	SEPTIN9	NM_001113491.2	12	<i>Ornithorhynchus anatinus</i>	Monotremata	5	0	0	108.65	111.5	123.5
				<i>Phascolarctos cinereus</i>	Diprotodontia	0	0	0			
				<i>Pan troglodytes</i>	Primates	0	0	0			
				<i>Mus musculus</i>	Rodentia	0	0	0			
				<i>Phyllostomus discolor</i>	Chiroptera	0	0	0			
				<i>Manis javanica</i>	Pholidota	0	0	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			
				<i>Bos taurus</i>	Artiodactyla	0	0	0			
20	DCTN2	NM_001261413.1	14	<i>Ornithorhynchus anatinus</i>	Monotremata	5	0	0	110.18	108.06	110.45
				<i>Monodelphis domestica</i>	Didelphimorphia	0	0	0			
				<i>Sarcophilus harrisii</i>	Dasyuromorpha	0	5	0			
				<i>Phascolarctos cinereus</i>	Diprotodontia	0	0	0			
				<i>Orycteropus afer afer</i>	Tubulidentata	0	0	0			
				<i>Chrysochloris asiatica</i>	Afrosoricida	0	0	0			
				<i>Elephantulus edwardii</i>	Macroscelidea	0	0	0			
				<i>Loxodonta africana</i>	Proboscidea	0	0	0			
				<i>Trichechus manatus latirostris</i>	Sirenia	0	0	0			
				<i>Tupaia chinensis</i>	Scandentia	0	0	0			
				<i>Pan troglodytes</i>	Primates	0	0	0			
				<i>Ochotona princeps</i>	Lagomorpha	0	0	0			
				<i>Mus musculus</i>	Rodentia	0	0	0			
				<i>Condylura cristata</i>	Eulipotyphla	0	0	0			
				<i>Phyllostomus discolor</i>	Chiroptera	0	0	0			
				<i>Canis lupus familiaris</i>	Carnivora	0	0	0			
				<i>Manis javanica</i>	Pholidota	0	0	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			
				<i>Bos taurus</i>	Artiodactyla	0	0	0			
21	RPL23	NM_000978.4	5	<i>Ornithorhynchus anatinus</i>	Monotremata	0	0	0	12.11	12.34	13.72
				<i>Monodelphis domestica</i>	Didelphimorphia	0	0	0			
				<i>Phascolarctos cinereus</i>	Diprotodontia	0	0	0			
				<i>Orycteropus afer afer</i>	Tubulidentata	0	0	0			
				<i>Chrysochloris asiatica</i>	Afrosoricida	0	0	0			
				<i>Elephantulus edwardii</i>	Macroscelidea	0	0	0			
				<i>Loxodonta africana</i>	Proboscidea	0	0	0			
				<i>Trichechus manatus latirostris</i>	Sirenia	0	0	0			
				<i>Dasypus novemcinctus</i>	Cingulata	0	0	0			
				<i>Tupaia chinensis</i>	Scandentia	5	0	0			
				<i>Pan troglodytes</i>	Primates	0	0	0			
				<i>Ochotona princeps</i>	Lagomorpha	0	0	0			
				<i>Mus musculus</i>	Rodentia	0	0	0			
				<i>Condylura cristata</i>	Eulipotyphla	0	0	0			
				<i>Phyllostomus discolor</i>	Chiroptera	0	0	0			
				<i>Canis lupus familiaris</i>	Carnivora	0	0	0			
				<i>Manis javanica</i>	Pholidota	0	0	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			

<i>n</i>	Gene	Isoform	E	Presumably functional in	Order	PI			TA (s)		
						CR	SD	BF	CR	SD	BF
				<i>Bos taurus</i>	Artiodactyla	0	0	0			
22	VIM	NM_003380.5	9	<i>Ornithorhynchus anatinus</i>	Monotremata	0	0	0	147.28	138.09	141.84
				<i>Monodelphis domestica</i>	Didelphimorphia	0	0	0			
				<i>Phascolarctos cinereus</i>	Diprotodontia	0	0	0			
				<i>Chrysochloris asiatica</i>	Afrosoricida	0	0	0			
				<i>Elephantulus edwardii</i>	Macroscelidea	0	0	0			
				<i>Loxodonta africana</i>	Proboscidea	0	0	0			
				<i>Trichechus manatus latirostris</i>	Sirenia	0	0	0			
				<i>Tupaia chinensis</i>	Scandentia	0	0	0			
				<i>Pan troglodytes</i>	Primates	0	0	0			
				<i>Ochotona princeps</i>	Lagomorpha	0	0	0			
				<i>Mus musculus</i>	Rodentia	0	0	0			
				<i>Condylura cristata</i>	Eulipotyphla	0	0	0			
				<i>Phyllostomus discolor</i>	Chiroptera	0	0	0			
				<i>Felis catus</i>	Carnivora	0	0	0			
				<i>Manis javanica</i>	Pholidota	0	0	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			
				<i>Bos taurus</i>	Artiodactyla	0	0	0			
23	TUBB4B	NM_006088.6	4	<i>Ornithorhynchus anatinus</i>	Monotremata	0	0	0	37.92	38.98	40.04
				<i>Monodelphis domestica</i>	Didelphimorphia	0	0	0			
				<i>Sarcophilus harrisii</i>	Dasyuromorpha	0	0	0			
				<i>Phascolarctos cinereus</i>	Diprotodontia	0	0	0			
				<i>Chrysochloris asiatica</i>	Afrosoricida	0	0	0			
				<i>Trichechus manatus latirostris</i>	Sirenia	0	0	0			
				<i>Dasyus novemcinctus</i>	Cingulata	0	0	0			
				<i>Tupaia chinensis</i>	Scandentia	0	0	0			
				<i>Pan troglodytes</i>	Primates	0	0	0			
				<i>Mus musculus</i>	Rodentia	0	0	0			
				<i>Phyllostomus discolor</i>	Chiroptera	0	0	0			
				<i>Canis lupus familiaris</i>	Carnivora	0	0	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			
				<i>Bos taurus</i>	Artiodactyla	0	0	0			
24	ATP5F1B	NM_001686.4	10	<i>Ornithorhynchus anatinus</i>	Monotremata	0	0	0	141.72	146.7	154.21
				<i>Monodelphis domestica</i>	Didelphimorphia	0	0	0			
				<i>Sarcophilus harrisii</i>	Dasyuromorpha	5	0	0			
				<i>Phascolarctos cinereus</i>	Diprotodontia	0	0	0			
				<i>Orycteropus afer afer</i>	Tubulidentata	0	0	0			
				<i>Chrysochloris asiatica</i>	Afrosoricida	0	0	0			
				<i>Elephantulus edwardii</i>	Macroscelidea	0	0	0			
				<i>Loxodonta africana</i>	Proboscidea	0	0	0			
				<i>Dasyus novemcinctus</i>	Cingulata	0	0	0			
				<i>Pan troglodytes</i>	Primates	0	0	0			
				<i>Ochotona princeps</i>	Lagomorpha	5	0	0			
				<i>Mus musculus</i>	Rodentia	0	0	0			
				<i>Condylura cristata</i>	Eulipotyphla	0	0	0			
				<i>Myotis davidii</i>	Chiroptera	0	0	0			
				<i>Canis lupus familiaris</i>	Carnivora	5	0	0			
				<i>Manis javanica</i>	Pholidota	0	0	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			

<i>n</i>	Gene	Isoform	E	Presumably functional in	Order	PI			TA (s)		
						CR	SD	BF	CR	SD	BF
				<i>Bos taurus</i>	Artiodactyla	0	0	0			
25	CALR	NM_004343.3	9	<i>Ornithorhynchus anatinus</i>	Monotremata	5	3	3	172.35	132.54	141.27
				<i>Monodelphis domestica</i>	Didelphimorphia	5	0	0			
				<i>Phascolarctos cinereus</i>	Diprotodontia	5	0	0			
				<i>Orycteropus afer afer</i>	Tubulidentata	0	0	0			
				<i>Chrysochloris asiatica</i>	Afrosoricida	0	0	0			
				<i>Loxodonta africana</i>	Proboscidea	0	0	0			
				<i>Trichechus manatus latirostris</i>	Sirenia	0	0	0			
				<i>Dasytus novemcinctus</i>	Cingulata	0	0	0			
				<i>Tupaia chinensis</i>	Scandentia	0	0	0			
				<i>Pan troglodytes</i>	Primates	0	0	0			
				<i>Ochotona princeps</i>	Lagomorpha	0	0	0			
				<i>Mus musculus</i>	Rodentia	0	0	0			
				<i>Condylura cristata</i>	Eulipotyphla	0	0	0			
				<i>Phyllostomus discolor</i>	Chiroptera	0	0	0			
				<i>Canis lupus familiaris</i>	Carnivora	0	0	0			
				<i>Manis javanica</i>	Pholidota	5	0	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			
				<i>Bos taurus</i>	Artiodactyla	0	0	0			
26	RER1	NM_007033.5	6	<i>Ornithorhynchus anatinus</i>	Monotremata	0	0	0	28.23	29.36	29.48
				<i>Monodelphis domestica</i>	Didelphimorphia	0	0	0			
				<i>Phascolarctos cinereus</i>	Diprotodontia	0	0	0			
				<i>Orycteropus afer afer</i>	Tubulidentata	0	0	0			
				<i>Chrysochloris asiatica</i>	Afrosoricida	0	0	0			
				<i>Loxodonta africana</i>	Proboscidea	0	0	0			
				<i>Dasytus novemcinctus</i>	Cingulata	0	0	0			
				<i>Tupaia chinensis</i>	Scandentia	0	0	0			
				<i>Pan troglodytes</i>	Primates	0	0	0			
				<i>Ochotona princeps</i>	Lagomorpha	0	0	0			
				<i>Mus musculus</i>	Rodentia	0	0	0			
				<i>Condylura cristata</i>	Eulipotyphla	0	0	0			
				<i>Phyllostomus discolor</i>	Chiroptera	0	0	0			
				<i>Canis lupus familiaris</i>	Carnivora	0	0	0			
				<i>Manis javanica</i>	Pholidota	0	0	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			
				<i>Bos taurus</i>	Artiodactyla	0	0	0			
27	AP2M1	NM_001025205.1	10	<i>Ornithorhynchus anatinus</i>	Monotremata	0	0	0	56.25	59.05	57.18
				<i>Phascolarctos cinereus</i>	Diprotodontia	0	0	0			
				<i>Orycteropus afer afer</i>	Tubulidentata	0	0	0			
				<i>Chrysochloris asiatica</i>	Afrosoricida	0	0	0			
				<i>Elephantulus edwardii</i>	Macroscelidea	0	0	0			
				<i>Loxodonta africana</i>	Proboscidea	0	0	0			
				<i>Trichechus manatus latirostris</i>	Sirenia	0	0	0			
				<i>Tupaia chinensis</i>	Scandentia	0	0	0			
				<i>Pan troglodytes</i>	Primates	0	0	0			
				<i>Ochotona princeps</i>	Lagomorpha	0	0	0			
				<i>Mus musculus</i>	Rodentia	0	0	0			
				<i>Condylura cristata</i>	Eulipotyphla	0	0	0			
				<i>Phyllostomus discolor</i>	Chiroptera	0	0	0			

n	Gene	Isoform	E	Presumably functional in	Order	PI			TA (s)		
						CR	SD	BF	CR	SD	BF
				<i>Felis catus</i>	Carnivora	0	0	0			
				<i>Manis javanica</i>	Pholidota	0	0	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			
				<i>Bos taurus</i>	Artiodactyla	0	0	0			
28	PSAP	NM_002778.4	14	<i>Ornithorhynchus anatinus</i>	Monotremata	5	5	5	273.69	289.63	353.46
				<i>Phascolarctos cinereus</i>	Diprotodontia	5	3	3			
				<i>Orycteropus afer afer</i>	Tubulidentata	0	0	0			
				<i>Elephantulus edwardii</i>	Macroscelidea	3	5	5			
				<i>Loxodonta africana</i>	Proboscidea	0	0	0			
				<i>Trichechus manatus latirostris</i>	Sirenia	0	0	0			
				<i>Tupaia chinensis</i>	Scandentia	0	0	0			
				<i>Pan troglodytes</i>	Primates	0	0	0			
				<i>Ochotona princeps</i>	Lagomorpha	5	5	5			
				<i>Mus musculus</i>	Rodentia	4	3	3			
				<i>Phyllostomus discolor</i>	Chiroptera	0	0	0			
				<i>Canis lupus familiaris</i>	Carnivora	0	0	0			
				<i>Manis javanica</i>	Pholidota	0	0	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			
				<i>Bos taurus</i>	Artiodactyla	0	0	0			
29	EMC4	NM_016454.4	5	<i>Ornithorhynchus anatinus</i>	Monotremata	0	0	0	12.60	12.93	13.42
				<i>Monodelphis domestica</i>	Didelphimorphia	0	0	0			
				<i>Phascolarctos cinereus</i>	Diprotodontia	0	0	0			
				<i>Chrysochloris asiatica</i>	Afrosoricida	0	0	0			
				<i>Elephantulus edwardii</i>	Macroscelidea	0	0	0			
				<i>Loxodonta africana</i>	Proboscidea	0	0	0			
				<i>Trichechus manatus latirostris</i>	Sirenia	0	0	0			
				<i>Dasypus novemcinctus</i>	Cingulata	0	0	0			
				<i>Pan troglodytes</i>	Primates	0	0	0			
				<i>Oryctolagus cuniculus</i>	Lagomorpha	0	0	0			
				<i>Mus musculus</i>	Rodentia	0	0	0			
				<i>Condylura cristata</i>	Eulipotyphla	0	0	0			
				<i>Phyllostomus discolor</i>	Chiroptera	0	0	0			
				<i>Canis lupus familiaris</i>	Carnivora	0	0	0			
				<i>Manis javanica</i>	Pholidota	0	0	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			
				<i>Bos taurus</i>	Artiodactyla	0	0	0			
30	CD63	NM_001780.5	10	<i>Ornithorhynchus anatinus</i>	Monotremata	0	0	0	36.52	36.08	37.56
				<i>Sarcophilus harrisii</i>	Dasyuromorpha	0	0	0			
				<i>Phascolarctos cinereus</i>	Diprotodontia	0	0	0			
				<i>Orycteropus afer afer</i>	Tubulidentata	0	0	0			
				<i>Chrysochloris asiatica</i>	Afrosoricida	0	0	0			
				<i>Elephantulus edwardii</i>	Macroscelidea	0	0	0			
				<i>Loxodonta africana</i>	Proboscidea	0	0	0			
				<i>Dasypus novemcinctus</i>	Cingulata	0	0	0			
				<i>Pan troglodytes</i>	Primates	0	0	0			
				<i>Oryctolagus cuniculus</i>	Lagomorpha	0	0	0			
				<i>Mus musculus</i>	Rodentia	0	0	0			
				<i>Phyllostomus discolor</i>	Chiroptera	0	0	0			
				<i>Felis catus</i>	Carnivora	0	0	0			

<i>n</i>	Gene	Isoform	E	Presumably functional in	Order	PI			TA (s)		
						CR	SD	BF	CR	SD	BF
				<i>Manis javanica</i>	Pholidota	0	0	0			
				<i>Equus caballus</i>	Perissodactyla	0	0	0			
				<i>Bos taurus</i>	Artiodactyla	0	0	0			

Appendix B

Raw data associated with Pseudo*Checker*'s performance evaluation over gene loss events

This appendix contains the raw data associated with the 90 analysis conducted during Pseudo*Checker*'s performance evaluation over previously known cases of gene loss events. It comprehends a main table that, besides other accessory information, includes, per tested gene, per target species, and per tested similarity scoring scheme underlying the coding sequence prediction step of Pseudo*Checker*'s pipeline, the corresponding assigned PseudoIndex. The time of execution regarding each of the 90 analysis is also disclosed.

Table B.1: Raw data underlying the 90 analysis conducted during PseudoChecker’s performance evaluation over previously known cases of gene loss events. This table comprehends per analysed gene (2nd column), the correspondent NCBI’s accession number of the retrieved human ortholog isoform (3rd column), the correspondent number of coding exons (4th column), the affected lineages by its loss and respective reference underlying the matching analysed study (5th and 6th columns, respectively), the PseudoIndex (PI) assigned for each affected species according to the three tested similarity scoring schemes (the one optimised for closely related species (CR), the one optimised for slightly divergent species (SD) and the best-fit similarity scoring scheme (BF)), and the corresponding time of analysis’ execution (TA), in seconds (last three columns, each per distinct tested similarity scoring scheme).

<i>n</i>	Gene	Isoform	E	Lost in	Ref.	PI			TA (s)		
						CR	SD	BF	CR	SD	BF
1	IL20	NM_018724.3	5	<i>Orcinus orca</i>	[108]	5	5	5	6.69	7.87	11.99
				<i>Tursiops truncatus</i>		5	5	5			
				<i>Physeter catodon</i>		0	0	0			
				<i>Delphinapterus leucas</i>		5	5	5			
				<i>Lipotes vexillifer</i>		5	5	5			
				<i>Balaenoptera acutorostrata scammoni</i>		5	5	5			
				<i>Nannospalax galili</i>		5	5	5			
2	AANAT	NM_001166579.1	4	<i>Lipotes vexillifer</i>	[111]	3	5	5	27.50	35.61	41.68
				<i>Balaenoptera acutorostrata scammoni</i>		3	5	5			
				<i>Orcinus orca</i>		3	5	5			
				<i>Lagenorhynchus obliquidens</i>		3	5	5			
				<i>Delphinapterus leucas</i>		5	5	5			
				<i>Neophocaena asiaeorientalis asiaeorientalis</i>		5	5	5			
				<i>Physeter catodon</i>		5	5	5			
3	ASMT	NM_001171038.2	9	<i>Lagenorhynchus obliquidens</i>	[111]	5	5	5	122.91	125.55	133.72
				<i>Delphinapterus leucas</i>		5	5	5			
				<i>Neophocaena asiaeorientalis asiaeorientalis</i>		5	5	5			
4	MTNR1B	NM_005959.3	2	<i>Balaenoptera acutorostrata scammoni</i>	[111]	5	5	5	13.66	13.20	24.62
				<i>Delphinapterus leucas</i>		5	5	5			
				<i>Physeter catodon</i>		5	5	5			
5	S100Z	NM_130772.4	2	<i>Odocoileus virginianus texanus</i>	[67]	5	5	5	10.46	10.47	19.44
				<i>Aotus nancymae</i>		5	5	5			
				<i>Canis lupus familiaris</i>		4	4	4			
				<i>Tursiops truncatus</i>		5	5	5			
				<i>Orcinus orca</i>		5	5	5			
				<i>Balaenoptera acutorostrata scammoni</i>		5	5	5			
				<i>Delphinapterus leucas</i>		5	5	5			
				<i>Pteropus alecto</i>		5	5	5			
				<i>Myotis davidii</i>		5	5	5			
				<i>Capra hircus</i>		5	5	5			
				<i>Neomonachus schauinslandi</i>		5	1	5			
				<i>Desmodus rotundus</i>		5	5	5			
				<i>Lipotes vexillifer</i>		5	5	5			
				<i>Myotis lucifugus</i>		5	5	5			

n	Gene	Isoform	E	Lost in	Ref.	PI			TA (s)		
						CR	SD	BF	CR	SD	BF
6	SYCN	NM_001080468.4	1	<i>Myotis brandtii</i>	[67]	5	5	5	1.72	1.88	10.18
				<i>Rousettus aegyptiacus</i>		5	5	5			
				<i>Heterocephalus glaber</i>		5	5	5			
				<i>Fukomys damarensis</i>		5	5	5			
7	PNLIPRP1	NM_006229.4	12	<i>Ceratotherium simum simum</i>	[67]	2	2	2	119.73	115.11	147.69
				<i>Heterocephalus glaber</i>		5	5	5			
				<i>Bos taurus</i>		5	5	5			
				<i>Bison bison bison</i>		5	5	5			
				<i>Bubalus bubalis</i>		5	5	5			
				<i>Ovis aries</i>		5	5	5			
				<i>Capra hircus</i>		5	5	5			
				<i>Loxodonta africana</i>		5	5	5			
				<i>Cavia porcellus</i>		5	5	5			
				<i>Bubalus bubalis</i>		5	5	5			
8	RFXP4	NM_181885.2	1	<i>Tursiops truncatus</i>	[67]	5	5	5	93.37	95.92	149.05
				<i>Ochotona princeps</i>		5	5	5			
				<i>Orcinus orca</i>		5	5	5			
				<i>Delphinapterus leucas</i>		5	5	5			
				<i>Lipotes vexillifer</i>		5	5	5			
				<i>Physeter catodon</i>		5	5	5			
				<i>Balaenoptera acutorostrata scammoni</i>		5	5	5			
				<i>Acinonyx jubatus</i>		3	3	3			
				<i>Eptesicus fuscus</i>		5	5	5			
				<i>Myotis davidii</i>		5	5	5			
				<i>Myotis brandtii</i>		5	5	5			
				<i>Myotis lucifugus</i>		5	5	5			
9	INSL5	NM_005478.4	2	<i>Myotis brandtii</i>	[67]	5	5	5	37.81	37.63	50.05
				<i>Odobenus rosmarus divergens</i>		5	5	5			
				<i>Mustela putorius furo</i>		5	5	5			
				<i>Enhydra lutris kenyon</i>		5	5	5			
				<i>Panthera tigris altaica</i>		5	5	5			
				<i>Panthera pardus</i>		5	5	5			
				<i>Leptonychotes weddellii</i>		5	5	5			
				<i>Neomonachus schauinslandi</i>		5	5	5			
				<i>Balaenoptera acutorostrata scammoni</i>		5	0	0			
				<i>Felis catus</i>		5	5	5			
				<i>Acinonyx jubatus</i>		5	5	5			
				<i>Eptesicus fuscus</i>		5	5	5			
10	NOX1	NM_007052.5	13	<i>Tursiops truncatus</i>	[67]	5	5	5	61.40	47.80	70.92
				<i>Delphinapterus leucas</i>		5	5	5			
				<i>Lipotes vexillifer</i>		5	5	5			
				<i>Manis javanica</i>		5	5	5			
11	KLK8	NM_144505.3	5	<i>Tursiops truncatus</i>	[66]	5	5	5	21.96	21.53	39.04
				<i>Physeter catodon</i>		5	5	5			
				<i>Trichechus manatus latirostris</i>		5	5	5			
12	ODAM	NM_017855.3	10	<i>Physeter catodon</i>	[180]	0	0	0	29.74	30.17	37.18
				<i>Orcinus orca</i>		5	5	3			
				<i>Balaenoptera acutorostrata scammoni</i>		5	5	5			
				<i>Callithrix jacchus</i>		5	5	5			

<i>n</i>	Gene	Isoform	E	Lost in	Ref.	PI			TA (s)		
						CR	SD	BF	CR	SD	BF
				<i>Lipotes vexillifer</i>		5	5	5			
				<i>Tursiops truncatus</i>		5	5	5			
				<i>Delphinapterus leucas</i>		5	5	5			
				<i>Neophocaena asiaeorientalis asiaeorientalis</i>		5	5	5			
				<i>Neomonachus schauinslandi</i>		5	5	5			
				<i>Dasyus novemcinctus</i>		5	5	5			
13	MX1	NM_001144925.2	13	<i>Tursiops truncatus</i>	[18]	5	5	5	128.37	221.65	225.44
				<i>Orcinus orca</i>		5	5	5			
				<i>Physeter catodon</i>		5	5	5			
14	MX2	NM_002463.2	13	<i>Tursiops truncatus</i>	[18]	5	5	5	124.01	149.97	168.65
				<i>Orcinus orca</i>		5	5	5			
				<i>Neophocaena asiaeorientalis asiaeorientalis</i>		5	5	5			
15	ALOXE3	NM_001165960.1	16	<i>Physeter catodon</i>	[172]	5	5	5	211.31	240.11	214.80
				<i>Tursiops truncatus</i>		5	5	5			
				<i>Orcinus orca</i>		5	5	5			
				<i>Balaenoptera acutorostrata scammoni</i>		5	5	5			
16	MOGAT3	NM_178176.4	7	<i>Orcinus orca</i>	[110]	5	5	5	13.07	13.02	21.61
				<i>Balaenoptera acutorostrata scammoni</i>		5	5	5			
				<i>Delphinapterus leucas</i>		5	5	5			
17	ELOVL3	NM_152310.3	4	<i>Orcinus orca</i>	[110]	5	5	5	6.98	7.68	11.01
				<i>Tursiops truncatus</i>		5	5	5			
				<i>Delphinapterus leucas</i>		5	5	5			
				<i>Physeter catodon</i>		5	5	5			
				<i>Lipotes vexillifer</i>		5	5	5			
18	CCL27	NM_006664.4	3	<i>Heterocephalus glaber</i>	[109]	5	5	5	5.03	6.04	6.84
				<i>Hipposideros armiger</i>		5	5	5			
				<i>Manis javanica</i>		5	5	5			
				<i>Balaenoptera acutorostrata scammoni</i>		5	5	5			
				<i>Lipotes vexillifer</i>		5	5	5			
				<i>Physeter catodon</i>		3	3	3			
				<i>Lagenorhynchus obliquidens</i>		5	5	5			
				<i>Tursiops truncatus</i>		5	5	5			
				<i>Delphinapterus leucas</i>		5	5	5			
19	KRT24	NM_019016.3	8	<i>Odobenus rosmarus divergens</i>	[46] [47]	5	5	5	48.57	52.53	60.31
				<i>Camelus bactrianus</i>		5	5	5			
				<i>Camelus dromedarius</i>		5	5	5			
				<i>Callorhinus ursinus</i>		5	5	5			
				<i>Zalophus californianus</i>		5	5	5			
				<i>Tursiops truncatus</i>		5	5	5			
				<i>Physeter catodon</i>		5	5	5			
				<i>Balaenoptera acutorostrata scammoni</i>		5	5	5			
				<i>Orcinus orca</i>		5	5	5			
20	KRT84	NM_033045.4	9	<i>Tursiops truncatus</i>	[46]	5	5	2	79.58	36.88	45.66
				<i>Physeter catodon</i>		5	5	5			
				<i>Lipotes vexillifer</i>		5	5	5			
21	PON1	NM_000446.7	9	<i>Physeter catodon</i>	[127]	5	5	5	78.15	76.44	90.62
				<i>Leptonychotes weddellii</i>		5	5	5			
				<i>Neomonachus schauinslandi</i>		5	5	5			
				<i>Trichechus manatus latirostris</i>		5	5	5			

<i>n</i>	Gene	Isoform	E	Lost in	Ref.	PI			TA (s)		
						CR	SD	BF	CR	SD	BF
				<i>Balaenoptera acutorostrata scammoni</i>		5	5	5			
				<i>Lipotes vexillifer</i>		5	5	5			
				<i>Tursiops truncatus</i>		5	5	5			
22	<i>ACP4</i>	NM_033068.3	11	<i>Orycteropus afer afer</i>	[172]	5	5	5	32.45	28.26	30.34
				<i>Dasypus novemcinctus</i>		5	5	5			
				<i>Balaenoptera acutorostrata scammoni</i>		5	5	5			
23	<i>DDB2</i>	NM_000107.2	10	<i>Manis javanica</i>	[172]	5	5	5	34.79	39.46	39.8
				<i>Dasypus novemcinctus</i>		5	5	5			
24	<i>MMP12</i>	NM_002426.6	10	<i>Trichechus manatus latirostris</i>	[172]	5	2	2	21.40	20.93	26.50
				<i>Balaenoptera acutorostrata scammoni</i>		5	5	5			
25	<i>DSC1</i>	NM_024421.2	16	<i>Physeter catodon</i>	[172]	5	5	5	263.83	241.37	247.17
				<i>Balaenoptera acutorostrata scammoni</i>		5	5	5			
				<i>Orcinus orca</i>		5	5	5			
26	<i>MOGAT2</i>	NM_025098.4	6	<i>Pteropus alecto</i>	[172]	5	5	5	6.71	5.73	10.25
				<i>Pteropus vampyrus</i>		5	5	5			
27	<i>FABP6</i>	NM_001040442.1	6	<i>Pteropus alecto</i>	[172]	5	5	5	13.57	9.36	12.07
				<i>Pteropus vampyrus</i>		5	5	5			
28	<i>AQP6</i>	NM_001652.4	4	<i>Pteropus alecto</i>	[172]	5	5	5	4.97	5.38	27.54
				<i>Pteropus vampyrus</i>		5	5	5			
29	<i>FFAR3</i>	NM_005304.5	1	<i>Pteropus alecto</i>	[172]	5	5	5	7.62	7.83	8.84
				<i>Pteropus vampyrus</i>		5	5	5			
30	<i>APMAP</i>	NM_001507.1	2	<i>Mus musculus</i>	[172]	5	5	5	13.21	8.35	14.61
				<i>Rattus norvegicus</i>		5	5	5			

Bibliography

- [1] National human genome research institute.
- [2] Federico Abascal, Rafael Zardoya, and Maximilian J Telford. Translatorex: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic acids research*, 38(suppl_2):W7–W13, 2010.
- [3] Sophie Abby and Vincent Daubin. Comparative genomics and the evolution of prokaryotes. *Trends in microbiology*, 15(3):135–141, 2007.
- [4] Khelifi Adel, Duret Laurent, and Mouchiroud Dominique. Hoppsigen: a database of human and mouse processed pseudogenes. *Nucleic acids research*, 33(suppl_1):D59–D66, 2005.
- [5] Ricard Albalat and Cristian Cañestro. Evolution by gene loss. *Nature Reviews Genetics*, 17(7):379, 2016.
- [6] Jessica Alföldi and Kerstin Lindblad-Toh. Comparative genomics as a tool to understand evolution and disease. *Genome research*, 23(7):1063–1068, 2013.
- [7] Joel Armstrong, Ian T Fiddes, Mark Diekhans, and Benedict Paten. Whole-genome alignment and comparative annotation. *Annual review of animal biosciences*, 7:41–64, 2019.
- [8] Ana B Asenjo, Jeanne Rim, and Daniel D Oprian. Molecular determinants of human red/green color discrimination. *Neuron*, 12(5):1131–1138, 1994.
- [9] Oswald T Avery, Colin M MacLeod, and Maclyn McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. *Journal of experimental medicine*, 79(2):137–158, 1944.
- [10] Robert Baertsch, Mark Diekhans, W James Kent, David Haussler, and Jürgen Brosius. Retrocopy contributions to the evolution of the human genome. *BMC genomics*, 9(1):466, 2008.
- [11] Evgeniy S Balakirev and Francisco J Ayala. Pseudogenes: are they “junk” or functional dna? *Annual review of genetics*, 37(1):123–151, 2003.

- [12] Suganthi Balasubramanian, Lukas Habegger, Adam Frankish, Daniel G MacArthur, Rachel Harte, Chris Tyler-Smith, Jennifer Harrow, and Mark Gerstein. Gene inactivation and its implications for annotation in the era of personal genomics. *Genes & development*, 25(1): 1–10, 2011.
- [13] Carl Barton, Tomáš Flouri, Costas S Iliopoulos, and Solon P Pissis. Global and local sequence alignment with a bounded number of gaps. *Theoretical Computer Science*, 582: 1–16, 2015.
- [14] Serafim Batzoglou. The many faces of sequence alignment. *Briefings in bioinformatics*, 6(1):6–22, 2005.
- [15] Olaf RP Bininda-Emonds. transalign: using amino acids to facilitate the multiple alignment of protein-coding dna sequences. *BMC bioinformatics*, 6(1):156, 2005.
- [16] Guillaume Blanc and Kenneth H Wolfe. Functional divergence of duplicated genes formed by polyploidy during arabidopsis evolution. *The Plant Cell*, 16(7):1679–1691, 2004.
- [17] Tine Blomme, Klaas Vandepoele, Stefanie De Bodt, Cedric Simillion, Steven Maere, and Yves Van de Peer. The gain and loss of genes during 600 million years of vertebrate evolution. *Genome biology*, 7(5):R43, 2006.
- [18] Benjamin A Braun, Amir Marcovitz, J Gray Camp, Robin Jia, and Gill Bejerano. Mx1 and mx2 key antiviral proteins are surprisingly lost in toothed whales. *Proceedings of the National Academy of Sciences*, 112(26):8036–8040, 2015.
- [19] Michael R Brent and Roderic Guigo. Recent advances in gene structure prediction. *Current opinion in structural biology*, 14(3):264–272, 2004.
- [20] Calvin B Bridges. The bar" gene" a duplication. *Science*, 83(2148):210–211, 1936.
- [21] Connor J Burgin, Jocelyn P Colella, Philip L Kahn, and Nathan S Upham. How many species of mammals are there? *Journal of Mammalogy*, 99(1):1–14, 2018.
- [22] M Buset, IA Seledtsov, and VV Solovyev. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic acids research*, 28(21):4364–4375, 2000.
- [23] Reed A Cartwright. Logarithmic gap costs decrease alignment accuracy. *BMC bioinformatics*, 7(1):527, 2006.
- [24] Lindsay Chaney, Aaron R Sharp, Carrie R Evans, and Joshua A Udall. Genome mapping in plant comparative genomics. *Trends in plant science*, 21(9):770–780, 2016.
- [25] Brad A Chapman, John E Bowers, Frank A Feltus, and Andrew H Paterson. Buffering of crucial functions by paleologous duplicated genes may contribute cyclicity to angiosperm genome duplication. *Proceedings of the National Academy of Sciences*, 103(8):2730–2735, 2006.

- [26] Eric CH Chen, Carlos Fernando Buen Abad Najar, Chunfang Zheng, Alex Brandts, Eric Lyons, Haibao Tang, Lorenzo Carretero-Paulet, Victor A Albert, and David Sankoff. The dynamics of functional classes of plant genes in rediploidized ancient polyploids. In *BMC bioinformatics*, volume 14, page S19. BioMed Central, 2013.
- [27] Hsun-Hua Chou, Toshiyuki Hayakawa, Sandra Diaz, Matthias Krings, Etty Indriati, Meave Leakey, Svante Paabo, Yoko Satta, Naoyuki Takahata, and Ajit Varki. Inactivation of cmp-n-acetylneuraminic acid hydroxylase occurred prior to brain expansion during human evolution. *Proceedings of the National Academy of Sciences*, 99(18):11736–11741, 2002.
- [28] S Clancy and W Brown. Translation: Dna to mrna to. 2008.
- [29] Suzanne Clancy. Rna splicing: introns, exons and spliceosome. *Nature Education*, 1(1):31, 2008.
- [30] Paul F Cliften, Robert S Fulton, Richard K Wilson, and Mark Johnston. After the duplication: gene loss and adaptation in saccharomyces genomes. *Genetics*, 172(2):863–872, 2006.
- [31] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.
- [32] Avril Coghlan, Evan E Eichler, Stephen G Oliver, Andrew H Paterson, and Lincoln Stein. Chromosome evolution in eukaryotes: a multi-kingdom perspective. *TRENDS in Genetics*, 21(12):673–682, 2005.
- [33] Nick Colegrave. Evolution. mark ridley. blackwell publishing. 2003. 751 pages. isbn 1 4051 0345 0. price£ 27.50. *Genetics Research*, 83(1):65–66, 2004.
- [34] Drosophila 12 Genomes Consortium et al. Evolution of genes and genomes on the drosophila phylogeny. *Nature*, 450(7167):203, 2007.
- [35] International Human Genome Sequencing Consortium et al. Initial sequencing and analysis of the human genome. *nature*, 409(6822):860, 2001.
- [36] Jasmin Coulombe-Huntington and Jacek Majewski. Characterization of intron loss events in mammals. *Genome research*, 17(1):23–32, 2007.
- [37] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561, 1970.
- [38] Thomas Czerny, Georg Halder, Urs Kloter, Abdallah Souabni, Walter J Gehring, and Meinrad Busslinger. twin of eyeless, a second pax-6 gene of drosophila, acts upstream of eyeless in the control of eye development. *Molecular cell*, 3(3):297–307, 1999.
- [39] Charles Darwin. *On the origin of species, 1859*. Routledge, 2004.

- [40] Riet De Smet, Keith L Adams, Klaas Vandepoele, Marc CE Van Montagu, Steven Maere, and Yves Van de Peer. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proceedings of the National Academy of Sciences*, 110(8):2898–2903, 2013.
- [41] Michael Dean, Mary Carrington, Cheryl Winkler, Gavin A Huttley, Michael W Smith, Rando Allikmets, James J Goedert, Susan P Buchbinder, Eric Vittinghoff, Edward Gomperts, et al. Genetic restriction of hiv-1 infection and progression to aids by a deletion allele of the ckr5 structural gene. *Science*, 273(5283):1856–1862, 1996.
- [42] Paramvir Dehal and Jeffrey L Boore. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS biology*, 3(10):e314, 2005.
- [43] Jeffery P Demuth, Tijn De Bie, Jason E Stajich, Nello Cristianini, and Matthew W Hahn. The evolution of mammalian gene families. *PloS one*, 1(1):e85, 2006.
- [44] Theodosius Dobzhansky. Nothing in biology makes sense except in the light of evolution. *The American Biology Teacher*, 35(3):125–129, 1973. ISSN: 0002-7685. doi:10.2307/4444260.
- [45] Robert C Edgar. Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*, 5(1):113, 2004.
- [46] Florian Ehrlich, Heinz Fischer, Lutz Langbein, Silke Praetzel-Wunder, Bettina Ebner, Katarzyna Figlak, Anton Weissenbacher, Wolfgang Sipos, Erwin Tschachler, and Leopold Eckhart. Differential evolution of the epidermal keratin cytoskeleton in terrestrial and aquatic mammals. *Molecular biology and evolution*, 36(2):328–340, 2018.
- [47] Florian Ehrlich, Maria Laggner, Lutz Langbein, Pamela Burger, Andreas Pollreis, Erwin Tschachler, and Leopold Eckhart. Comparative genomics suggests loss of keratin k24 in three evolutionary lineages of mammals. *Scientific reports*, 9(1):10924, 2019.
- [48] Janice A Fischer and Tom Maniatis. Structure and transcription of the drosophila mulleri alcohol dehydrogenase genes. *Nucleic acids research*, 13(19):6899–6917, 1985.
- [49] Robert D Fleischmann, Mark D Adams, Owen White, Rebecca A Clayton, Ewen F Kirkness, Anthony R Kerlavage, Carol J Bult, Jean-Francois Tomb, Brian A Dougherty, Joseph M Merrick, et al. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, 269(5223):496–512, 1995.
- [50] Pauline A Fujita, Brooke Rhead, Ann S Zweig, Angie S Hinrichs, Donna Karolchik, Melissa S Cline, Mary Goldman, Galt P Barber, Hiram Clawson, Antonio Coelho, et al. The ucsc genome browser database: update 2011. *Nucleic acids research*, 39(suppl_1):D876–D882, 2010.
- [51] Douglas J Futuyma and Anurag A Agrawal. Evolutionary history and species interactions. 2009.

- [52] Nicolas Galtier, Manolo Gouy, and Christian Gautier. Seaview and phylo_win: two graphic tools for sequence alignment and molecular phylogeny. *Bioinformatics*, 12(6):543–548, 1996.
- [53] Greg Gibson. *A primer of human genetics*. Sinauer Associates, Incorporated, Publishers, 2015.
- [54] Don Gilbert. Bioinformatics software resources. *Briefings in bioinformatics*, 5(3):300–304, 2004.
- [55] Walter Gilbert. Why genes in pieces? *Nature*, 271(5645):501, 1978.
- [56] Benjamin Goldman-Huertas, Robert F Mitchell, Richard T Lapoint, Cécile P Faucher, John G Hildebrand, and Noah K Whiteman. Evolution of herbivory in drosophilidae linked to loss of behaviors, antennal responses, odorant receptors, and ancestral diet. *Proceedings of the National Academy of Sciences*, 112(10):3026–3031, 2015.
- [57] Verne Grant. Plant speciation. Technical report, Columbia University Press,, 1981.
- [58] Anthony JF Griffiths, William M Gelbart, Richard C Lewontin, and Jeffrey H Miller. *Modern genetic analysis: integrating genes and genomes*, volume 1. Macmillan, 2002.
- [59] Zhenglong Gu, Andre Cavalcanti, Feng-Chi Chen, Peter Bouman, and Wen-Hsiung Li. Extent of gene duplication in the genomes of drosophila, nematode, and yeast. *Molecular biology and evolution*, 19(3):256–262, 2002.
- [60] Yoonsoo Hahn and Byungkook Lee. Human-specific nonsense mutations identified by genome sequence comparisons. *Human genetics*, 119(1-2):169–178, 2006.
- [61] Ross C Hardison, Eugene T Butler III, Elizabeth Lacy, Tom Maniatis, Nadia Rosenthal, and Argiris Efstratiadis. The structure and transcription of four linked rabbit β -like globin genes. *Cell*, 18(4):1285–1297, 1979.
- [62] Paul Harrison, Anuj Kumar, Ning Lan, Nathaniel Echols, Michael Snyder, and Mark Gerstein. A small reservoir of disabled orfs in the yeast genome and its implications for the dynamics of proteome evolution. *Journal of molecular biology*, 316(3):409–419, 2002.
- [63] Paul M Harrison, Duncan Milburn, Zhaolei Zhang, Paul Bertone, and Mark Gerstein. Identification of pseudogenes in the drosophila melanogaster genome. *Nucleic acids research*, 31(3):1033–1037, 2003.
- [64] Jennifer Harrow, Alinda Nagy, Alexandre Reymond, Tyler Alioto, Laszlo Patthy, Stylianos E Antonarakis, and Roderic Guigó. Identifying protein-coding genes in genomic sequences. *Genome Biology*, 10(1):201, 2009.
- [65] Yehudit Hasin, Marcus Seldin, and Aldons Lusi. Multi-omics approaches to disease. *Genome biology*, 18(1):83, 2017.

- [66] Nikolai Hecker, Virag Sharma, and Michael Hiller. Transition to an aquatic habitat permitted the repeated loss of the pleiotropic *klk8* gene in mammals. *Genome biology and evolution*, 9(11):3179–3188, 2017.
- [67] Nikolai Hecker, Virag Sharma, and Michael Hiller. Convergent gene losses illuminate metabolic and physiological changes in herbivores and carnivores. *Proceedings of the National Academy of Sciences*, 116(8):3036–3041, 2019.
- [68] Andreas Hejnol and Casey W Dunn. Animal evolution: Are phyla real? *Current Biology*, 26(10):R424–R426, 2016.
- [69] Javier Herrero, Matthieu Muffato, Kathryn Beal, Stephen Fitzgerald, Leo Gordon, Miguel Pignatelli, Albert J Vilella, Stephen MJ Searle, Ridwan Amode, Simon Brent, et al. Ensembl comparative genomics resources. *Database*, 2016, 2016.
- [70] Desmond G Higgins, Alan J Bleasby, and Rainer Fuchs. Clustal v: improved software for multiple sequence alignment. *Bioinformatics*, 8(2):189–191, 1992.
- [71] Maria Elena Hoballah, Thomas Gübitz, Jeroen Stuurman, Larissa Broger, Mario Barone, Therese Mandel, Alexandre Dell’Olivo, Maeva Arnold, and Cris Kuhlemeier. Single gene-mediated shift in pollinator attraction in petunia. *The Plant Cell*, 19(3):779–790, 2007.
- [72] Linda Z Holland and Daniel Ocampo Daza. A new look at an old question: when did the second whole genome duplication occur in vertebrate evolution? *Genome biology*, 19(1):209, 2018.
- [73] Alison K Hottes, Peter L Freddolino, Anupama Khare, Zachary N Donnell, Julia C Liu, and Saeed Tavazoie. Bacterial adaptation through loss of function. *PLoS genetics*, 9(7):e1003617, 2013.
- [74] Shengfeng Huang, Haozheng Tian, Zelin Chen, Ting Yu, and Anlong Xu. The evolution of vertebrate tetraspanins: gene loss, retention, and massive positive selection after whole genome duplications. *BMC evolutionary biology*, 10(1):306, 2010.
- [75] Weichun Huang, David M Umbach, and Leping Li. Accurate anchoring alignment of divergent sequences. *Bioinformatics*, 22(1):29–34, 2005.
- [76] Laura A Hug, Brett J Baker, Karthik Anantharaman, Christopher T Brown, Alexander J Probst, Cindy J Castelle, Cristina N Butterfield, Alex W Hernsdorf, Yuki Amano, Kotaro Ise, et al. A new view of the tree of life. *Nature microbiology*, 1(5):16048, 2016.
- [77] Austin L Hughes. Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. *Journal of molecular evolution*, 48(5):565–576, 1999.
- [78] Austin L Hughes and Robert Friedman. 2r or not 2r: testing hypotheses of genome duplication in early vertebrates. In *Genome Evolution*, pages 85–93. Springer, 2003.

- [79] Hideki Innan and Fyodor Kondrashov. The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics*, 11(2):97, 2010.
- [80] Jun Inoue, Yukuto Sato, Robert Sinclair, Katsumi Tsukamoto, and Mutsumi Nishida. Rapid genome reshaping by multiple-gene loss after whole-genome duplication in teleost fish suggested by mathematical modeling. *Proceedings of the National Academy of Sciences*, 112(48):14918–14923, 2015.
- [81] Manuel Irimia, David Penny, and Scott W Roy. Coevolution of genomic intron number and splice sites. *Trends in Genetics*, 23(7):321–325, 2007.
- [82] Claude Jacq, JR Miller, and GG Brownlee. A pseudogene structure in 5s dna of xenopus laevis. *Cell*, 12(1):109–120, 1977.
- [83] Erich D Jarvis, Siavash Mirarab, Andre J Aberer, Bo Li, Peter Houde, Cai Li, Simon YW Ho, Brant C Faircloth, Benoit Nabholz, Jason T Howard, et al. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215):1320–1331, 2014.
- [84] David Jebb and Michael Hiller. Recurrent loss of hmgcs2 shows that ketogenesis is not essential for the evolution of large mammalian brains. *Elife*, 7:e38906, 2018.
- [85] William R Jeffery. Regressive evolution in astyanax cavefish. *Annual review of genetics*, 43: 25–47, 2009.
- [86] Roy A Jensen. Enzyme recruitment in evolution of new function. *Annual Reviews in Microbiology*, 30(1):409–425, 1976.
- [87] Peihua Jiang, Jesusa Josue, Xia Li, Dieter Glaser, Weihua Li, Joseph G Brand, Robert F Margolskee, Danielle R Reed, and Gary K Beauchamp. Major taste loss in carnivorous mammals. *Proceedings of the National Academy of Sciences*, 109(13):4956–4961, 2012.
- [88] Roddy Jorquera, Rodrigo Ortiz, F Ossandon, Juan Pablo Cárdenas, Rene Sepúlveda, Carolina González, and David S Holmes. Sinex db: a database for single exon coding sequences in mammalian genomes. *Database*, 2016, 2016.
- [89] Mustapha Kandouz, Andrew Bier, George D Carystinos, Moulay A Alaoui-Jamali, and Gerald Batist. Connexin43 pseudogene is expressed in tumor cells and inhibits growth. *Oncogene*, 23(27):4763, 2004.
- [90] Philipp Kapranov, Jill Cheng, Sujit Dike, David A Nix, Radharani Duttgupta, Aaron T Willingham, Peter F Stadler, Jana Hertel, Jörg Hackermüller, Ivo L Hofacker, et al. Rna maps reveal new rna classes and a possible function for pervasive transcription. *Science*, 316(5830):1484–1488, 2007.
- [91] John E Karro, Yangpan Yan, Deyou Zheng, Zhaolei Zhang, Nicholas Carriero, Philip Cayting, Paul Harrision, and Mark Gerstein. Pseudogene. org: a comprehensive database

- and comparison platform for pseudogene annotation. *Nucleic acids research*, 35(suppl_1): D55–D60, 2006.
- [92] Matthew Kearse, Richard Moir, Amy Wilson, Steven Stones-Havas, Matthew Cheung, Shane Sturrock, Simon Buxton, Alex Cooper, Sidney Markowitz, Chris Duran, et al. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12):1647–1649, 2012.
- [93] Eugene V Koonin. Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, 39: 309–338, 2005.
- [94] Sergei A Korneev, Ji-Ho Park, and Michael O’Shea. Neuronal expression of neural nitric oxide synthase (nnos) protein is suppressed by an antisense rna transcribed from an nos pseudogene. *Journal of Neuroscience*, 19(18):7711–7720, 1999.
- [95] R Daniel Kortschak, Gabrielle Samuel, Robert Saint, and David J Miller. Est analysis of the cnidarian acropora millepora reveals extensive gene loss and rapid sequence divergence in the model invertebrates. *Current Biology*, 13(24):2190–2195, 2003.
- [96] Martin Krzywinski, Jacqueline Schein, Inanc Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9):1639–1645, 2009.
- [97] Arne Kusserow, Kevin Pang, Carsten Sturm, Martina Hroudá, Jan Lentfer, Heiko A Schmidt, Ulrich Technau, Arndt Von Haeseler, Bert Hobmayer, Mark Q Martindale, et al. Unexpected complexity of the wnt gene family in a sea anemone. *Nature*, 433(7022):156, 2005.
- [98] Mariana Lagos-Quintana, Reinhard Rauhut, Winfried Lendeckel, and Thomas Tuschl. Identification of novel genes coding for small expressed rnas. *Science*, 294(5543):853–858, 2001.
- [99] Jean Baptiste Lamarck. *Zoological philosophy*. University of Chicago, 1984.
- [100] Jong S Lee, Gregory G Brown, and Desh Pal S Verma. Chromosomal arrangement of leghemoglobin genes in soybean. *Nucleic acids research*, 11(16):5541–5553, 1983.
- [101] Wen Li, Wei Yang, and Xiu-Jie Wang. Pseudogenes: pseudo or real functional elements? *Journal of Genetics and Genomics*, 40(4):171–177, 2013.
- [102] Yi Li, Meng-Chen Zhang, Xiao-Kang Xu, Yang Zhao, Chato Mahanand, Tao Zhu, Hong Deng, Eviatar Nevo, Ji-Zeng Du, and Xue-Qun Chen. Functional diversity of p53 in human and wild animals. *Frontiers in endocrinology*, 10, 2019.
- [103] Pablo Librado, Filipe G Vieira, Alejandro Sanchez-Gracia, Sergios-Orestis Kolokotronis, and Julio Rozas. Mycobacterial phylogenomics: an enhanced method for gene turnover analysis reveals uneven levels of gene gain and loss among species and gene families. *Genome biology and evolution*, 6(6):1454–1465, 2014.

- [104] Kerstin Lindblad-Toh, Manuel Garber, Or Zuk, Michael F Lin, Brian J Parker, Stefan Washietl, Pouya Kheradpour, Jason Ernst, Gregory Jordan, Evan Mauceli, et al. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370):476, 2011.
- [105] Carolus Linnaeus. *Systema naturae per regna tria naturae*, 10th. *Edition. Laurentius Salvius, Holmiae, Stockholm, Sweden.[In Latin]*, 1758.
- [106] Yang Liu, Paul M Harrison, Victor Kunin, and Mark Gerstein. Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. *Genome biology*, 5(9):R64, 2004.
- [107] Brendan Loftus, Iain Anderson, Rob Davies, U Cecilia M Alsmark, John Samuelson, Paolo Amedeo, Paola Roncaglia, Matt Berriman, Robert P Hirt, Barbara J Mann, et al. The genome of the protist parasite *entamoeba histolytica*. *Nature*, 433(7028):865, 2005.
- [108] Mónica Lopes-Marques, André M Machado, Susana Barbosa, Miguel M Fonseca, Raquel Ruivo, and L Filipe C Castro. Cetacea are natural knockouts for il20. *Immunogenetics*, 70(10):681–687, 2018.
- [109] Monica Lopes-Marques, Luis Q Alves, Miguel Fonseca, Giulia Secci-Petretto, Andre M Machado, Raquel Ruivo, and Filipe Castro. Convergent inactivation of the skin-specific cc motif chemokine ligand 27 in mammalian evolution. *Immunogenetics*, 2019.
- [110] Mónica Lopes-Marques, André M Machado, Luís Q Alves, Miguel M Fonseca, Susana Barbosa, Mikkel-Holger S Sinding, Marianne Helene Rasmussen, Maria Refsgaard Iversen, Mads Frost Bertelsen, Paula F Campos, et al. Complete inactivation of sebum-producing genes parallels the loss of sebaceous glands in cetacea. *Molecular biology and evolution*, 2019.
- [111] Monica Lopes-Marques, Raquel Ruivo, Luís Q Alves, Nelson Sousa, André M Machado, and L Filipe C Castro. The singularity of cetacea behavior parallels the complete inactivation of melatonin gene modules. *Genes*, 10(2):121, 2019.
- [112] Mónica Lopes-Marques. *The evolution of fatty acid metabolism in chordates*, 2017.
- [113] Jonathan B Losos, Stevan J Arnold, Gill Bejerano, ED Brodie III, David Hibbett, Hopi E Hoekstra, David P Mindell, Antónia Monteiro, Craig Moritz, H Allen Orr, et al. Evolutionary biology for the 21st century. *PLoS Biology*, 11(1):e1001466, 2013.
- [114] Edward J Louis. Evolutionary genetics: making the most of redundancy. *Nature*, 449(7163):673, 2007.
- [115] Mark Lundstrom. Moore’s law forever? *Science*, 299(5604):210–211, 2003.
- [116] Charles Lyell and Gérard Paul Deshayes. *Principles of geology: being an attempt to explain the former changes of the earth’s surface, by reference to causes now in operation*, volume 1. J. Murray, 1830.

- [117] Michael Lynch. Streamlining and simplification of microbial genome architecture. *Annu. Rev. Microbiol.*, 60:327–349, 2006.
- [118] Michael Lynch and John S Conery. The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–1155, 2000.
- [119] Thomas MacCarthy and Aviv Bergman. The limits of subfunctionalization. *BMC evolutionary biology*, 7(1):213, 2007.
- [120] Steven Maere, Stefanie De Bodt, Jeroen Raes, Tineke Casneuf, Marc Van Montagu, Martin Kuiper, and Yves Van de Peer. Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences*, 102(15):5454–5459, 2005.
- [121] Elaine R Mardis. A decade’s perspective on dna sequencing technology. *Nature*, 470(7333):198, 2011.
- [122] Andrew Martin. Is tetralogy true? lack of support for the “one-to-four rule”. *Molecular biology and evolution*, 18(1):89–93, 2001.
- [123] Jane Masterson. Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. *Science*, 264(5157):421–424, 1994.
- [124] Carolyn S McBride and J Roman Arguello. Five drosophila genomes reveal nonneutral evolution and the signature of host specialization in the chemoreceptor superfamily. *Genetics*, 177(3):1395–1416, 2007.
- [125] John P McCutcheon and Nancy A Moran. Extreme genome reduction in symbiotic bacteria. *Nature Reviews Microbiology*, 10(1):13, 2012.
- [126] Camille Meslin, Sylvie Mugnier, Isabelle Callebaut, Michel Laurin, Géraldine Pascal, Anne Poupon, Ghylène Goudet, and Philippe Monget. Evolution of genes involved in gamete interaction: evidence for positive selection, duplications and losses in vertebrates. *PloS one*, 7(9):e44548, 2012.
- [127] Wynn K Meyer, Jerrica Jamison, Rebecca Richter, Stacy E Woods, Raghavendran Partha, Amanda Kowalczyk, Charles Kronk, Maria Chikina, Robert K Bonde, Daniel E Crocker, et al. Ancient convergent losses of paraoxonase 1 yield potential risks for modern marine mammals. *Science*, 361(6402):591–594, 2018.
- [128] AJ Mighell, NR Smith, PA Robinson, and AF Markham. Vertebrate pseudogenes. *FEBS letters*, 468(2-3):109–114, 2000.
- [129] Flavio Mignone, Carmela Gissi, Sabino Liuni, and Graziano Pesole. Untranslated regions of mrnas. *Genome biology*, 3(3):reviews0004–1, 2002.
- [130] Madison Miller, Nitin Shirole, Ruxiao Tian, Debjani Pal, and Raffaella Sordella. The evolution of tp53 mutations: from loss-of-function to separation-of-function mutants. *Journal of cancer biology & research*, 4(4), 2016.

- [131] Webb Miller, Kateryna D Makova, Anton Nekrutenko, and Ross C Hardison. Comparative genomics. *Annu. Rev. Genomics Hum. Genet.*, 5:15–56, 2004.
- [132] Peter Mombaerts. The human repertoire of odorant receptor genes and pseudogenes. *Annual review of genomics and human genetics*, 2(1):493–510, 2001.
- [133] Damian Moran, Rowan Softley, and Eric J Warrant. The energetic cost of vision and the evolution of eyeless mexican cavefish. *Science advances*, 1(8):e1500363, 2015.
- [134] Burkhard Morgenstern, Oliver Rinner, Saïd Abdeddaïm, Dirk Haase, Klaus FX Mayer, Andreas WM Dress, and Hans-Werner Mewes. Exon discovery by genomic sequence alignment. *Bioinformatics*, 18(6):777–787, 2002.
- [135] Olena Morozova and Marco A Marra. Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 92(5):255–264, 2008.
- [136] Andrés Moya, Juli Peretó, Rosario Gil, and Amparo Latorre. Learning how to live together: genomic insights into prokaryote–animal symbioses. *Nature Reviews Genetics*, 9(3):218, 2008.
- [137] Fatimah Noni Muhamad, RB Ahmad, S Mohd Asi, and MN Murad. Reducing the search space and time complexity of needleman-wunsch algorithm (global alignment) and smith-waterman algorithm (local alignment) for dna sequence alignment. *Jurnal Teknologi*, 77(20), 2011.
- [138] Paul Muir, Shantao Li, Shaoke Lou, Daifeng Wang, Daniel J Spakowicz, Leonidas Salichos, Jing Zhang, George M Weinstock, Farren Isaacs, Joel Rozowsky, et al. The real cost of sequencing: scaling computation to keep pace with data generation. *Genome biology*, 17(1):53, 2016.
- [139] Yoichiro Nakatani, Hiroyuki Takeda, Yuji Kohara, and Shinichi Morishita. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome research*, 17(9):1254–1265, 2007.
- [140] SB Needleman. Needleman-wunsch algorithm for sequence similarity searches. *J Mol Biol*, 48:443–453, 1970.
- [141] Mariana F Nery, José Ignacio Arroyo, and Juan C Opazo. Increased rate of hair keratin gene loss in the cetacean lineage. *BMC genomics*, 15(1):869, 2014.
- [142] Martin A Nowak, Maarten C Boerlijst, Jonathan Cooke, and John Maynard Smith. Evolution of genetic redundancy. *Nature*, 388(6638):167, 1997.
- [143] Genome 10K Community of Scientists. Genome 10k: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. *Journal of Heredity*, 100(6):659–674, 2009.
- [144] Susumu Ohno. *Evolution by gene duplication*. Springer Science & Business Media, 1970.

- [145] Maynard V Olson. When less is more: gene loss as an engine of evolutionary change. *The American Journal of Human Genetics*, 64(1):18–23, 1999.
- [146] LE Orgel. Gene-duplication and the origin of proteins with novel functions. *Journal of theoretical biology*, 67(4):773, 1977.
- [147] Georgia Panopoulou, Steffen Hennig, Detlef Groth, Antje Krause, Albert J Poustka, Ralf Herwig, Martin Vingron, and Hans Lehrach. New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. *Genome research*, 13(6a):1056–1066, 2003.
- [148] George H Perry, Raul Y Tito, and Brian C Verrelli. The evolutionary history of human and chimpanzee y-chromosome gene loss. *Molecular biology and evolution*, 24(3):853–859, 2007.
- [149] Eric Peyretailade, Hicham El Alaoui, Marie Diogon, Valérie Polonais, Nicolas Parisot, David G Biron, Pierre Peyret, and Frédéric Delbac. Extreme reduction and compaction of microsporidian genomes. *Research in microbiology*, 162(6):598–606, 2011.
- [150] Ernesto Picardi and Graziano Pesole. Computational methods for ab initio and comparative gene finding. pages 269–284, 2010.
- [151] Ryan Charles Pink, Kate Wicks, Daniel Paul Caley, Emma Kathleen Punch, Laura Jacobs, and David Raul Francisco Carter. Pseudogenes: pseudo-functional or key regulators in health and disease? *Rna*, 17(5):792–798, 2011.
- [152] Ondrej Podlaha and Jianzhi Zhang. Pseudogenes and their evolution. *e LS*, 2001.
- [153] Kornelia Polyak and Matthew Meyerson. Overview: gene structure. In *Holland-Frei Cancer Medicine. 6th edition*. BC Decker, 2003.
- [154] Meredith E Protas, Candace Hersey, Dawn Kochanek, Yi Zhou, Horst Wilkens, William R Jeffery, Leonard I Zon, Richard Borowsky, and Clifford J Tabin. Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. *Nature genetics*, 38(1):107, 2006.
- [155] Nick Proudfoot. Pseudogenes. *Nature*, 286(5776):840, 1980.
- [156] Nicholas H Putnam, Thomas Butts, David EK Ferrier, Rebecca F Furlong, Uffe Hellsten, Takeshi Kawashima, Marc Robinson-Rechavi, Eiichi Shoguchi, Astrid Terry, Jr-Kai Yu, et al. The amphioxus genome and the evolution of the chordate karyotype. *Nature*, 453(7198):1064, 2008.
- [157] Wenfeng Qian, Ben-Yang Liao, Andrew Ying-Fei Chang, and Jianzhi Zhang. Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends in Genetics*, 26(10):425–430, 2010.
- [158] Anand Ramanathan, G Brett Robb, and Siu-Hong Chan. mrna capping: biological functions and applications. *Nucleic acids research*, 44(16):7511–7526, 2016.

- [159] Vincent Ranwez, Sébastien Harispe, Frédéric Delsuc, and Emmanuel JP Douzery. Macse: Multiple alignment of coding sequences accounting for frameshifts and stop codons. *PLoS one*, 6(9):e22594, 2011.
- [160] Vincent Ranwez, Emmanuel JP Douzery, Cédric Cambon, Nathalie Chantret, and Frédéric Delsuc. Macse v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Molecular biology and evolution*, 35(10):2582–2584, 2018.
- [161] Gene E Robinson, Kevin J Hackett, Mary Purcell-Miramontes, Susan J Brown, Jay D Evans, Marian R Goldsmith, Daniel Lawson, Jack Okamuro, Hugh M Robertson, and David J Schneider. Creating a buzz about insect genomes. *Science*, 331(6023):1386–1386, 2011.
- [162] Eric C Rouchka and I Elizabeth Cha. Current trends in pseudogene detection and characterization. *Current Bioinformatics*, 4(2):112–119, 2009.
- [163] Lee Rowen. Gene structure and organization. *e LS*, 2001.
- [164] Naruya Saitou. *Introduction to evolutionary genomics*. Springer, 2013.
- [165] Meena Kishore Sakharkar, Vincent TK Chow, and Pandjassaram Kanguane. Distributions of exons and introns in the human genome. *In silico biology*, 4(4):387–393, 2004.
- [166] Alex Sánchez-Pla, Ferran Reverter, M Carme Ruíz de Villa, and Manuel Comabella. Transcriptomics: mrna and alternative splicing. *Journal of neuroimmunology*, 248(1-2): 23–31, 2012.
- [167] Andrea Sboner, Xinmeng Jasmine Mu, Dov Greenbaum, Raymond K Auerbach, and Mark B Gerstein. The real cost of sequencing: higher than you think! *Genome biology*, 12(8):125, 2011.
- [168] Kathy Scienski, Justin C Fay, and Gavin C Conant. Patterns of gene conversion in duplicated yeast histones suggest strong selection on a coadapted macromolecular complex. *Genome biology and evolution*, 7(12):3249–3258, 2015.
- [169] Cathal Seoighe and Chris Gehring. Genome duplication led to highly selective expansion of the arabidopsis thaliana proteome. *Trends in Genetics*, 20(10):461–464, 2004.
- [170] Virag Sharma and Michael Hiller. Loss of enzymes in the bile acid synthesis pathway explains differences in bile composition among mammals. *Genome biology and evolution*, 10(12):3211–3217, 2018.
- [171] Virag Sharma, Anas Elghafari, and Michael Hiller. Coding exon-structure aware realigner (cesar) utilizes genome alignments for accurate comparative gene annotation. *Nucleic acids research*, 44(11):e103–e103, 2016.
- [172] Virag Sharma, Nikolai Hecker, Juliana G Roscito, Leo Foerster, Bjoern E Langer, and Michael Hiller. A genomics approach reveals insights into the importance of gene losses for mammalian adaptations. *Nature communications*, 9(1):1215, 2018.

- [173] Virag Sharma, Thomas Lehmann, Heiko Stuckas, Liane Funke, and Michael Hiller. Loss of *rxfp2* and *insl3* genes in afrotheria shows that testicular descent is the ancestral condition in placental mammals. *PLoS biology*, 16(6):e2005293, 2018.
- [174] Patricia Sieber, Matthias Platzer, and Stefan Schuster. The definition of open reading frame revisited. *Trends in Genetics*, 34(3):167–170, 2018.
- [175] Selvarajan Sivashankari and Piramanayagam Shanmughavel. Comparative genomics-a perspective. *Bioinformation*, 1(9):376, 2007.
- [176] Lucy Skrabanek and Kenneth H Wolfe. Eukaryote genome duplication-where’s the evidence? *Current opinion in genetics & development*, 8(6):694–700, 1998.
- [177] Guy St C Slater and Ewan Birney. Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics*, 6(1):31, 2005.
- [178] Wendy Weijia Soon, Manoj Hariharan, and Michael P Snyder. High-throughput sequencing for biology and medicine. *Molecular systems biology*, 9(1):640, 2013.
- [179] Pietro D Spanu, James C Abbott, Joelle Amselem, Timothy A Burgis, Darren M Soanes, Kurt Stüber, Emiel Ver Loren van Themaat, James KM Brown, Sarah A Butcher, Sarah J Gurr, et al. Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. *Science*, 330(6010):1543–1546, 2010.
- [180] Mark S Springer, Christopher A Emerling, John Gatesy, Jason Randall, Matthew A Collin, Nikolai Hecker, Michael Hiller, and Frédéric Delsuc. Odontogenic ameloblast-associated (odam) is inactivated in toothless/enamelless placental mammals and toothed whales. *BMC evolutionary biology*, 19(1):31, 2019.
- [181] Hallam Stevens. *Life out of sequence: a data-driven history of bioinformatics*. University of Chicago Press, 2013.
- [182] Wing-Kin Sung. *Algorithms in bioinformatics: A practical introduction*. CRC Press, 2009.
- [183] Eörs Szathmáry, Ferenc Jordán, and Csaba Pál. Can genes explain biological complexity? *Science*, 292(5520):1315–1316, 2001.
- [184] Oliver H Tam, Alexei A Aravin, Paula Stein, Angelique Girard, Elizabeth P Murchison, Sihem Cheloufi, Emily Hodges, Martin Anger, Ravi Sachidanandam, Richard M Schultz, et al. Pseudogene-derived small interfering rnas regulate gene expression in mouse oocytes. *Nature*, 453(7194):534, 2008.
- [185] Ulrich Technau, Stephen Rudd, Peter Maxwell, Paul MK Gordon, Michael Saina, Laretta C Grasso, David C Hayward, Christoph W Sensen, Robert Saint, Thomas W Holstein, et al. Maintenance of ancestral complexity and non-metazoan genes in two basal cnidarians. *TRENDS in Genetics*, 21(12):633–639, 2005.

- [186] Françoise Thibaud-Nissen, Shu Ouyang, and C Robin Buell. Identification and characterization of pseudogenes in the rice gene complement. *BMC genomics*, 10(1):317, 2009.
- [187] David Torrents, Mikita Suyama, Evgeny Zdobnov, and Peer Bork. A genome-wide survey of human pseudogenes. *Genome research*, 13(12):2559–2567, 2003.
- [188] Christophe Tournamille, Yves Colin, Jean Pierre Cartron, and Caroline Le Van Kim. Disruption of a gata motif in the duffy gene promoter abolishes erythroid gene expression in duffy-negative individuals. *Nature genetics*, 10(2):224, 1995.
- [189] Roel Van Assche, Valérie Broeckx, Kurt Boonen, Evelyne Maes, Wouter De Haes, Liliane Schoofs, and Liesbet Temmerman. Integrating-omics: systems biology as explored through c. elegans research. *Journal of molecular biology*, 427(21):3441–3451, 2015.
- [190] Marijke J van Baren and Michael R Brent. Iterative gene prediction and pseudogene removal improves genome annotation. *Genome research*, 16(5):678–685, 2006.
- [191] Elio F Vanin. Processed pseudogenes: characteristics and evolution. *Annual review of genetics*, 19(1):253–272, 1985.
- [192] Byrappa Venkatesh, Alison P Lee, Vydianathan Ravi, Ashish K Maurya, Michelle M Lian, Jeremy B Swann, Yuko Ohta, Martin F Flajnik, Yoichi Sutoh, Masanori Kasahara, et al. Elephant shark genome provides unique insights into gnathostome evolution. *Nature*, 505(7482):174, 2014.
- [193] Andreas Wagner. Evolution of gene networks by gene duplications: a mathematical model and its implications on genome organization. *Proceedings of the National Academy of Sciences*, 91(10):4387–4391, 1994.
- [194] Andreas Wagner. Distributed robustness versus redundancy as causes of mutational robustness. *Bioessays*, 27(2):176–188, 2005.
- [195] Alfred Russel Wallace. *Darwinism: an exposition of the theory of natural selection with some of its applications*. Cosimo, Inc., 2007.
- [196] Xiaoxia Wang, Wendy E Grus, and Jianzhi Zhang. Gene losses during human origins. *PLoS biology*, 4(3):e52, 2006.
- [197] Yan Wang, Jing Liu, BO Huang, Yan-Mei Xu, Jing Li, Lin-Feng Huang, Jin Lin, Jing Zhang, Qing-Hua Min, Wei-Ming Yang, et al. Mechanism of alternative splicing and its regulation. *Biomedical reports*, 3(2):152–158, 2015.
- [198] James D Watson and Francis HC Crick. The structure of dna. In *Cold Spring Harbor symposia on quantitative biology*, volume 18, pages 123–131. Cold Spring Harbor Laboratory Press, 1953.

- [199] Rasmus Wernersson and Anders Gorm Pedersen. Revtrans: multiple alignment of coding dna from aligned amino acid sequences. *Nucleic acids research*, 31(13):3537–3539, 2003.
- [200] Travis J Wheeler and John D Kececioglu. Multiple alignment by aligning alignments. *Bioinformatics*, 23(13):i559–i568, 2007.
- [201] Yuri I Wolf and Eugene V Koonin. Genome reduction as the dominant mode of evolution. *Bioessays*, 35(9):829–837, 2013.
- [202] Lu Xiao-Jie, Gao Ai-Mei, Ji Li-Juan, and Xu Jiang. Pseudogene in cancer: real functions and promising signature. *Journal of medical genetics*, 52(1):17–24, 2015.
- [203] Ziheng Yang. Paml 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8):1586–1591, 2007.
- [204] Alison Yao, Rosane Charlab, and Peter Li. Systematic identification of pseudogenes through whole genome expression evidence profiling. *Nucleic acids research*, 34(16):4477–4485, 2006.
- [205] Jianzhi Zhang. Evolution by gene duplication: an update. *Trends in ecology & evolution*, 18(6):292–298, 2003.
- [206] Jianzhi Zhang, Helene F Rosenberg, and Masatoshi Nei. Positive darwinian selection after gene duplication in primate ribonuclease genes. *Proceedings of the National Academy of Sciences*, 95(7):3708–3713, 1998.
- [207] Michael Q Zhang. Computational prediction of eukaryotic protein-coding genes. *Nature Reviews Genetics*, 3(9):698, 2002.
- [208] Xinmin Zhang and Stuart Firestein. The olfactory receptor gene superfamily of the mouse. *Nature neuroscience*, 5(2):124, 2002.
- [209] ZhaoLei Zhang and Mark Gerstein. Large-scale analysis of pseudogenes in the human genome. *Current opinion in genetics & development*, 14(4):328–335, 2004.
- [210] Zhaolei Zhang, Paul M Harrison, Yin Liu, and Mark Gerstein. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome research*, 13(12):2541–2558, 2003.
- [211] Zhaolei Zhang, Nick Carriero, and Mark Gerstein. Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends in Genetics*, 20(2):62–67, 2004.
- [212] Zhaolei Zhang, Nicholas Carriero, Deyou Zheng, John Karro, Paul M Harrison, and Mark Gerstein. Pseudopipe: an automated pseudogene identification pipeline. *Bioinformatics*, 22(12):1437–1439, 2006.
- [213] Zhengdong D Zhang, Adam Frankish, Toby Hunt, Jennifer Harrow, and Mark Gerstein. Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. *Genome biology*, 11(3):R26, 2010.

-
- [214] Yi Zhao, Liang Tang, Zhe Li, Jinpu Jin, Jingchu Luo, and Ge Gao. Identification and analysis of unitary loss of long-established protein-coding genes in poaceae shows evidences for biased gene loss and putatively functional transcription of relics. *BMC evolutionary biology*, 15(1):66, 2015.
- [215] Deyou Zheng and Mark B Gerstein. A computational approach for identifying pseudogenes in the encode regions. *Genome biology*, 7(1):S13, 2006.
- [216] Deyou Zheng, Adam Frankish, Robert Baertsch, Philipp Kapranov, Alexandre Reymond, Siew Woh Choo, Yontao Lu, France Denoeud, Stylianos E Antonarakis, Michael Snyder, et al. Pseudogenes in the encode regions: consensus annotation, analysis of transcription, and evolution. *Genome research*, 17(6):839–851, 2007.
- [217] Rebecca A Zufall and Mark D Rausher. Genetic changes associated with floral adaptation restrict future evolutionary potential. *Nature*, 428(6985):847, 2004.