Aus der Klinik für ……..
Direktor: Frau/ Herr Prof. Dr. …….

A novel approach to infer orthologs and produce gene annotations at scale

D i s s e r t a t i o n s s c h r i f t
zur Erlangung des akademischen Grades
Doctor of Philosophy (Ph.D.)
vorgelegt
der Medizinischen Fakultät Carl Gustav Carus
der Technischen Universität Dresden

von

Dipl.-Biol. Bogdan M. Kirilenko

aus Moskau

Dresden 2021

2. Blatt (2. Seite)

1. Gutachter: Prof. Michael Hiller, Senckenberg Society for Nature Research & Goethe University
2. Gutachter: Prof. Marino Zerial, Max Planck Institute of Molecular Cell Biology and Genetics

Tag der mündlichen Prüfung: (Verteidigungstermin)

gez.: ----------
Vorsitzender der Promotionskommission

Anmerkung:

Die Eintragung der Gutachter und Tag der mündlichen Prüfung (Verteidigung) erfolgt nach Festlegung von Seiten der Medizinischen Fakultät Carl Gustav Carus der TU Dresden. Sie wird durch die Promovenden nach der Verteidigung zwecks Übergabe der fünf Pflichtexemplare an die Zweigbibliothek Medizin in gedruckter Form oder handschriftlich vorgenommen.

Contents

List of abbreviations

aa - amino acid
BLOSUM - BLOcks SUbstitution Matrix
bp - base pairs
CDS - protein-CoDing Sequence
CESAR - Coding Exon-Structure Aware Realigner
DAG - Directed Acyclic Graph
EST - Expressed Sequence Tag
FPR - False Positive Rate
HGT - Horizontal Gene Transfer
HMM - Hidden Markov Model
HPC - High-Performance Computing
indel - insertion/deletion
Mya - Million years ago
NMD - Nonsense-Mediated mRNA Decay
NN - Neural Network
nt - nucleotides
OR - Odorant Receptor
ORF - Open Reading Frame
PP - Processed Pseudogene
ROC - Receiver Operating Characteristic
sps - substitutions per site
TPR - True Positive Rate
UCSC - University of California, Santa Cruz
UTR - UnTranslated Region
ZNF - Zinc Finger

List of figures

List of tables

Abstract

Due to advances in DNA sequencing technologies, the number of available genome assemblies rapidly grows over the last decades. Nowadays, thousands of genomes representing a great variety of clades are available for the scientific community. Genomic data allows deep and detailed comparative analysis, which are important for relevant research questions such as discovery connections between genotype and phenotype, exploring particularities in complex proteins, and advancing medical applications. To address all these questions, it is necessary to annotate protein-coding genes and infer orthologs in newly sequenced genomes. However, the existing methods for genome analysis are not adapted for increased data scales. Therefore, the major challenge in comparative genomics is not the number of available genome assemblies but rather the development of high-throughput data analysis methods.

To address these issues, I propose a novel paradigm of genome annotation and orthology inference that uses whole-genome alignments. Whereas the currently applied gene annotation and orthology inference methods only rely on presumably coding sequences, expanding input data to neutrally evolving sequences would provide a genomic context for better and more complete annotations by highlighting orthologous loci, which can be predicted before gene annotation. By using pairwise genome alignments, the proposed methodology could be easily scaled to thousands of considered genomes.

In this work, I present TOGA (Tool to infer Orthologs from Genome Alignments) - a bioinformatics method that implements the proposed concept, combining orthology inference and gene annotation in a single pipeline. TOGA uses machine learning to distinguish orthologs from paralogs or processed pseudogenes based on alignments of intronic and intergenic regions.

The quality evaluation results show that the TOGA pipeline outcompetes the traditional approaches within the Placentals clade. It provides excellent orthology loci classification quality and detects loss genes with a high degree of accuracy. Earlier TOGA versions have been applied in several studies resulting in two publications so far. Using TOGA, we generated gene annotations for 500 mammalian genomes, creating the most extensive comparative dataset.

The results suggest that TOGA has the potential to become a widely accepted gene annotation method. I show that it can perfectly complement the currently applied techniques to create comprehensive genome annotations.

Zusammenfassung

Aufgrund von Fortschritten im Bereich der DNA-Sequenzierung hat die Anzahl verfügbarer Genome in den letzten Jahrzehnten rapide zugenommen. Tausende bereits heute zur Verfügung stehende Genome ermöglichen detaillierte vergleichende Analysen, welche für die Beantwortung relevanter Fragestellungen essentiell sind. Dies betrifft die Assoziation von Genotyp und Phänotyp, die Erforschung der Besonderheiten komplexer Proteine und die Weiterentwicklung medizinischer Anwendungen. Um all diese Fragen zu beantworten ist es notwendig, proteinkodierende Gene in neu sequenzierten Genomen zu annotieren und ihre Homologieverhältnisse zu bestimmen. Die bestehenden Methoden der Genomanalyse sind jedoch nicht für Menge heutzutage anfallender Datenmengen ausgelegt. Daher ist die zentrale Herausforderung in der vergleichenden Genomik nicht die Anzahl der verfügbaren Genome, sondern die Entwicklung neuer Methoden zur Datenanalyse im Hochdurchsatz.

Um diese Probleme zu adressieren, schlage ich ein neues Paradigma der Annotation von Genomen und der Inferenz von Homologieverhältnissen vor, welches auf dem Alignment gesamter Genome basiert. Während die derzeit angewendeten Methoden zur Gen-Annotation und Bestimmung der Homologie ausschließlich auf codierenden Sequenzen beruhen, könnten durch die Einbeziehung des umgebenden neutral evolvierenden genomischen Kontextes bessere und vollständigere Annotationen vorgenommen werden. Die Verwendung von Genom-Alignments ermöglicht eine beliebige Skalierung der vorgeschlagenen Methodik auf Tausende Genome.

In dieser Arbeit stelle ich TOGA (Tool to infer Orthologs from Genome Alignments) vor, eine bioinformatische Methode, welche dieses Konzept implementiert und Homologie-Klassifizierung und Gen-Annotation in einer einzelnen Pipeline kombiniert. TOGA verwendet Machine-Learning, um Orthologe von Paralogen basierend auf dem Alignment von intronischer und intergener Regionen zu unterscheiden.

Die Ergebnisse des Benchmarkings zeigen, dass TOGA die herkömmlichen Ansätze innerhalb der Placentalia übertrifft. TOGA klassifiziert Homologieverhältnisse mit hoher Präzision und identifiziert zuverlässig inaktivierte Gene als solchet. Frühere Versionen von TOGA fanden in mehreren Studien Anwendung und wurden in zwei Publikationen verwendet. Außerdem wurde TOGA erfolgreich zur Annotation von 500 Säugetiergeenomen verwendet, dies ist der bisher umfangreichste solche Datensatz.

Diese Ergebnisse zeigen, dass TOGA das Potenzial hat, sich zu einer etablierten Methode zur Gen-Annotation zu entwickeln und die derzeit angewandten Techniken zu ergänzen.

Acknowledgments

First and foremost, I would like to express my gratefulness to my supervisor, *Michael Hiller*, for the opportunity to work on this exciting project. His guidance, support, motivation, and overall insights in the field have made this project an inspirational experience for me. Moreover, I like to acknowledge *Marino Zerial* and *Jochen Rink* for their intellectual contributions and helpful suggestions at our TAC meetings.

I cannot leave MPI-CBG's computer department staff without mentioning, especially *Oscar Gonzales,* for their patience and for keeping our computational infrastructure intact. In this context, I like to thank the personnel of MPI-CBG's international and PhD offices for making the official aspects of my life in Dresden much more effortless than this could be.

My completion of this project could not have been accomplished without my labmates. Thank you for the comfortable, supportive, and stimulating atmosphere. It was a pleasure to work with you.

Especially, I would like to acknowledge my colleagues who tested my software in practice: *David Jebb, Moritz Blumer, Katya Osipova, Ariadna Morales*, *Alexis Ahmed*, and others. It is hard to overrate the importance of your feedback and suggestions to implement the final pipeline and consequently finish the project.

I would like to extend my sincere thanks to my parents, *Mikhail and Tatyana Kirilenko*, for their great support and motivation. And remarkably, I want to thank my brother *Ilya Kirilenko* for his valuable suggestions regarding my thesis.

Last but not least, a special thank you to my wife, *Anastasiya Kirilenko*, for being with me during this period. For your support, understanding, and patience.

List of own publications

1. TOGA: a novel machine-learning approach to infer orthologs and integrate gene annotation with orthology inference at scale. *In preparation.* Bogdan M. Kirilenko, Ekaterina Osipova, David Jebb, Chetan Munegowda, Virag Sharma, Michael Hiller
2. Six reference-quality genomes reveal evolution of bat adaptations. Nature 583, 578-584 (2020) David Jebb, Zixia Huang, Martin Pippel, Graham M. Hughes, Ksenia Lavrichenko, Paolo Devanna, Sylke Winkler, Lars S. Jermiin, Emilia C. Skirmuntt, Aris Katzourakis, Lucy Burkitt-Gray, David A. Ray, Kevin A. M. Sullivan, Juliana G. Roscito, Bogdan M. Kirilenko, Liliana M. Dávalos, Angelique P. Corthals, Megan L. Power, Gareth Jones, Roger D. Ransome, Dina K. N. Dechmann, Andrea G. Locatelli, Sébastien J. Puechmaille, Olivier Fedrigo, Erich D. Jarvis, Michael Hiller, Sonja C. Vernes, Eugene W. Myers, Emma C. Teeling.
3. Evolutionary Analysis of Bile Acid-Conjugating Enzymes Reveals a Complex Duplication and Reciprocal Loss History. Genome Biol Evol, 11(11) 3256-3268 (2019) Bogdan Kirilenko, Lee R Hagey, Stephen Barnes, Charles N Falany, Michael Hiller.
4. Molecular parallelism in fast-twitch muscle proteins in echolocating mammals. Sci Adv, 4(9) Art. No eaat9660 (2018). Jun Hoe Lee, Kevin M Lewis, Timothy W Moural, Bogdan Kirilenko, Barbara Borgonovo, Gisa Prange, Manfred Koessl, Stefan Huggenberger, ChulHee Kang, Michael Hiller.
5. Inhibition of the expression of proteasomal genes Saccharomyces cerevisiae by artificial transcriptional repressor. Mol Biol (Mosk). 2016 Jul-Aug;50(4):703-712. Kirilenko BM, Grineva EN, Karpov DS, Karpov VL.

“A lot of times, people don’t know what they want until you show it to them.”

Steve Jobs (1997)

1. General Introduction

1.1 The genome encodes most phenotypes

Evolution has led to a great diversity of living forms that inhabit every single corner of the Earth. These forms vary in different aspects, such as physical scale, appearance, accepted energy sources, and responsiveness to different stimuli. Various creatures represent the diversity of life, including unicellular photosynthetic algae, flying frugivorous drosophilae, or gigantic marine predators, which are well adapted to their environments.

This tremendous biological multifariousness is encoded in the DNA sequences of four characters: A, T, G, and C. Each adaptation and peculiarity is reflected in the species' genome. Nowadays, the genomes of thousands of different species are sequenced, ranging from relatively primitive unicellular life forms to profoundly complex mammals, and this number continues to grow. Plenty of available genomic data allows us to connect differences in genomic sequences with specific phenotypic traits. Discovering these connections, being the fundamental challenge of comparative genomics, requires identifying the genetic origin of phenotypic variation, and reveals insights into the underlying molecular and cellular mechanisms.

1.1.1 Overview of genomic changes related to phenotypic variation

Numerous studies published in scientific literature describe various connections between genomic changes and phenotypic adaptations in different species. Genomic DNA comprises many types of sequences, such as repeats (e.g., transposons, tandem repeats, centromeres, telomeres), regulatory elements (e.g., enhancers, transcription factor binding sites, nrRNAs), RNA- and protein-coding genes. However, the vast majority of these studies focus on changes in protein-coding genes because they can significantly affect phenotypes (Brandes et al., 2020).

Therefore, annotating protein-coding genes is typically the first step to be performed on a newly sequenced genome. Even though only a minor fraction of the genomic DNA encodes proteins (for human, ~1% (International Human Genome Sequencing Consortium, 2001)), mutations in these regions could affect virtually any system of the organism. For example, these mutations can influence the performance of the immune system (Jebb et al., 2020), affect teeth development (Springer et al., 2019), or contribute to the evolution of echolocation (Li et al., 2010; Liu et al., 2014; Lee et al., 2018). Also, many human diseases are associated with mutations in protein-coding genes such as *GNAT2*, *TJP2*, *BAAT*, or *TTN*

(Kohl et al., 2002, p. 2; Carlton et al., 2003; Lange et al., 2005). In the following subsections, I give a brief overview of mutations affecting protein-coding genes.

1.1.1.1 Amino acid changes in the functionally critical regions

Mutations that alter the amino acid sequence, such as substitutions, insertions, or deletions, are random and have a different impact on the protein functionality, depending on mutation localization and radicality (Choi et al., 2012). Radical amino acid changes occurring in critical protein regions are likely to influence biological function. For instance, mutations in the active center of an enzyme might affect substrate specificity towards another compound (Smooker et al., 2000; Kirilenko et al., 2019). Moreover, they can diversify the DNA-binding protein specificity (Jalal et al., 2020). Consequently, this class of mutations has a great potential to affect the phenotype.

1.1.1.2 Evolutionary origins of new genes

Another evolutionary process that deserves particular attention is the creation of new genes. Mechanisms that enrich gene repertoire include gene duplication, gene fusion/fission, horizontal gene transfer, and even the emergence of entirely new genes from previously non-coding regions (Kaessmann, 2010; Neme and Tautz, 2014).

Gene duplication (figure 1.1, A) is one major mechanism that provides gene repertoire variation (Taylor and Raes, 2004). Such events as DNA reparation errors or the activity of mobile genetic elements may duplicate regions containing genes (Zhang, 2003). Duplications of these regions introduce a redundant coding sequence into a genome. Further, the redundant copies can escape the original gene's functional limitations and develop new activities while the original copy retains the same role (Peterson et al., 2009). For example, entire gene families such as myosins or histones were created through this mechanism - each family originates from a single ancestral gene duplicated multiple times, followed by a functional adjustment (Thompson and Langford, 2002; Malik and Henikoff, 2003).

New genes can also originate from existing ones through processes of gene fusion and fission where previously independent genes are joined or separated, respectively (figure 1.1, B and C). Given the combinatory nature of these processes, they play a crucial role in proteome evolution - combining various protein domains leads to the emergence of novel functionality. These processes are known to be significant contributors to protein variability in bacteria (Pasek et al., 2006) and fungi (Leonard and Richards, 2012). Besides, many remarkable genes in the human genome originate from these mechanisms, including Ubiquitin Specific Peptidase 6 (*USP6*) (Paulding et al., 2003) and ATP citrate lyase (*ACL*) (Gawryluk et al., 2015).

Species could also exchange genetic material between each other through horizontal gene transfer. This mechanism has paramount importance for the evolution of unicellular forms

of life. For example, HGT is considered the primary mechanism for acquiring antibiotic resistance in bacteria (Kay et al., 2002; Gyles and Boerlin, 2014). Additionally, this process had a crucial importance in the early evolution of eukaryotes - many mitochondrial and plastid genes have derived from the bacterial endosymbiotic ancestors (Blanchard and Lynch, 2000; Mower et al., 2010). However, for more complex species, it is mainly limited to rare events associated with endosymbiosis and parasitism (Keeling and Palmer, 2008; Xia et al., 2021).

In previous paragraphs, I provided a brief overview of genetic innovation mechanisms based on already existing coding sequences such as gene duplication or fusion. However, new genes may also evolve from ancestrally non-coding DNA sequences (Carvunis et al., 2012) (figure 1.1, D). Non-coding regions occupy most of the genome and provide plenty of raw material for novel transcriptional events. These events lead to an entirely new protein sequence and, therefore, are considered the primary driver of evolutionary innovation (Van Oss and Carvunis, 2019).

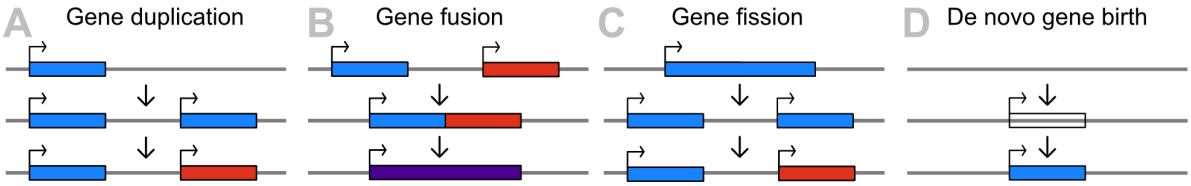

Figure 1.1 Different scenarios of the novel gene emergence

Panel A illustrates a schematic representation of the gene duplication process. After the duplication, a redundant copy can develop new activities. Panels B and C illustrate processes of gene fusion and fission, respectively. These processes could also be followed by functional change. Panel D illustrates a de novo gene birth from non-coding DNA sequences.

1.1.1.3 Gene loss

During evolution, protein-coding genes not only emerge but sometimes they also get lost. Lost gene, also known as a unitary pseudogene, implies the absence of previously functional protein and affects the repertoire of gene functions (Zhang et al., 2010). Loss of protein-coding genes is a radical genetic change that proved to have the potential to affect phenotypic evolution (Sharma et al., 2018). For example, the loss of the egg yolk genes contributed to the development of the lactation mechanism in mammals (Brawand et al., 2008). Moreover, this event played an important role in the phenotypic evolution of birds: instead of the lost teeth, they use beak and muscular gizzard for food collecting and processing and birds have lost six genes essential for the proper formation of dentin and enamel (Meredith et al., 2014).

Gene loss may occur through various mechanisms. The simplest, although not the most widespread, gene loss mechanism is the deletion of the gene-containing locus. Not only the complete deletion is necessary for gene inactivation - if this event affects a significant part

of the coding sequence, it could also be considered a loss-of-function mutation (Hahn et al., 2007). Nevertheless, gene inactivation events usually occur by accumulating loss-of-function mutations.

Loss-of-function (inactivating) mutations such as frameshifting indels or premature stop codons disorganize the reading frame in different manners (figure 1.2). It is doubtful that a gene affected by such an ORF-disrupting mutation encodes a functional protein (Behe, 2010). Such nonfunctional regions resembling protein-coding genes are termed "pseudogenes."

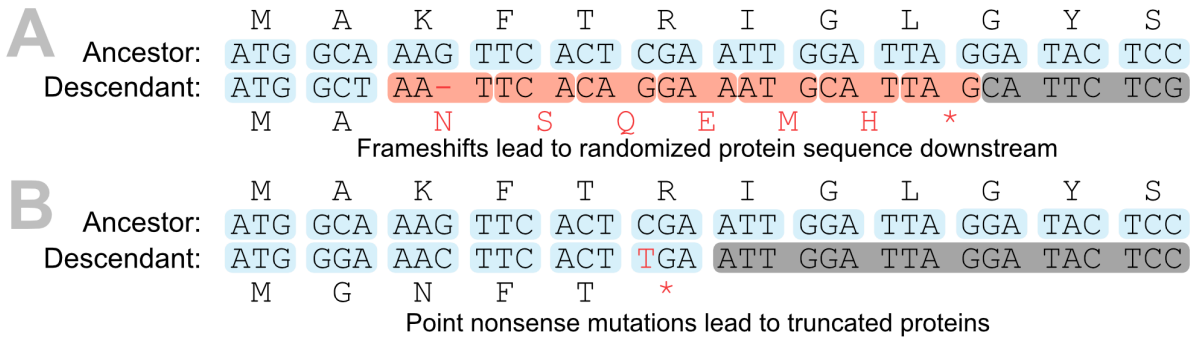

Figure 1.2 Frameshifting and nonsense mutations

The figure shows two pairwise alignments of protein-coding sequences between an ancestor and descendant. Panel A illustrates the consequences of frameshifting mutations resulting in entirely different protein sequence downstream. Panel B illustrates the premature stop codon resulting in protein truncation.

Usually, the following mutations are considered to be gene-inactivating:

1) Frameshifting mutations imply deletions or insertions of the number of base pairs that are not a multiple of three. Given that the ribosome reads the coding sequence nucleotides three-by-three, frameshift leads to a translation of entirely different and randomized amino acid sequences downstream. Moreover, randomized nucleotide sequence contains abundant stop codons, which means that the affected protein sequence is most likely also truncated. There is a chance that a gene corrupted by such mutation is partially or fully inactivated.
2) Point-nonsense mutations that terminate a translation process prematurely, leading to the synthesis of incomplete protein chains. Usually, heavily truncated proteins lack the functionality of the ancestral protein and therefore are inactivated.
3) Disruption of consensus splice sites in multi-exon genes confuses the exon boundaries recognition process leading to incorrect splicing of premature mRNA. It can result in a complete exon exclusion from the mature mRNA, which has the potential of inactivating the gene if the excluded region contributes to the functionality of the encoded protein. Another possible consequence of this mutation is the retention of an intron in the

mature mRNA. This event is expected to be deleterious because it implies the insertion of non-homologous and nearly random sequences in the mature mRNA.

4) Large-scale frame-preserving insertions and deletions could also be deleterious for the functionality of a protein. These mutations diminish the sequence similarity in the affected region. Consequently, it could lead to improper protein folding or corruption of the active center. However, this mutation is less drastic than previously listed ones because such mutations indeed appear sometimes in conserved genes.
5) Loss of start codon might be considered as another loss-of-function mutation, but with particular reservations. Indeed, in the absence of the start site, the translation process is doubtful to be initiated. However, for many genes, it is hard to determine whether it is truly absent. The actual start codon could be shifted outside the locus, where its presence was expected. Therefore, the presumably absent ATG codon could be, in fact, present but not detected.

Any of the mutations enlisted above disrupt the ancestral coding sequence and are likely to reduce the encoded protein functionality drastically or even eliminate it entirely. However, occasionally these mutations can contribute to functional variation. For instance, the Ornithine decarboxylase antizyme 1 (*OAZ1*) encoding gene comprises two overlapping ORFs where the choice of the particular frame is regulated at the translational level by the concentration of polyamines. Specifically, the increased polyamine levels induce programmed +1 ribosomal frameshift resulting in the full-length functional *OAZ1* bypassing a nonsense codon (Kurian et al., 2011).

1.1.2 Identification of mutations connected to phenotype

To recognize the evolutionary events listed above, this is essential to identify protein-coding genes in the considered genomes and comprehensively annotate them. An overview of protein-coding gene annotation methods is provided in part 1.2 of the present work. Another crucial step following gene annotation is the separation of paralogous from orthologous genes since the latter usually utilize the same function (Tatusov, 1997; Altenhoff et al., 2012). Part 1.3 explains the differences between orthologs and paralogs, including an overview of methods to distinguish them.

1.2 Gene annotation

Protein-coding gene annotation is one of the earliest and longstanding challenges in bioinformatics and genomics. Pioneering gene annotation methods have been designed to operate on early bacteria genomes, while the modern methods advanced to annotate more

complex eukaryotic genes and continue to improve (Brent, 2005; Zerbino et al., 2020). Besides, the present work also contributes to gene annotation methods development.

Gene annotation provides the gene repertoire of the analyzed species, which is the starting point of virtually any comparative genomics research. This could include comparative studies revealing the genetic origin of various adaptations within entire clades (Huelsmann et al., 2019; Jebb et al., 2020), research of distinct metabolic pathways (Pizarro et al., 2020), or even the analysis of individual genes in the evolutionary context (Jebb and Hiller, 2018; Kirilenko et al., 2019).

Section 1.2.1 describes the main structural elements of a protein-coding gene that annotation methods aim to identify. Then, section 1.2.2 provides an overview of practically applied techniques of gene annotation. In detail, I covered the following classes: (i) *ab initio* methods, (ii) reference-based approaches, and (iii) transcriptome-based methods.

1.2.1 Eukaryotic gene structure and expression

Broadly, a protein-coding gene is a DNA region coding for the RNA, which consequently encodes the amino acid sequence in a polypeptide chain. Making an RNA copy of a gene sequence is called transcription; the following protein synthesis process is called translation. Transcribed regions comprise two types of sequence: (i) introns and (ii) exons. Introns are excluded from the mature mRNA during splicing, whereas exons constitute the resulting processed mRNA molecule. While eukaryotic genes often have at least one intron, about a third of mammalian genes do not have introns and consist of a single exon.

Furthermore, exons also split between two sequence types: (i) coding sequence (CDS), which actually encodes the polypeptide chain, and (ii) untranslated regions (UTR), those flanking the CDS and regulating the process of translation. CDS is subdivided into triplets of nucleotides, known as codons, where each codon encodes an amino acid or the signal of protein synthesis termination. Typically, CDS starts with methionine-encoding “ATG”-codon, which designates the translation initiation, and ends with one of three stop codons: “TGA”, “TAG”, or “TAA”. This structure of consecutive nucleotide triplets that encode polypeptide chains is called the open reading frame (ORF).

Additionally, non-coding intronic sequences also expose some particular characteristics. The intronic sequence starts and ends with a specific sequence of nucleotides necessary for proper recognition by the spliceosome. For protein synthesis, this is crucial to recognize the boundary between coding and non-coding gene regions accurately. Usually, the intron sequence starts with a "GT" dinucleotide, which determines the 5' splice junction, also known as the donor splice site. At the 3' terminus, the intron sequence typically ends with the "AG" dinucleotide, which defines the acceptor splice site.

The vast majority (about 99%) of introns are flanked with conserved “GT”-”AG” dinucleotides and are recognized by the major U2-spliceosome. However, a minority of introns are spliced by a minor U12-spliceosome and are flanked with alternative sets of dinucleotides. Concretely, U12 introns splice sites can comprise various dinucleotide pairs except for the canonical “GT-AG” pair, including “AT”-”AC”, “GT”-”AG”, “GT”-”GG”, “AT”-”AT”, or “AT”-”AA” (Levine and Durbin, 2001).

Figure 1.3 demonstrates the main structural elements of a multi-exon gene mapped to a hypothetical genomic sequence. Throughout this thesis, I apply the same visual conception to illustrate protein-coding genes consistent with the UCSC genome browser style of data representation. Specifically, the conception implies the following: boxes depict exons, and lines between them represent introns. Arrows covering intronic regions illustrate the direction of transcription and translation. Thinner parts of the exons represent the UTR regions, whereas a thicker part illustrates the CDS.

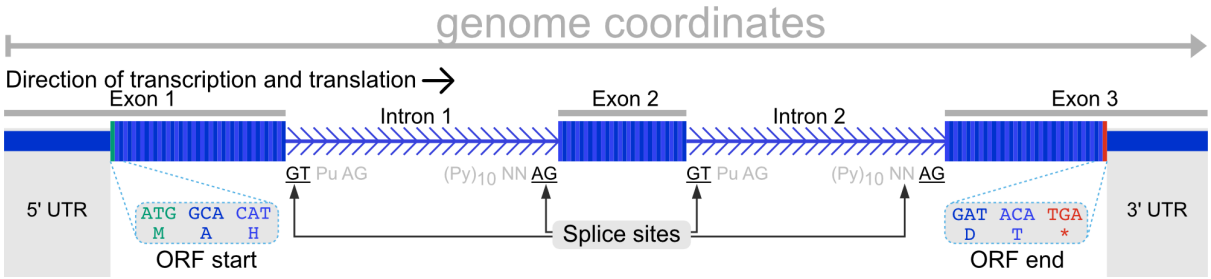

Figure 1.3 Eukaryotic gene structure

The figure shows basic structural elements of a protein-coding gene mapped to genome coordinates. Boxes illustrate exons, whereas lines between them show introns. The thinner fraction of exons represents UTRs. In contrast, thicker fractions represent CDS, and each stripe (different shades of blue) shows an individual codon. The start codon is marked with a green, whereas the red color indicates a stop codon (ORF start and end are magnified). Arrows covering intronic fractions indicate the direction of gene expression. Additionally, this figure shows canonical splice sites flanking introns.

1.2.1.1 Alternative splicing isoforms

It is not compulsory that an individual gene is invariably translated into a single protein. Instead, most eukaryotic genes are usually translated into different proteins because of the alternative splicing process, which is considered the principal mechanism for producing a complex proteome from a limited set of genes (Roy et al., 2013). During the splicing process, not only introns can be excluded from the mature mRNA, but also some exons. This process can produce multiple ORFs from the same set of exons by combining them in various ways, thus increasing the variability of the encoded proteins. This process of building alternative ORFs from a limited set of exons is called alternative splicing (Baralle and Giudice, 2017). Each distinct ORF produced by combining different exons through alternative splicing is called

an isoform or a transcript variant. Protein isoforms can differ in various aspects, such as substrate specificity, protein domain composition, or cellular localization (Yang et al., 2016).

To provide one example, different isoforms of the Fas Cell Surface Death Receptor (*FAS*) gene are produced by an exon-skipping mechanism (figure 1.4). There are two isoforms of the Fas receptor, typically occurring in humans: a longer one that includes exon number six and the alternative, where this exon is skipped. Since the 6th exon encodes a transmembrane domain, the isoforms differ in terms of cellular localization: the protein encoded by the longer isoform, which includes this exon, is membrane-binding, whereas the product of the shorter isoform is water-soluble. Thus, the longer isoform can bind transmembrane ligands, which promotes the apoptosis process. It was shown that the expression of the longer isoform is increased in skin cells chronically exposed to the sun, which suggests that this may be important to eliminate pre-cancerous cells (Hughes and Crispe, 1995).

Summarizing that, annotating a gene implies identifying the locations of all constituting exons in the considered genome. In addition to this, it is necessary to distinguish discrete isoforms of the respected gene. The following section introduces conventional techniques of gene annotation.

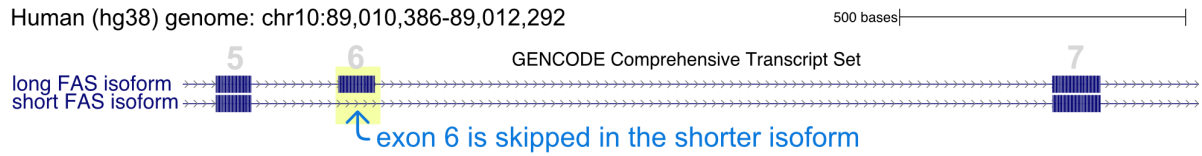

Figure 1.4 Alternative splicing isoforms of the FAS gene

The figure shows a locus in the human genome comprising exons 5, 6, and 7 of the FAS gene. Two FAS gene isoforms naturally occurring in humans are mapped to this region. Exon 6 is skipped in the shorter isoform due to the alternative splicing mechanism.

1.2.2: Overview of gene annotation methods

The gene annotation challenge has a long history, and many different methods and approaches for this problem have been developed up to this date. It is hard to undervalue the significance of gene annotation for bioinformatics and comparative biology. On account of this importance, the scientific community continues to advance the existing techniques, inventing various heuristics to improve the annotation quality (Ejigu and Jung, 2020). Ideally, a comprehensive gene annotation should satisfy the following criteria:

1. Sensitivity: correctly annotate all protein-coding genes and their isoforms existing in the considered species.
2. Specificity: do not annotate regions that do not belong to protein-coding genes in the considered species.

The existing gene annotation techniques could be divided into three major branches, particularly:

1. *Ab initio* methods (subsection 1.2.2.1). These methods exploit various heuristics such as the maximum likelihood to identify the regions that statistically are likely to be protein-coding genes. In general, these methods show high sensitivity but report a considerable number of false annotations.
2. Reference-based methods (subsection 1.2.2.2). These methods employ already existing annotations of closely related (reference) species to detect similar genes in the annotated (query) genome. Reference-based methods have reasonable sensitivity and usually outperform *ab initio* methods in terms of specificity.
3. Transcriptome-based methods. These methods use RNA sequencing data to annotate genomes. The mRNA sequences that designate actually translated regions are mapped back to the genome sequence. These methods provide the best sensitivity; however, they strongly depend on sample collecting quality. These methods are explained in detail in subsection 1.2.2.3

1.2.2.1 *Ab initio* method predictions

Ab initio methods are the most straightforward to use because they require only a genome sequence and a pre-trained model describing characteristic properties of genes. Moreover, there are plenty of available pre-trained models for all major clades. This category includes popular and widely applied methods such as AUGUSTUS (Nachtweide and Stanke, 2019), GeneScan (Burge, 1997), and SNAP (Korf, 2004). In general, these methods are based on the evidence that protein-coding sequences have distinct statistical properties that could be utilized to discriminate them from non-coding regions.

Ab initio methods detect specific coding sequence patterns such as codon and hexamer usage, exon and intron lengths, GC content, nucleotide periodicity, or the compositional bias between codon positions. Additionally, they rely on signal detection, recognizing different ORF features, such as start and stop codons, donor and acceptor splice sites, CpG islands, promoter and terminator sequences (Huang et al., 2016, p.).

These methods typically employ statistical models, such as Hidden Markov Models or Support Vector Machines, to predict potential coding genes. The biggest challenge in applying these methods is the training of probabilistic models, implying the extraction of statistical properties from already annotated genes. However, in-depth details of this problem are beyond the scope of this thesis.

Since these methods do not depend on external evidence, they can identify previously unknown genes or genes that diverged beyond the recognition limits of homology-based approaches. Thus, they provide a convenient approach to obtain an exhaustive genome

annotation. However, this category of annotation methods has certain limitations: they regularly suffer from overprediction, implying the low specificity of the results.

1.2.2.2 Reference-based annotation methods

Reference-based approaches, in contrast to *ab initio* methods, rely on external evidence to annotate genes. In view of the fact that tremendous resources have already been invested in annotating genomes of model organisms, it is reasonable to utilize this information to annotate other species (Yates et al., 2019; Zerbino et al., 2020; Frankish et al., 2021). For example, highly-complete human and mouse genome annotations provide an outstanding source of external evidence. The methods that exploit external gene annotations to identify genes in the analyzed species belong to the reference-based class and are described in this subsection.

The idea behind these methods is that gene sequences necessary for species survival are usually conserved, especially in closely related species (Clark et al., 2019). Hence, similar protein-coding sequences could be identified in different species employing nucleotide alignment tools. This class incorporates methods such as AgenDA (Taher et al., 2003), GenomeThreader (Gremme et al., 2005), TWINSCAN (Korf et al., 2001), and SGP2 (Parra et al., 2003). Typically, those methods align reference protein-coding sequences to the query genome utilizing local sequence aligners such as BLAST (Altschul et al., 1990). The significant matches then are processed using numerous heuristics to annotate the regions that likely represent protein-coding genes.

However, this class of methods has specific caveats. The overall performance of such methods significantly depends on the reference annotation quality. Additionally, the high evolutionary distance between the reference and query genomes reduces the overall method accuracy, therefore for some clades, the search for suitable reference could be challenging. Moreover, the prediction of genes with different properties between reference and the query could be challenging due to reference bias (Chen et al., 2021).

For instance, if annotation of some gene includes only a reference-specific isoform, it involves a risk of incorrect gene annotation. Furthermore, annotated genes are limited to reference gene homologs - query-specific genes are obscure for these methods. In contrast to *ab initio* methods, homology-based methods ensure that the number of false discoveries would be minimal, especially if the reference and query species are close relatives.

1.2.2.3 Transcriptome-based approaches

As opposed to previously introduced methods, this class consists of very reliable approaches supported by experimental data. The idea behind these methods is that messenger RNA, which is usually translated from coding regions, can be sequenced and then

aligned back to the genomic DNA sequence to recognize exons of a gene. This approach allows revealing the exon-intron structure of the genes that are indeed expressed in the organism, approaching nearly perfect specificity (Kuo et al., 2020).

However, those methods come with several conditions. First, some genes are expressed at shallow levels or are strictly tissue-specific, so there is a high chance that RNA-sequencing will not capture them (Wang et al., 2016). To capture all genes actually expressed in the considered species, the RNA-seq data should be genuinely comprehensive and cover all possible tissues and development stages, which could be highly complicated in practice for most species. Second, sequenced RNA might represent incorrectly spliced transcripts, NMD-targets, degraded RNA, or to be a product of background translation (Minoche et al., 2015).

Moreover, these methods are much more expensive and challenging to apply than the other methods listed in this subsection. As for extinct species, acquiring this data appears to be nearly impossible. Nevertheless, only the experimental support can theoretically provide the ideal annotation quality.

Although, the common practice to produce reliable and comprehensive annotations is combining independent methods (Haas et al., 2008). This practice allows to neutralize the limitations of each particular method and increase the overall gene annotation performance. Moreover, the annotation of the most critical genes can be curated manually or involve additional experiments. The subsequent challenge after protein-coding gene prediction is to identify homologous genes that share common evolutionary ancestry and classify them into orthologs and paralogs, as explained in the following part.

1.3 Gene homology

Homology is a property that describes evolutionary history. Orthologs and paralogs are two fundamentally different types of homologs diverged by speciation and duplication events, respectively. Proper identification of orthologs is crucial for various comparative studies because orthologs are generally assumed to carry out biologically equivalent functions in different organisms (Tatusov, 1997; Altenhoff et al., 2012). In contrast, paralogs can functionally diverge after duplication and therefore alter the functionality.

This part gives a detailed explanation of orthology and paralogy concepts (section 1.3.1) and provides an overview of modern techniques to distinguish orthologs from paralogs (section 1.3.2).

1.3.1 Differences between orthologs and paralogs

Orthology and paralogy concepts focus on the mode of descent from their common ancestor (Fitch, 1970; Fitch, 2000), not on the level of sequence or function conservation. However, sequence similarity is a plausible indicator of common ancestry (Pearson, 2013). Orthologs are a subset of homologs that evolved from the same ancestral sequence and are separated by a speciation event. It means when a species diverges into two separate ones, the descendants of the single ancestral gene remaining in these species are called orthologs (Koonin, 2005). In contrast to *orthologs*, *paralogous* sequences are separated by a duplication event that occurred within the ancestral genome. Sequences diverged from different copies are said to be paralogous (Koonin, 2005). The figure below illustrates these concepts on a theoretical gene tree (figure 1.5).

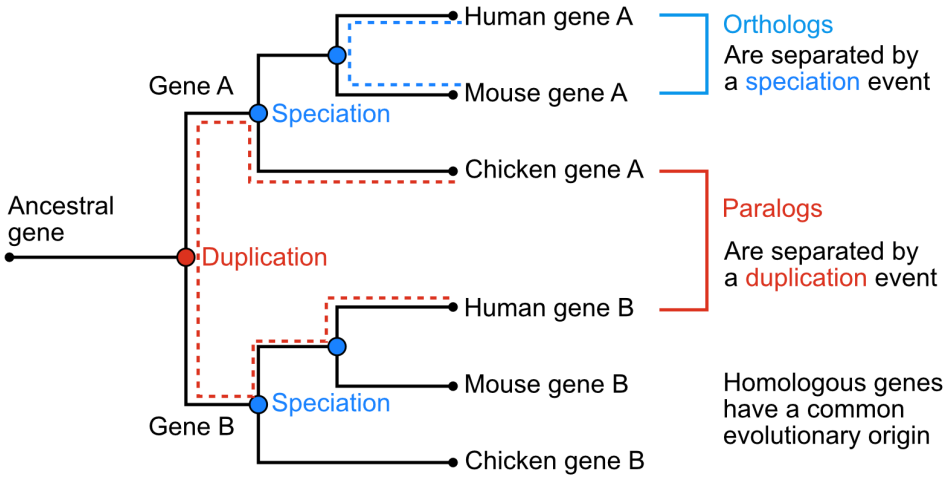

Figure 1.5 Difference between orthology and paralogy

In this particular example, a hypothetical ancestral gene is duplicated within the ancestor of mammals and birds. The resulting copies of this duplication are called A and B. Then, after two speciation events, we may detect A and B copies in the exemplified species. The path between the human and mouse copies of gene A goes through the speciation event (shown in blue); thus, these genes are orthologous. However, the evolutionary path between the chicken copy of gene A and the human gene goes through the gene duplication in the ancestral genome (shown in red); therefore, these genes are said to be paralogs.

It is worth considering a slightly more complicated evolutionary structure. The case illustrated above (figure 1.5) represents the case of one-to-one orthology, implying that each considering species has a single copy of the ortholog. However, the duplication event might happen after the speciation, as exemplified in Figure 1.6.

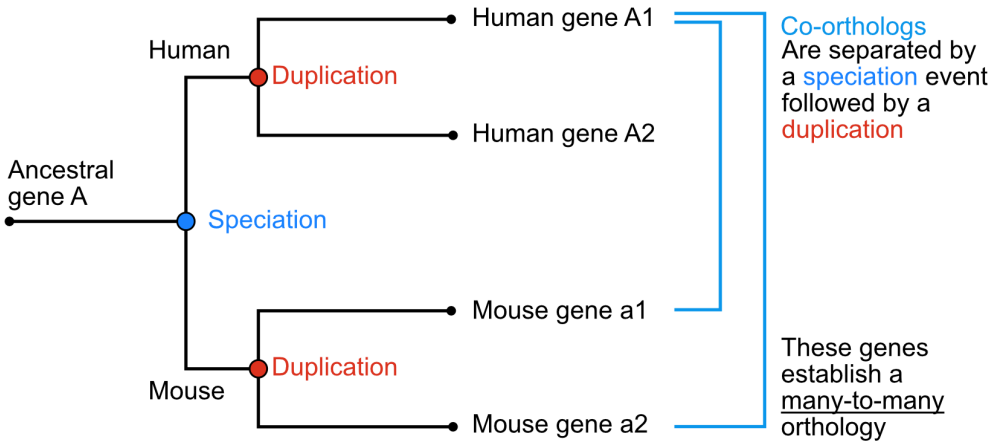

Figure 1.6 Example of co-orthology

The phylogenetic tree shows a theoretical gene A that underwent independent duplications after the speciation event. In such a case, mouse genes a1 and a2 are said to be co-orthologs of human genes A1 and A2.

In this example, a hypothetical gene underwent independent duplications in the human and mouse lineages. In this case, mouse genes a1 and a2 are said to be co-orthologs regarding each human gene, and vice versa. Together, these genes establish a many-to-many orthology connection, or comprise an orthologous group (Gabaldón and Koonin, 2013).

Notably, the applicability of orthology and paralogy concepts is not restricted to protein-coding regions. In fact, it applies to any class of sequences that allows the determination of evolutionary relationships. For instance, that could be non-coding RNA regions (Bryzghalov et al., 2019) or regulatory elements (Hallikas et al., 2006). One may potentially detect the same evolutionary origin and inspect for such sequences, whether they arose because of speciation or duplication events, and therefore classify them as orthologs and paralogs. However, this work principally focuses on protein-coding genes. In the following section, I outline the principal methodological approaches of orthology inference.

1.3.2 State of art methods to distinguish orthologs from paralogs

During the last decades, plenty of approaches for inferring orthology were developed. However, they have a common feature - as input, they require a set of already annotated genes with their coding and protein sequences for each considered species. This observation explains why **gene annotation generally precedes orthology inference.** Moreover, gene annotation quality has a substantial influence on the accuracy of the orthology inference process. In general, the assortment of orthology inference methods could be divided into two

major types: graph-based methods, introduced in subsection 1.3.2.1, and tree-based approaches, described in subsection 1.3.2.2.

1.3.2.1 Graph-based methods

Graph-based, also known as sequence-similarity methods, are built on the assumption that the protein sequences originated from orthologous genes exhibit a higher degree of sequence similarity since they are by definition more closely related. These methods cluster annotated genes into groups based on pairwise sequence similarity to employ this concept. This approach is implemented in a great variety of methods, such as OrthoFinder (Emms and Kelly, 2015), OrthoMCL (Li, 2003), and OrthoVenn (Wang et al., 2015).

Usually, graph-based methods use techniques such as BLAST search to identify gene pairs having the highest sequence identity. As a result of this search, a graph-based approach generates a graph where nodes stand for genes, and edges characterize sequence similarities between genes. Then, clustering techniques are applied to segregate groups of orthologous genes. However, the premise that orthologs always show substantially higher sequence similarity than paralogs is not always correct. Therefore, to increase the accuracy, most of these methods further include post-processing steps involving various heuristics.

1.3.2.2 Tree-based methods

Tree-based methods implement the classical approach for orthology identification. Methods of this category determine whether a pair of genes coalesce in speciation or a duplication node in the gene tree (Zmasek and Eddy, 2001; Li, 2003; van der Heijden et al., 2007; Huerta-Cepas et al., 2007; Vilella et al., 2008; Emms and Kelly, 2015; Herrero et al., 2016). To separate orthologs and paralogs, these methods construct a tree of considered genes, using various approaches such as maximum parsimony (Felsenstein, 1978), maximum likelihood (Vandamme, 2009), or Bayesian algorithms (Yang and Rannala, 2012). Then, they map the resulting gene tree to an already established phylogeny of species that host the analyzed genes. Since the gene and species trees have different topologies due to evolutionary events performing particularly on genes, such as duplications or losses, the problem of orthology inference is reduced to the reconciliation of combined phylogenetic trees. Typically, these methods search for orthology/paralogy configuration that could be explained by the smallest number of evolutionary events to reconcile the tree (Altenhoff et al., 2019).

It is worth mentioning that gene tree-based algorithms could be confused by evolutionary events such as reciprocal gene deletion (Gabaldón, 2008). Figure 1.7 provides a hypothetical example of such an event. In this case, the deletion of mouse gene A and human gene B may cause the ancestral duplication event undetected, consequently leading to the erroneous assignment of human gene A and mouse gene B as orthologs.

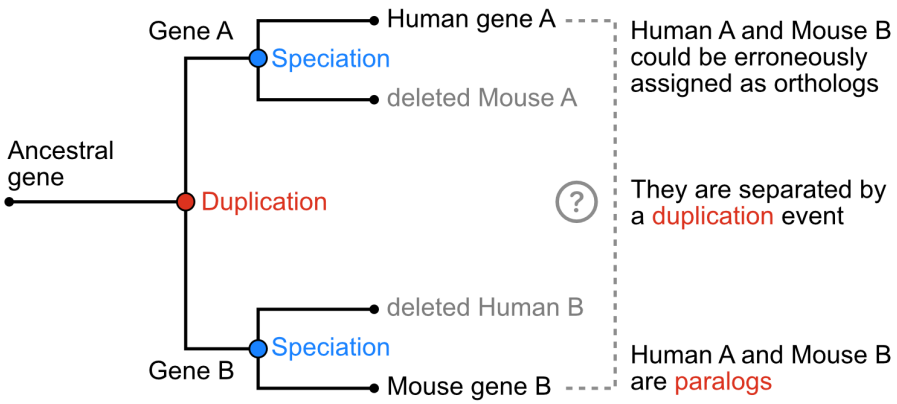

Figure 1.7 Gene deletion leading to inaccurate orthology inference

In this case, human genes A and mouse gene B are split by a duplication event that happened in the ancestor, and therefore are paralogs. However, since the orthologs of these genes that arose by speciation event (human gene B and mouse gene A) are deleted, the duplication event can be undetected by a tree reconciliation algorithm. Consequently, human A and mouse B genes can be erroneously classified as orthologs.

The methods of orthology inference introduced above established themselves as reliable approaches with discovered advantages and downsides. The orthology connections found using these methods shaped the comparative biology research of the last decades. However, nowadays, they are subjected to novel challenges, which are described in the next part.

1.4 Major challenges in the field

On account of advances in DNA sequencing technologies, the number of available genome assemblies rapidly grows in the last years. Indeed, for certain clades such as mammals (appendix B) there are several hundred genomes available nowadays. The scientific community encountered unprecedented challenges in adapting the existing methodology to meet the increased data scales (Sonnhammer et al., 2014; Nevers et al., 2020). Moreover, the requirements for genome annotation quality and completeness have been increased (Salzberg, 2019). To meet the challenges related to the latest large-scale genome sequencing projects such as the EBP (Lewin et al., 2018), we crucially need the next generation of high-throughput annotation methods that scale to the abundance of data.

It is worth considering that currently applied orthology inference and gene annotation approaches were developed in entirely different circumstances, where smaller amounts of data were accessible. Moreover, the overall quality and completeness of data have significantly improved since then. Most of the underlying techniques are adapted to process

dozens of genomes and require an unreasonable amount of computational resources when it comes to hundreds of them. Thus, they lack scalability and cannot meet the increased data amounts. Thus, the bottleneck of comparative genomics is shifting from the amount of available assembled genomes to the throughput of computational methods.

In this work, I present TOGA (Tool to infer Orthologs from Genome Alignments) - a bioinformatics method that implements a novel paradigm of genome annotation and orthology inference at scale. In contrast to other methodologies, this paradigm does not exclusively rely on the protein-coding sequences to infer orthology. Instead, it uses whole-genome alignments to integrate detecting intact or lost genes, determining gene orthology, and annotating orthologous genes in a comparative manner. In contrast to current orthology-detection methods that rely on the similarity between gene sequences, TOGA extracts rich information on how the genomic context around genes aligns between species and uses machine learning to distinguish orthologs from paralogs accurately.

Exploiting the whole genomic context, we can approach the classic challenges from new directions, and the proposed design demonstrated solid performance compared to hi-end techniques (see part 3.4). Additionally, the TOGA method can be effortlessly scaled to hundreds of genomes. Background on additional concepts required by TOGA is explained in the following part (1.5), while the TOGA method itself is described in chapter 2.

1.5 TOGA method novelty

To the best of my knowledge, TOGA is the first method that integrates inferring orthologs and annotating genes in a single pipeline. Historically, these two steps were considered separate since orthology inference methods require already existing gene annotation. Instead, TOGA solves the problem in the reversed order. First, TOGA accurately identifies orthologous loci in the annotated genome using whole-genome alignments. To infer orthology ahead of gene annotation, it employs divergence of neutrally evolving sequences such as introns and intergenic regions, in contrast to other methods that rely exclusively on coding sequence alignments. The idea behind this is that neutrally evolving sequences in orthologous regions tend to be less diverged than in the paralogous loci. As shown below (part 3.1), the utilization of non-coding regions of a gene enables TOGA to infer orthologs at high accuracy. Second, TOGA realigns reference genes to predicted orthologous loci using the HMM-based method CESAR (Sharma et al., 2016; Sharma et al., 2017), and third, eventually resolves the orthology relationships using the graph method.

This design provides TOGA various advantages compared to other methods. For instance, previous orthology inference methods are insufficient for genes that have virtually identical protein sequences. For TOGA, which uses additional information, this is not a

problem. In addition to this, the TOGA pipeline is able to detect and annotate lost genes, representing orthologous but inactivated (non-functional) genes. Also, TOGA can identify orthologs in highly fragmented assemblies where genes are often split between different scaffolds.

In the following sections, I provide an explanation of how exactly the whole-genome alignments can highlight the orthologous loci in the query genome. First of all, I introduce the approach of genome alignment representation that TOGA actively uses - the genome alignment chains. Then, I show the differences between chains that indicate alignments to orthologous or paralogous loci in the annotated genome.

1.5.1 Pairwise genome-wide alignments

Alignments between entire genomes are the foundation for most comparative genomics analyses, and a key input for TOGA. Pairwise genome alignments establish the correspondence between different regions in the reference and query genomes based on the sequence similarity. In other words, they provide a collection of local alignments distributed along with the entire genome sequences. Usually, these alignments connect homologous sequences such as genic regions or regulatory elements between each other.

Building genome alignments is a computationally heavy task, and many various methods to solve it are currently available. For instance, BLAT (Kent, 2002) is a famous and fast local alignment tool that could be applied to produce whole-genome alignments. Also, BLASTZ (Schwartz, 2003) and its successor LASTZ (Harris, 2007) are local aligners adjusted to provide a higher sensitivity than standard BLAST (Ma et al., 2002). Similarly, LAST (Kielbasa et al., 2011) could compete with other methods applying numerous heuristics to approach the genome alignment challenge. The resulting genome alignments could be represented in many different ways and formats. To perform the computations, TOGA utilizes chains (Kent et al., 2003) of collinear local alignments described in the following section.

1.5.2 Alignment chains introduction

Chains of co-linear local alignments are a form of pairwise genome alignment representation. Local alignments, also known as aligning blocks that occur in the same order and orientation (strand) in both reference and query genome, are chained together to build alignment chains (Kent et al., 2003). Aligning blocks can be separated by insertions and deletions that occurred in either the reference or query genome. Also, blocks can be separated by regions where the reference or the query genome underwent a deletion, inversion, or translocation. In case the local alignments of the rearranged regions are not co-linear with the primary chain, they can form a separate one.

Each individual chain establishes the one-to-one correspondence between aligned regions in the reference and query genome. Therefore, having a region in the reference overlapped by an aligning block, it is possible to obtain the aligned region's coordinates in the query. Technically, the chain consists of the following elements:

1) Aligning blocks representing an aligned ungapped region between the reference and the query. Since an alignment block establishes an explicit connection between two genomes, it has two sets of coordinates: in the reference and the query genomes, respectively.

2) Gaps between blocks representing either a not-aligned region or a deletion of the corresponding region in the query genome.

Figure 1.8 below shows the main elements of the alignment chain. The visualization style is consistent with the UCSC genome browser (Kent et al., 2002) representation.

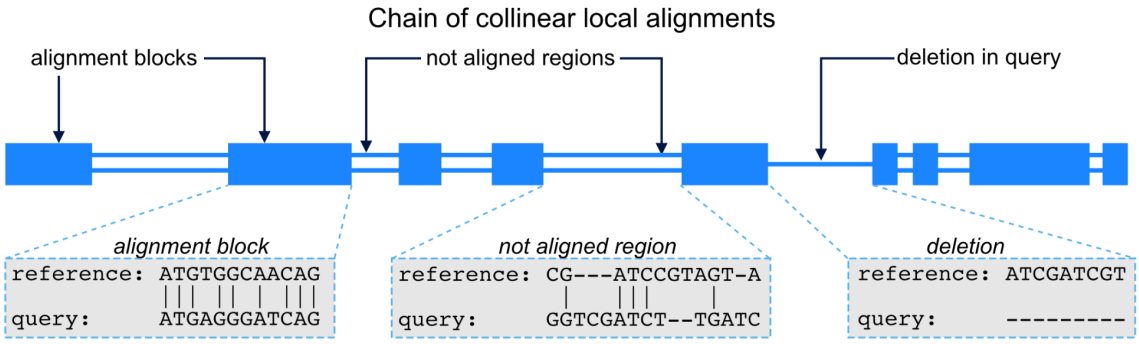

Figure 1.8 Alignment chain structure

Each box represents a non-gapped alignment between the reference and query genome (alignment block). Additionally, double lines between blocks indicate that corresponding regions do not align between reference and query genomes. Furthermore, a single line shown between aligning blocks signifies that the corresponding region is deleted in the query (or inserted in the reference; a pairwise comparison cannot distinguish between both).

To make the concept of alignment chains clearer, figure 1.9 below illustrates a genome alignment between two theoretical genomes represented by a chain. It shows that alignment blocks establish associations between genomic regions in the reference and query based on the sequence similarity. The single line implies that regions corresponding to flanking alignment blocks are adjacent in the query genome. Alternatively, they might be considered as non-homologous insertions in the reference genome from the query perspective. Not aligned regions may have different lengths in the reference and query; therefore, they do not determine reciprocal connections between genome coordinates unequivocally like the alignment blocks.

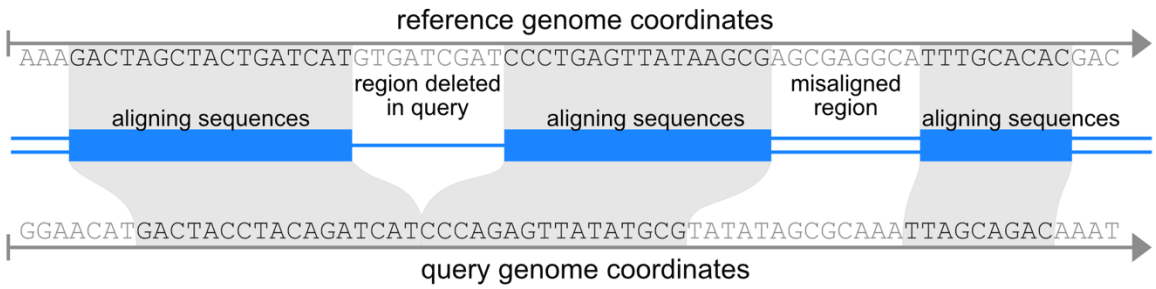

Figure 1.9 Alignment chains on genomic sequences

The figure shows two loci in a reference and query genomes and a chain representing alignment between these regions.

Moreover, a single region in the reference genome could be aligned to several regions in the query genome. In this case, multiple chains cover this region simultaneously, pointing to different aligned loci, revealing that the locus underwent a duplication. Furthermore, alignment chains can represent other evolutionary events such as translocations or inversions, as shown in Figure 1.10.

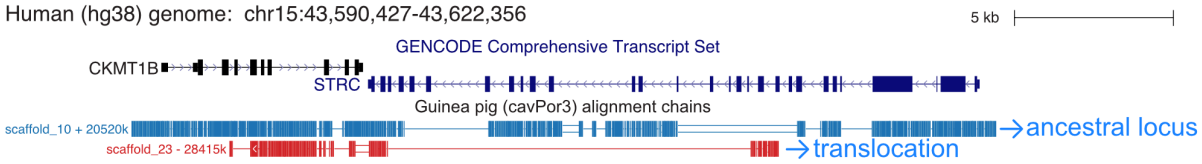

Figure 1.10 Chains indicating translocations

UCSC genome browser screenshot shows the locus in the human genome containing CKMT1B and STRC genes. Also, it illustrates two alignment chains to the guinea pig genome. The top-level chain (blue) aligns to the ancestral locus in the guinea pig genome. The second-level chain (red) is shorted and indicates a translocated region in the guinea pig. Thus, multiple chains can align to the same locus in the reference genome.

1.5.3 Differences between orthologous and paralogous alignment chains

In general, for related species where neutrally-evolving regions are still partially alignable, the key distinguishing feature of chains representing alignments between orthologous loci is that alignment blocks cover not only conserved coding exons but also some intronic and intergenic regions that typically evolve neutrally. Additionally, such chains are often characterized by conserved gene order (synteny): orthologous chains tend to cover multiple consecutive genes because orthologs usually appear in the conserved order. In contrast, alignment blocks of non-orthologous chains mainly cover the regions that evolve under purifying selection, such as exons or regulatory elements. Usually, such chains illustrate that neutrally evolving regions are misaligned.

To illustrate these characteristic differences of chains representing orthologous and non-orthologous alignments, Figure 1.11 shows a UCSC genome browser screenshot of a locus in the human genome that contains the *EHD1* gene. This gene belongs to a conserved gene family of *EPS15* homology domain-containing proteins, which also includes *EHD1* paralogs, named *EHD2*, *3*, and *4*. The top annotation track on this screenshot illustrates the exon-intron structure of three isoforms of the *EHD1* gene. A track below shows several alignment chains corresponding to different loci in the mouse genome. However, only one of these chains aligns to the ortholog in the mouse genome (*Ehd1*). This chain is vividly exhibited contrary to the rest by aligning blocks intersecting both exons and introns. The rest of the chains show that respective loci in the query align only to exons - they represent other genes belonging to the *EHD* family, which are clearly paralogous. Also, one of the chains shows that in the corresponding locus, introns are completely deleted - this indicates that this chain aligns to a processed pseudogene copy.

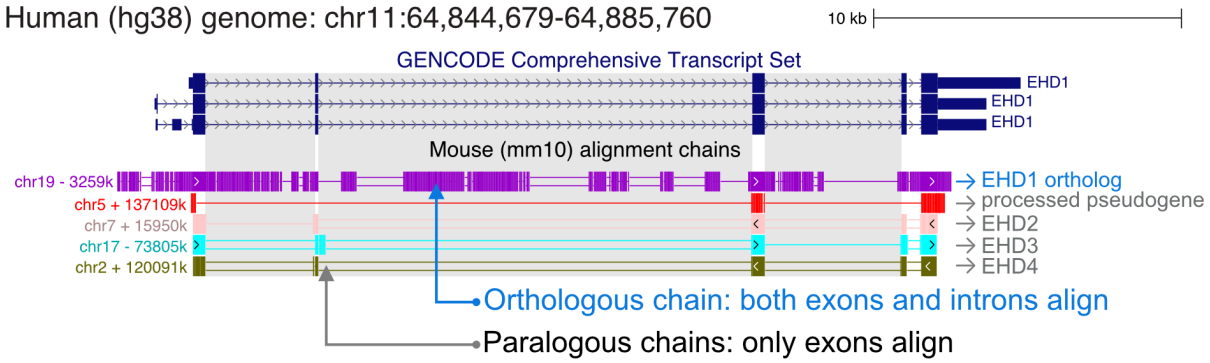

Figure 1.11 Differences between orthologous and paralogous chains

UCSC genome browser shows the locus in the human genome comprising the EHD1 gene. Also, it shows 5 chains representing alignments to different loci in the mouse genome. The top-level chain represents the alignment to the mouse Edh1 gene (ortholog). This chain shows that (i) exons, (ii) introns (highlighted), and (iii) intergenic regions align between orthologous regions in the human and mouse. Other chains show that corresponding loci in the query genome align only to exons, and therefore, they represent non-orthologous regions in the mouse: paralogs and processed pseudogene copies.

The differences between the appearance of orthologous and paralogous chains are explained by molecular divergence. The molecular distance between orthologous genes is substantially shorter than between paralogs. For instance, the speciation event that separated *EHD1* genes in the human and mouse happened ~96 Million years ago (Nei et al., 2001). However, since both human and chicken have an *EDH1* and *EDH2* gene, the duplication that separated *EHD1* and *EHD2* happened in the Vertebrata ancestor, at least 450Mya.

In molecular terms, neutrally evolving sequences in orthologous genes did not have enough time to diverge strongly. Contrary, the molecular distance between paralogs is

substantially larger. More precisely, we expect 0.5 substitutions per neutral site between orthologous loci in the human and mouse genomes (Hiller et al., 2013), which implies approximately 50% of neutral sequence identity. Practically, it suggests that some parts of intronic and intergenic regions will still align.

In contrast, we expect much more than one substitution per neutral site between paralogous sequences, which means that neutral regions are fully randomized and therefore do not align. Summarizing that, we expect that neutral sequence similarity remains only for orthologous sequences.

Figure 1.12 illustrates this principle on an *EDH1/2* gene tree. Split between *EDH1* orthologs in the human and mouse happened simultaneously with the speciation event. However, the evolutionary distance between paralogs is significantly longer than between orthologs because the path between paralogs crosses the duplication event that happened in the vertebrates' ancestors.

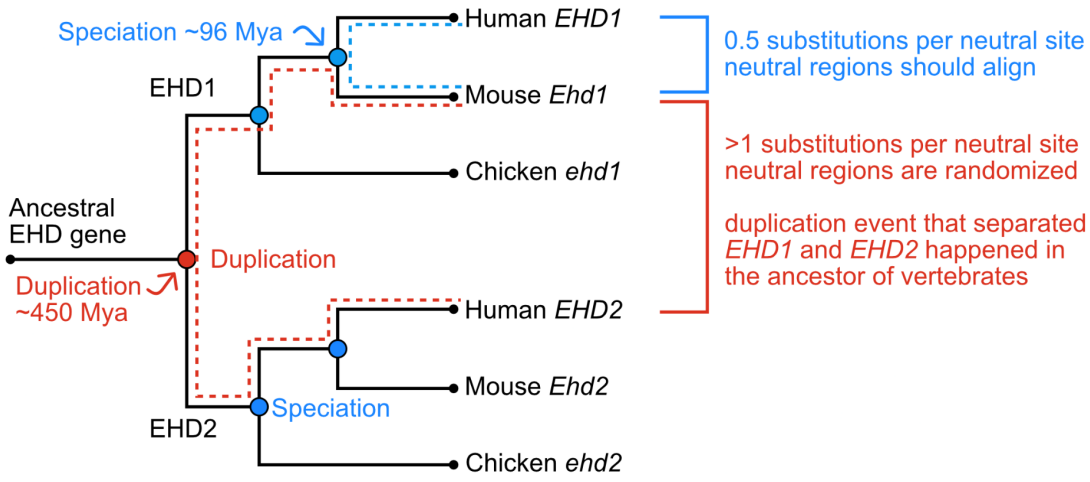

Figure 1.12 Molecular divergence explains the differences between orthologous and paralogous chains

Split between EDH1 orthologs in the human and mouse happened simultaneously with the speciation event (~96Mya, blue dotted line). Thus, we expect 0.5 substitutions per neutral site - introns and intergenic regions still align between orthologous loci. However, the evolutionary distance between paralogs is significantly longer than between orthologs because the path between paralogs crosses the duplication event that happened in the vertebrates' ancestors (~490 Mya, red dotted line). Therefore, neutral sequences in paralogous loci are randomized and do not align.

1.5.4 TOGA conceptual idea

TOGA captures these characteristic differences that separate orthologous from paralogous chains in features that measure e.g., alignment coverage of introns, alignments of gene-flanking regions, and synteny. In a machine learning framework, these features are used to classify each chain as orthologous and non-orthologous. This concept enables TOGA to recognize orthologous loci with a great degree of accuracy.

It is worth mentioning that in contrast to traditional methods, TOGA does not rely on coding sequence alignment features - neutral sequence divergence is sufficient enough to identify orthology accurately. As demonstrated in the following chapters, TOGA can compete with and even outperform traditional approaches that rely on coding sequence alignments and gene trees.

2. TOGA pipeline

This chapter covers the details of the TOGA computational pipeline implementation. Briefly, the genome annotation pipeline implemented in TOGA consists of the following steps:

1. At first, TOGA checks the input data (described in part 2.1) for correctness, and if it does not contain any mistakes, it continues the algorithm execution. For each coding gene annotated in the reference genome, TOGA detects alignment chains that intersect it and extracts numeric features that describe the appearance of alignment. Then it applies a gradient boosting binary classification algorithm to determine chains that represent orthologous loci in the query genome. This procedure is explained in part 2.2. Then, part 2.3 introduces the TOGA annotations naming convention.
2. To recognize all coding exons of a reference gene in the corresponding orthologous loci, TOGA uses the CESAR2.0 method (Sharma et al., 2016; Sharma et al., 2017). In parallel, TOGA scans the predicted reading frame for inactivating mutations and determines whether any exons are missing due to assembly gaps. Then it classifies the predicted transcripts as intact, missing, or inactivated. This pipeline step is detailed in part 2.4.
3. In the end, TOGA infers the orthology type between genes and resolves spurious N-to-M relationships that are only supported by weak orthology. Part 2.5 contains the details of this pipeline step.

The TOGA pipeline steps listed above are briefly illustrated in figure 2.1. Later in this chapter, parts 2.6 and 2.7 outline the pipeline output and a procedure of reference gene set filtering, respectively.

The pipeline is implemented in Python and C languages and requires a minimal number of external dependencies, which improves the tool accessibility. Feature extraction and realigning reference genes in the query genome are heavy computational tasks recommended to be performed on an HPC system. To handle cluster-dependent steps, TOGA uses Nextflow, which maximizes the set of compatible HPC systems.

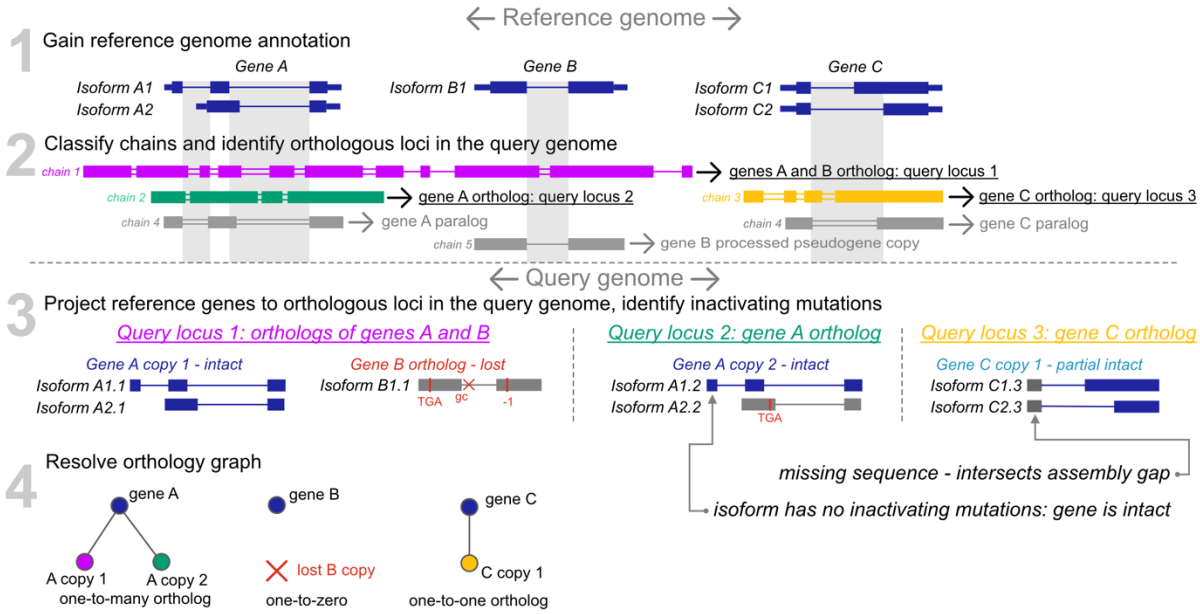

Figure 2.1 TOGA pipeline overview

1: As input, TOGA accepts reference genome annotation; several isoforms might represent each gene. 2: Given alignment chains, TOGA identifies those that likely represent orthologous loci (3 orthologous chains shown in different colors, non-orthologous chains are shown in grey). 3: TOGA projects reference genes to predicted orthologous loci and scans predicted reading frames for inactivating mutations. Then it classifies predicted genes as intact, missing, or lost. Note that TOGA does not annotate UTR regions, only the CDS. 4: At the end, TOGA performs the final orthology inference, establishing orthologous connections between reference and query genes.

2.1 TOGA pipeline input

The genome annotation pipeline implemented in TOGA requires the following files as input:

1) Reference and query genome sequences. The TOGA implementation presented in this thesis requires the genome to be provided in the 2bit format (https://genome.ucsc.edu/goldenPath/help/twoBit.html), containing a compressed and indexed sequence of the entire genome. Users can easily convert a genome in a multi-fasta format to a 2bit using the faToTwoBit program from the UCSC genome browser toolset.

2) Annotation of coding genes in the reference assembly in bed-12 format. If reference genome annotation provides more than one isoform for a gene, TOGA also considers this if the appropriate file is provided. Additionally, a user may provide information about U12 introns localization as a part of reference genome annotation, which is non-mandatory but recommended. The quality of reference genome annotation has a significant input on the TOGA output. Part 2.7 of this chapter provides information on how the annotation could be filtered to avoid input-related errors.

3) Genome alignment file containing chains of collinear local alignments between the reference and query genomes - this file could be extracted from any pairwise genome alignment. It is worth mentioning that chain quality influences the orthology loci classification step. Part 3.5 of chapter 3 discusses this influence.

Compared to other genome annotation methods, the required set of input files for TOGA is minimalistic. Nevertheless, it provides a sufficient amount of information to perform all pipeline steps with a high degree of accuracy.

2.2 Inferring orthologous loci from pairwise genome alignments

To infer orthologous loci, TOGA uses pairwise chains of collinear local alignments, computed between a reference and query genome (see part 1.5), and the gene annotation of the reference genome. Each gene-chain pair uniquely determines a single locus in the query genome. At the first step, TOGA identifies candidate chains for each reference gene, then extracts numeric features from each gene-chain pair, and at the end applies machine learning to distinguish orthologous loci in the query (figure 2.2).

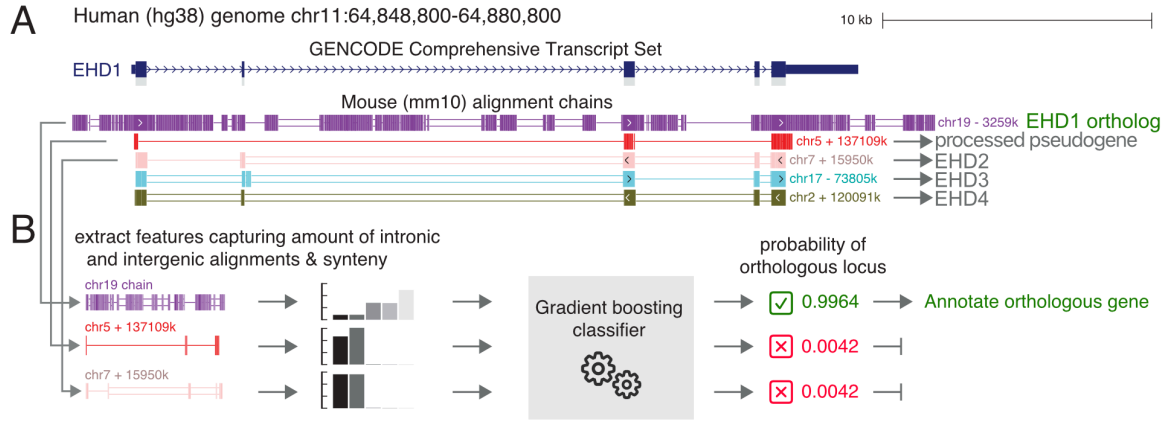

Figure 2.2 Inference of orthologous loci

UCSC browser screenshots show a region in the human genome containing the EHD1 gene. A: Five mouse chains align to this region. B: TOGA extracts characteristic features for each chain and applies a gradient boosting model to identify orthologous chains. For the human EDH1 gene, the top-level alignment chain represents the alignment to the orthologous locus (on chromosome 19). TOGA then annotates the ortholog in the respective locus in the query genome.

2.2.1 Identifying candidate chains

At the very beginning, for each provided reference transcript, TOGA identifies intersected alignment chains. This procedure is necessary to establish the set of potential gene-chain pairs for further classification and analysis. Since a naïve algorithm that iterates

over all possible gene-chain pairs and checks whether they intersect is time-consuming, TOGA implements a faster method that utilizes the sorting of gene and chain regions on the same chromosome:

1. Specifically, on each reference chromosome/scaffold, TOGA sorts the genomic regions of all genes and chains by the start coordinate.
2. Then, for each chain, TOGA iterates over the sorted list of genes, starting with the first gene that intersected the previous chain. Since the list of genes is sorted, all genes that precede the selected one indeed do not intersect the considered chain. Therefore, they can be excluded from the computation, which saves computational time.
3. We determine whether the chain overlaps or spans at least one coding exon for each gene, making this chain a candidate chain. The iteration is stopped at the first gene that starts downstream of the current chain end. Again, since the genes' list is sorted, all following genes will start upstream of the current chain end and, consequently, indeed do not intersect the considered chain.

As a result of this algorithm, we quickly acquire a list of chains that intersect each gene. As well as the naïve approach, this procedure also has an asymptotic worst-case $O(N^2)$ runtime. However, for the naïve approach, the worst and average case scenarios are identical: it iterates over all possible gene-chain pairs, and the runtime only depends on the number of genes and chains. Our optimized approach has the quadratic runtime only in the worst case where every chain overlaps every gene. In practice, we found that this procedure results in a speedup of ~70 fold because it avoids examining numerous upstream or downstream genes of a considered chain.

2.2.1.1 Filtering low-scored chains

It is also worth mentioning that alignment chains undergo the filtering procedure to spare computational runtime further. Genome alignment tools do not take into account the function of the aligned sequence. Therefore, alignment chains may appear in any fraction of the genome unless they are masked. Thus, the vast majority of chains are very short and cover minor regions of a few hundred bp long and are characterized by low alignment scores. It can be expected that most of these chains are very unlikely to align any significant fraction of a protein-coding gene. However, extracting features from each chain requires computational resources. To avoid unnecessary computations, TOGA considers only the chains with any prospect to align with protein-coding regions in the query genome.

In particular, TOGA first removes chains with alignment scores below 5000 and then selects those that span at least one coding exon for a given protein-coding transcript (figure

2.3). This filter allows TOGA to optimize the runtime because the low-scored chains comprise up to 90% of the total number of chains and do not contribute to the pipeline results.

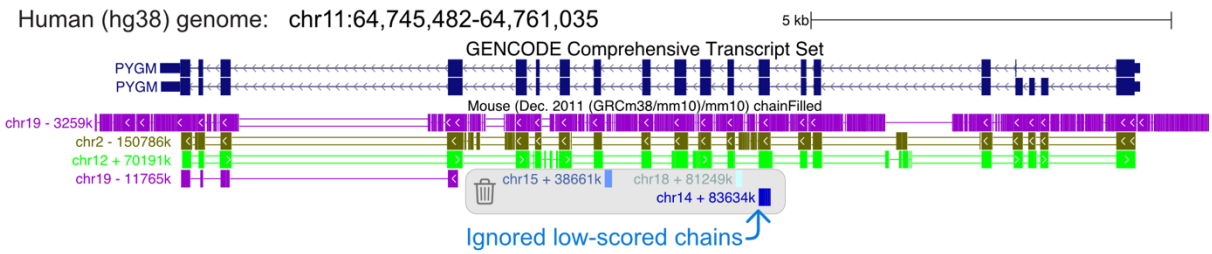

Figure 2.3 Low-scoring alignment chains

Chains that are shown here on the grey background are ignored as they are unlikely to represent complete transcripts.

2.2.2 Features extraction for machine learning

At this step, TOGA computes numeric features describing alignment appearance for each gene-chain pair. To compute these features, it intersects the reference coordinates of chain aligning blocks with different gene regions, such as CDS, UTRs, and introns. In particular, TOGA extracts the following values:

- c: number of reference bases in the intersection between chain aligning blocks and CDS of the analyzed gene.
- C: number of reference bases in the intersection between chain blocks and coding exons of all genes covered by the analyzed gene. If the chain overlaps a single gene, this value is equal to "c."
- a: number of reference bases in the intersection between chain blocks and coding exons and introns of the gene under consideration.
- A: number of reference bases in the intersection between chain blocks and coding exons and introns of all genes and intergenic regions covered by the considered chain. It is worth mentioning that this value could not be equal to value "a" even if the chain covers a single gene because value a does not include intergenic regions.
- i: number of reference bases in the intersection between chain blocks and introns of the gene under consideration.
- I: the sum of all intron lengths of the gene under consideration.
- CDS: length of the CDS of the gene under consideration.
- f: number of reference bases in chain blocks overlapping the 10 kb flanks of the gene under consideration. Alignment blocks overlapping exons of another gene located in these 10 kb flanks are ignored.

Using these values, TOGA computes the following characteristic features:

- "global CDS fraction" as C / A. Chains with a C / A value closer to one have alignments that largely overlap coding exons, which is a signature of paralogous or processed pseudogene chains. In contrast, chains with the C /A value closer to 0 also align many intronic and intergenic regions, which is a hallmark of orthologous chains.
- "local CDS fraction" as c / a. Orthologous chains tend to have this value closer to 0, as intronic regions partially align. For paralogous chains, the c /a value is usually closer to 1. This feature is not computed for single-exon genes since they have no intronic fraction.
- "local intron fraction" as i / I. This feature shows the fraction of aligned introns. Orthologous chains tend to have a higher i / I ratio. Like the previous, this feature is not computed for single-exon genes.
- "flank fraction" as f / 20000. Orthologous chains tend to have higher values, as flanking intergenic regions usually align. This feature is essential to detect orthologous loci of single-exon genes since we cannot rely on intron-related features in this case.
- "synteny" as log10 of the number of genes, whose coding exons overlap by at least one base aligning blocks of this chain. Orthologous chains tend to cover several genes located in a conserved order, resulting in higher synteny values.
- "local CDS coverage" as c / CDS, which is only used for single-exon genes.

It is necessary to remark that the term 'global' refers to features computed from all genes that overlap the chain, whereas 'local' refers to features calculated from the single gene under consideration. For example, if a chain covers two genes, 'global' features computed for both genes will be identical, but 'local' will differ. Most of these features quantify how well neutrally evolving intronic and intergenic regions align in comparison to coding exons, which mainly evolve under purifying selection. Because selection in UTR exons is variable, alignments overlapping UTR exons are ignored for feature computation. All features are visually explained in figure 2.4.

2.2.3 Machine learning classification of chains

To classify the extracted transcript-chain pairs as orthologous and not, we apply a pre-trained gradient boosting model. Model selection, training, and testing are described in part 3.1 of this thesis. Model input is a predefined set of features per transcript-chain pair, and output is a number ranging from 0 to 1, which indicates the probability of orthology. We consider each transcript-chain pair that gained the orthology score of ≥ 0.5 as pointing to an orthologous locus. To extract high-confidence orthologs for phylogeny inference, we use a higher score threshold of ≥0.95 (part 4.1).

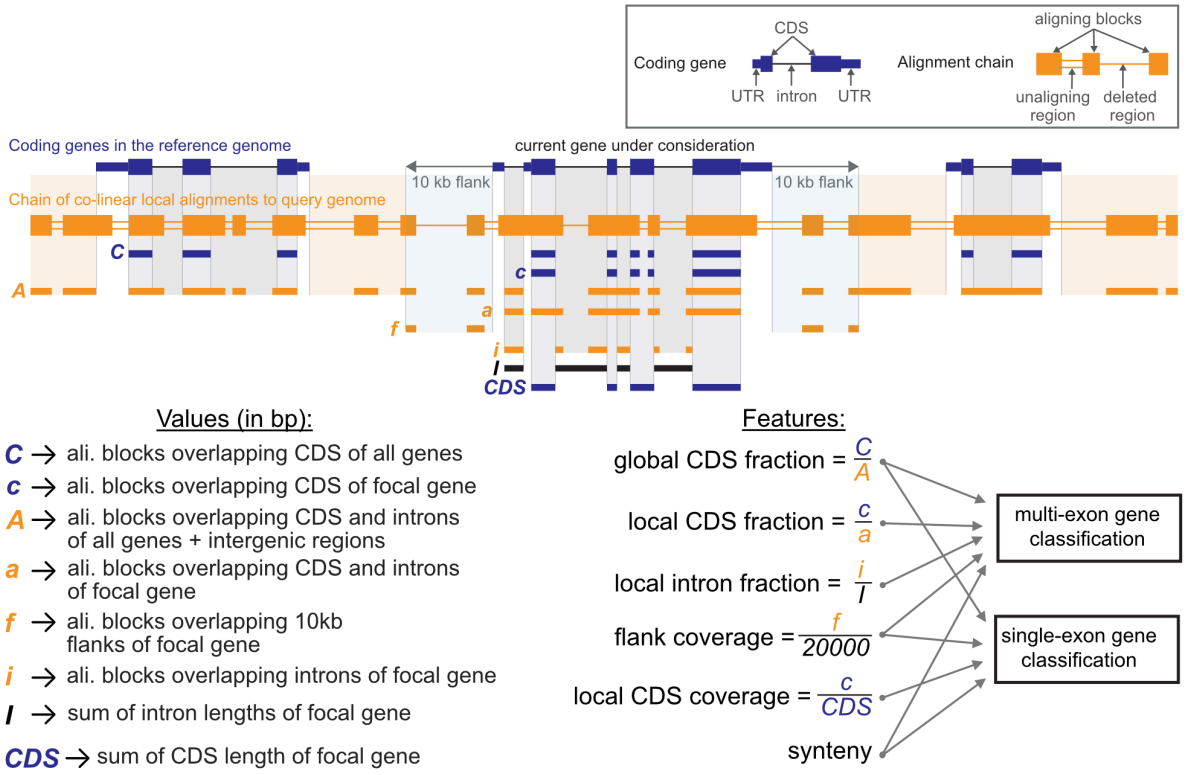

Figure 2.4 Graphical representation of extracted features

The figure shows a locus in the reference genome comprising three genes and an aligning chain. Additionally, it provides graphical representations of "C", "c", "A", "a", "f"," "I", "I", and CDS values (details in the main text).

2.2.4 Handling gene-spanning chains

We use the term spanning for the chains that cover entirely deleted, missing due to assembly gaps, or heavily diverged genes. In this case, the chain blocks are located up and downstream of the considered gene (figure 2.5).

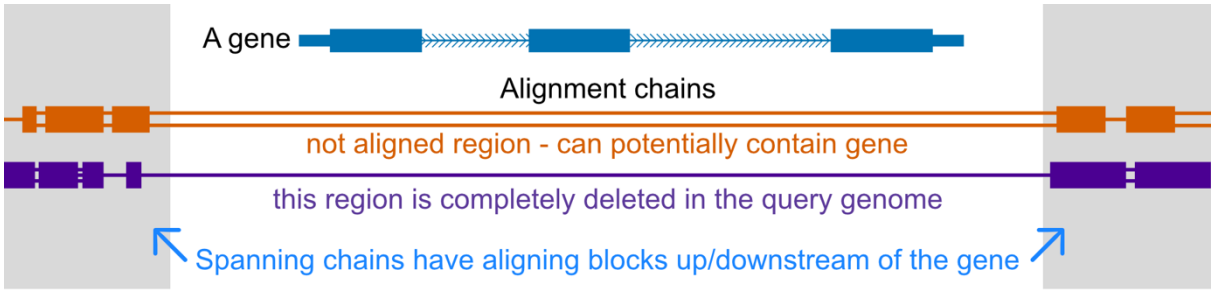

Figure 2.5 Gene-spanning alignment chains

Chains that align up/downstream of the gene and do not have aligning blocks intersecting exons are processed separately because the feature extraction for such chains is impossible. Usually, they represent either the deletion or high sequence divergence in the respective locus in the query genome.

In the absence of aligning blocks intersecting exons, it is impossible to adequately compute the “local” features listed above for such transcript-chain pairs. Therefore, our machine learning classification procedure is inapplicable in this case, so TOGA processes these chains separately. To deal with these chains, TOGA extracts them before classification and treats them as follows: If aligning blocks of this chain overlap coding exons of less than two other genes, TOGA excludes it from further computations. Otherwise, we consider it an orthologous chain candidate. For such chains, we run CESAR 2.0, as described below, on the query locus defined by the closest upstream and downstream aligning blocks.

2.2.5 Annotating processed pseudogenes

Non-orthologous chains do not necessarily align to paralogous copies - paralogs are only a subset of the possibilities. For example, they could also indicate alignments to processed pseudogene (PP) copies. Alignment chains leading to PP copies are characterized by the deletion of intronic regions (figure 2.6). This feature is derived from the fact that, by definition, processed pseudogenes are copies of reverse-transcribed mRNAs (where introns are deleted) that have been inserted back into the genome (Kabza et al., 2014). TOGA exploits this key feature to separate processed pseudogene from the rest of non-orthologous chains. This enables TOGA to enrich the query genome annotation by including processed pseudogenes to the output.

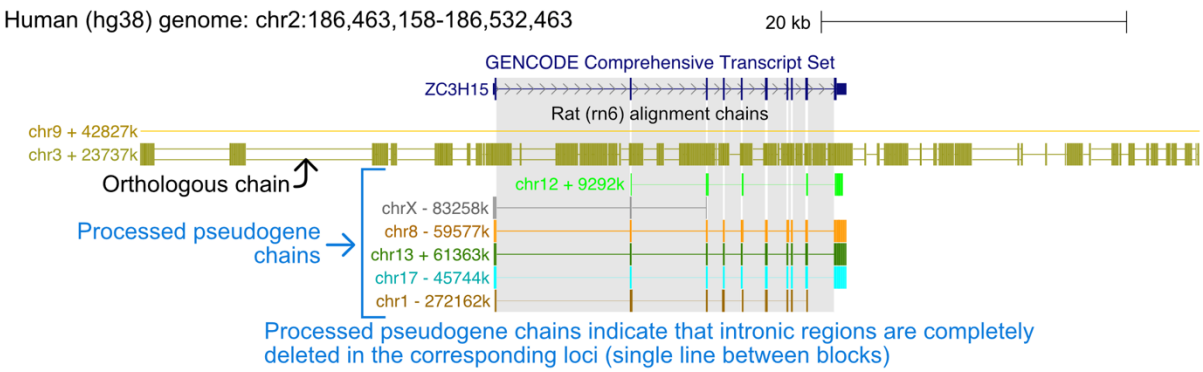

Figure 2.6 Chains aligning to processed pseudogene copies

UCSC browser screenshot shows locus in the human genome containing ZC3H15 gene and rat aligning chains. Intronic regions are highlighted in grey. Chains aligning to processed pseudogene copies indicate that respective intronic regions are entirely deleted in the query genomes. This feature is used in TOGA to separate these chains from the rest of the non-orthologous ones.

To distinguish processed pseudogenes chains, TOGA computes the “alignment to query span” value for multi-exon genes. Defining e as the number of reference bases in the intersection between chain blocks and exons (here using both UTR and CDS) and defining Q as the span of the chain in the query genome, “alignment to query span” is computed as e / Q. For chains that represent processed pseudogenes copies, this value is close to 1 because the intronic fraction in such copies is deleted, by definition. Thus, the cumulative length of exon alignments is close to the total chain length in the query. Therefore, non-orthologous chains that overlap only one multi-exon gene and have the "alignment to query span" value greater than 0.95 are classified as processed pseudogene chains. To annotate PP copies in the query genome, TOGA extracts the corresponding coordinates directly from chain blocks, omitting the CESAR realignment step.

Two examples of processed pseudogene annotations extracted by TOGA are shown on UCSC genome browser screenshots below (figure 2.7). In panel A, the TOGA prediction of the processed pseudogene coincides with the PP annotation track from Ensembl. In the example presented in panel B, TOGA identified a processed pseudogene in an intergenic region which is not detected by Ensembl.

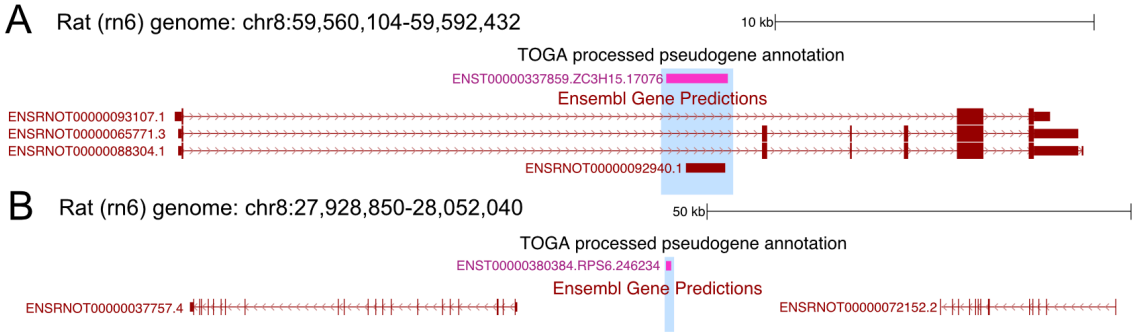

Figure 2.7 TOGA annotation of processed pseudogene copies

UCSC genome browser screenshots show TOGA annotation track in two loci in the rat genome. On panel A, the TOGA-predicted processed pseudogene copy intersects PP copy annotated by Ensembl. On panel B, TOGA annotated a processed pseudogene copy undetected by the Ensembl pipeline.

2.2.6 Assembly of fragmented genes

While contemporary genome projects often aim at producing complete, chromosome-scale scaffolds, some of the species are still represented by strongly fragmented genome assemblies. In such fragmented assemblies, protein-coding genes can be split between different scaffolds. In this case, multiple non-overlapping chains representing alignments residing on different scaffolds overlap the gene and could be misclassified as multiple partial duplications.

Since we observed that TOGA could detect partial gene duplications (Figure 1.10), I implemented a feature that enables TOGA to identify the specific patterns exhibited by alignments to fragmented orthologs. If this is the case, TOGA joins the respective orthologous chains and reassembles the fragmented gene, annotating it as a whole.

Figure 2.8 provides two specific examples of joined genes. Panel A demonstrates that gene fragmentation could occur even in high-quality genome assemblies such as the rat's, where the *MKX* gene is split into two pieces. Panel B shows a side-by-side comparison between genome assemblies of very closely related species: the Kogia and the sperm whale on the example of the human *LRCH3* gene. In the Kogia, this gene is split into six pieces located on different assemblies, whereas in the sperm whale, a single orthologous aligns. Since these species are closest relatives, it suggests that Kogia actually possesses a single undivided ortholog but not six partial copies.

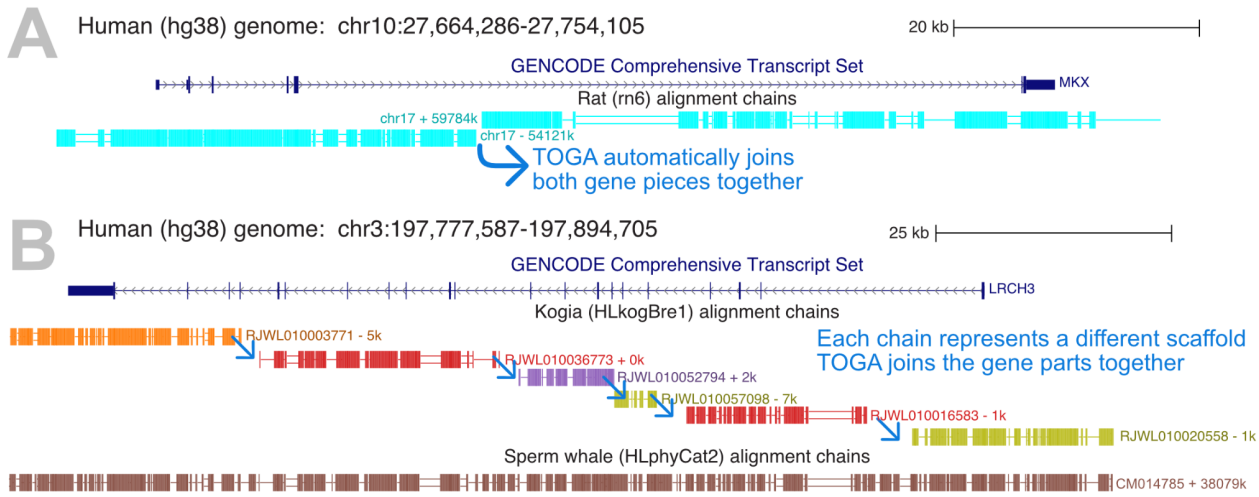

Figure 2.8 Gene residing on multiple scaffolds

UCSC browser screenshot on panel A shows the human MKX gene and rat alignment chains. Notably, the rat ortholog of this gene is split into two pieces. TOGA detects such an event and annotates this ortholog as a whole. Panel B illustrates Kogia and sperm whale alignment chains covering the human LRCH3 gene. This gene is split into six pieces located on different scaffolds in the Kogia genome, whereas in the closest relative (sperm whale), this gene is complete, suggesting that Kogia also possesses a single orthologous copy. As in the first example, TOGA recognizes this pattern and annotates the LRCH3 ortholog in the Kogia as a whole.

To achieve this, TOGA verifies whether multiple orthologous chains overlapping a gene represent a single ortholog residing on multiple scaffolds after the chain classification step. TOGA performs the inspection unless no orthologous chain covers the gene entirely, but the number of orthologous chains is greater than 1.

Based on the start and end coordinates of each orthologous chain along with the orthology confidence score, TOGA builds a directed acyclic graph (DAG) with chains as vertices and orthology score as the edges' weight. Nodes representing adjacent chains are connected. Next, TOGA performs a search for the highest-scoring path in the graph connecting chains from the beginning of the gene to the end. Then, the corresponding loci of the chains comprising the detected path are merged. This merged sequence is further used as an input for the CESAR realignment step. Therefore, the orthologous gene is being annotated as a whole.

To be conservative, TOGA tries to recover only one-to-one orthologs using this methodology. This limitation is justified because a highly complex chain configuration implies multiple optimal paths and, consequently, a high probability of incorrect gene reassembly. However, the vast majority of orthologs are one-to-one, hence this rule does not lead to drastic information loss. With this method, TOGA can extract many additional orthologs from strongly fragmented genomes, which would be otherwise classified as missing.

2.3 Naming convention of the predicted transcripts

Potentially, TOGA could detect multiple orthologous chains for any reference transcript, which means that those transcripts are projected to the query genome multiple times. Therefore, we need a naming convention that uniquely identifies each projected transcript without loss of information. Each alignment chain is associated with a unique identifier, which means that a combination of a transcript name with chain ID uniquely determines projection and could be used to identify TOGA annotations.

Inside TOGA, we use the following notation: transcript ID and chain ID are separated by a dot character. For example, a transcript A projected through chain 9 will be named “A.9” in the query genome. It also allows for any transcript annotated by TOGA to trace which exact reference transcript was projected, and which chain was used for the projection. Figure 2.9 explains this naming convention.

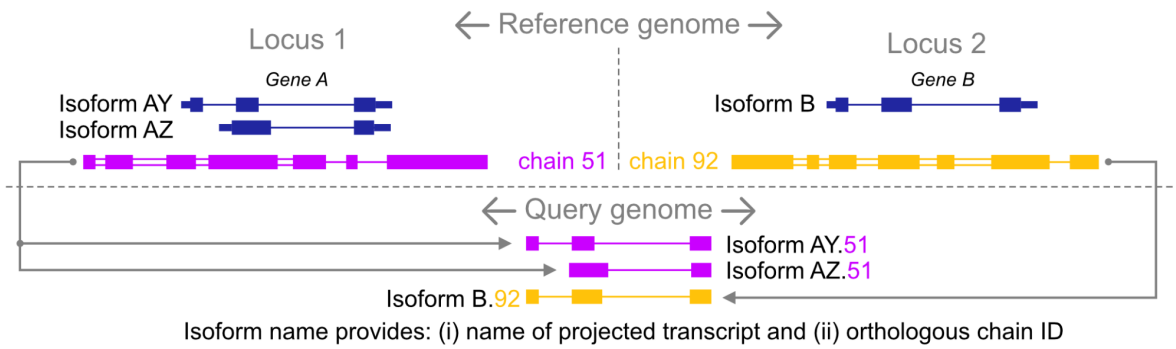

Figure 2.9 Naming convention of the predicted transcripts

TOGA combines the projected reference transcript and orthologous chain identifiers to create unique identifiers for predicted transcripts in the query. For example, reference transcript "AY" projected to the query through chain 51 is annotated as AY.51.

2.4 Aligning reference transcripts to orthologous query loci

The first step of the TOGA pipeline yields a set of transcript-chain pairs classified as orthologous. Each of those transcript-chains unambiguously determines the locus in the query genome where the ortholog of the respected transcript could be detected. In the second step of the pipeline, TOGA projects reference transcripts to orthologous loci in the query to produce query genome annotation. To recognize reference transcripts in the query genome, TOGA applies CESAR (Codon Exon Structure Aware Realigner) version 2.0 in multi-exon mode (Sharma et al., 2016; Sharma et al., 2017).

2.4.1 CESAR approach

Briefly, CESAR 2.0 is a Hidden Markov model-based method that takes the reference exons together with the query sequence as input and produces a pairwise nucleotide alignment, detecting reference exons in the query sequence. In contrast to other nucleotide aligners, CESAR considers reading frame and splice site information to generate the alignments. This method was created principally to align protein-coding sequences, taking specific features of them into account.

This method has high precision in aligning shifted splice sites, can detect precise intron deletions that merge two neighboring exons, and generates alignments of intact exons whenever possible (Sharma et al., 2016; Sharma et al., 2017). CESAR provides a more suitable option to create protein-coding gene annotations than universal nucleotide alignment methods such as BLAST (Sharma and Hiller, 2019).

2.4.1.1 Handling selenocysteine-coding reference codons

CESAR examines the input exon sequences to ensure that they constitute an intact reading frame. For example, CESAR performs inspections for sequence lengths (must be multiple of three), the correct location of split codons, and the absence of premature stop codons. If any of the conditions are violated, then CESAR terminates and shows an error message.

In general, the requirement for the absence of in-frame stop codons is justified because typically intact protein-coding genes do not have them. However, internal TGA codons are used to encode a selenocysteine in a few dozen eukaryotic genes, and consequently, CESAR cannot correctly process such genes. To handle selenocysteine-coding genes, TOGA temporarily replaces in-frame TGA codons in the reference sequence with NNN while running CESAR. This replacement enables CESAR to align such stop codons to sense or stop codons.

2.4.1.2 U2 and U12 splice sites.

Another potential issue is that CESAR tries to find the most optimal alignment, assuming that all introns are flanked with canonical splice sites by default. Indeed, 99% of mammalian introns have canonical “GT”-”AG” or “GC”-”AG” termini spliced by the major U2-dependent spliceosome. Albeit, the minority of introns in higher eukaryotes have different termini spliced by the U12-spliceosome (Burge et al., 1998; Patel and Steitz, 2003).

Thus, for U12-spliced introns, this premise could result in (i) the emission of a corrupted splice site or (ii) incorrect insertions or deletion on exon borders to detect a canonical dinucleotide. However, the information about U12 introns locations in the reference can be

provided to TOGA as input. TOGA can further pass this information to CESAR to handle U12 splice sites properly.

As U12 intron splice sites can comprise a great variety of dinucleotides, I have changed the U12 donor and acceptor splice site profile in CESAR to capture this splice site diversity with a uniform nucleotide distribution. Since information about U12 introns in the reference genome may be incomplete or not available, TOGA considers every intron in the reference without canonical “GT”/”GC”-”AG” splice sites as a putative U12 intron. To generate TOGA annotations with human and mouse genomes as the reference (part 3.7), we used U12 data from U12DB (Alioto, 2007).

2.4.2 Individual exon classification: remaining, deleted, missing

After parsing the CESAR output, TOGA classifies each predicted exon in the query as present (P), missing (M), or deleted (D). This step is necessary since the Viterbi algorithm implemented in CESAR outputs alignments of all input reference exons, including those that do not actually exist in the query locus. For instance, it might produce alignments for cases where the exon is truly deleted or diverged to the extent that no meaningful alignment is possible (class D) or because the exon overlaps an assembly gap in the query genome (class M).

To distinguish between class P, M, and D, TOGA leverages that an orthologous chain provides not only the orthologous query locus, but the aligning chain blocks also provide information about the location of individual exons, as illustrated in figure 2.10.

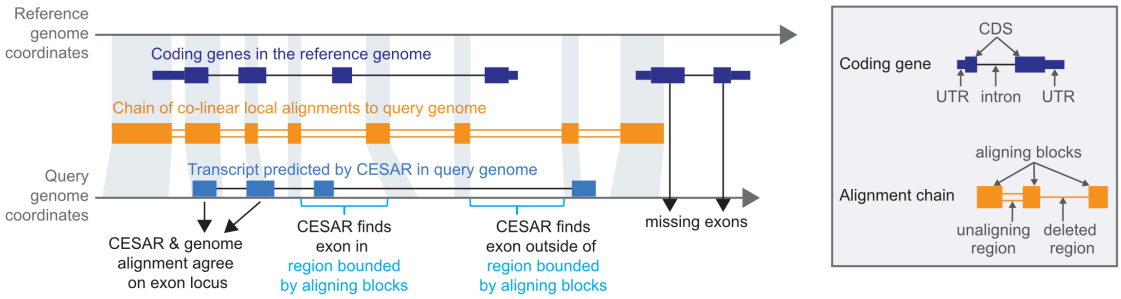

Figure 2.10 Exon alignment classification

The illustration shows an alignment between query and reference genomes. The reference region contains two genes. Using a chain, TOGA could identify for each exon where it is expected to be found (light blue curly bracket). Then, TOGA can check whether CESAR-predicted exons are located in the expected region. Four different scenarios are considered: (i) CESAR and genome alignment agree on exon locus in the query, (ii) CESAR finds exon in the region bounded by aligning blocks (within the expected region), (iii) CESAR finds the exon outside the expected region. If the exon is not covered by chain span (iv), TOGA considers such exons as missing.

TOGA determines whether the predicted exon coordinates overlap the query genome locus that should contain the exon according to the genome alignment chain. In case both the nucleotide-based genome alignment chain and the codon-based CESAR alignment agree on the exon location in the query, TOGA classifies these exons as present (P). For exons, where the chain and CESAR disagree on the exon location or the exon aligns only with the more sensitive CESAR method, TOGA uses two metrics to evaluate whether the exon aligns better than randomized exons.

The first metric is the %nucleotide identity, defined as the percentage of identical bases in the CESAR alignment. The second metric, %BLOSUM, measures the amino acid similarity between reference and query using the BLOSUM62 matrix. The %BLOSUM value is computed as follows. Let S_{RQ} be the sum of BLOSUM scores for each amino acid pair between reference (R) and query (Q), using a score of -1 for insertions and deletions. As S_{RQ} depends on the exon's length, we also determine the maximum score possible for this exon by comparing the reference sequence to itself, thus computing S_{RR}. Thus, %BLOSUM is defined as S_{RQ} / S_{RR} * 100. It is perhaps possible that this value exceeds 100, then we forcibly set it to 100 for numeric consistency.

To determine thresholds that separate real and randomized exon alignments, we extracted 137935 exons of human-mouse one-to-one orthologous genes for which the TOGA-annotated exon overlaps an Ensembl-annotated exon. This resulting set presumably consists of real exons since they are supported by two independent methods. To obtain randomized exons, we reversed the exon sequence and aligned it to the factual query sequence with CESAR. By comparing %nucleotide identity and %BLOSUM between actual and random CESAR exon alignments, we defined thresholds as %nucleotide identity ≥ 45% and %BLOSUM ≥ 20%. These thresholds correspond to a sensitivity of 0.9808 and a precision of 0.99075. The plot of these values for actual and randomized exons is illustrated in figure 2.11.

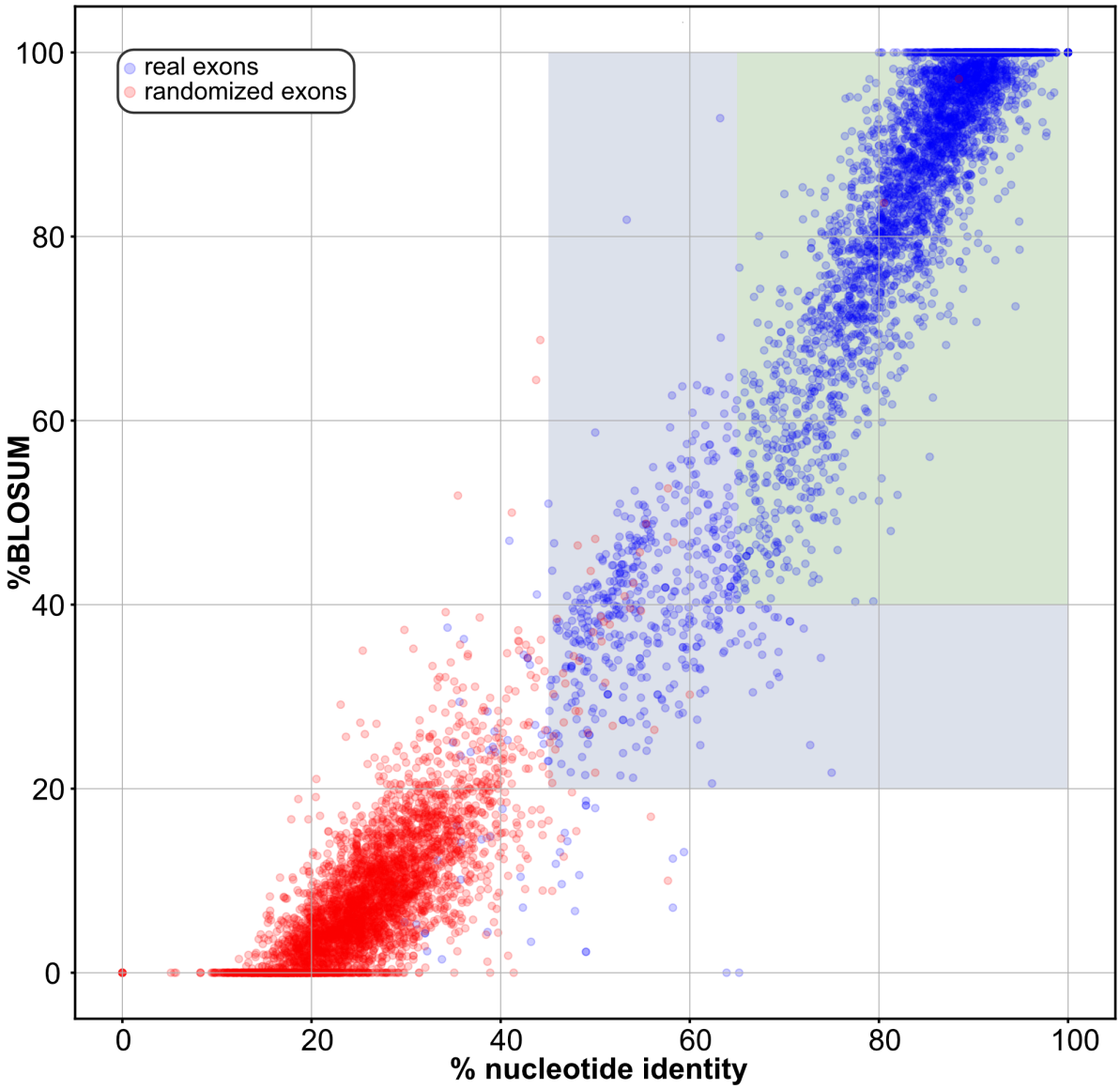

Figure 2.11 Percent nucleotide identity and BLOSUM thresholds obtained from alignments between real and randomized exons

Blue dots illustrate real exons, whereas red dots represent random exon alignments. As expected, randomized exons exhibit lower sequence similarity (nucleotide %identity and BLOSUM score) values. We defined exon annotation thresholds as %nucleotide identity ≥ 45% and %BLOSUM ≥ 20% taking this data into account (blue region). %BLOSUM and %nucleotide identity greater than 50 and 65, respectively, are used to identify high-confidence exon projections.

Exons that exceed these thresholds are classified as present (P). For all other exons, TOGA determines whether the query locus expected to contain this exon overlaps an assembly gap (10 consecutive N characters) in the query genome. If so, TOGA classifies the exon as missing (M). Otherwise, it is classified as deleted (class D). Exons not spanned by an

orthologous chain are also classified as missing (M), as such cases are often due to assembly incompleteness. The decision process is illustrated in figure 2.12.

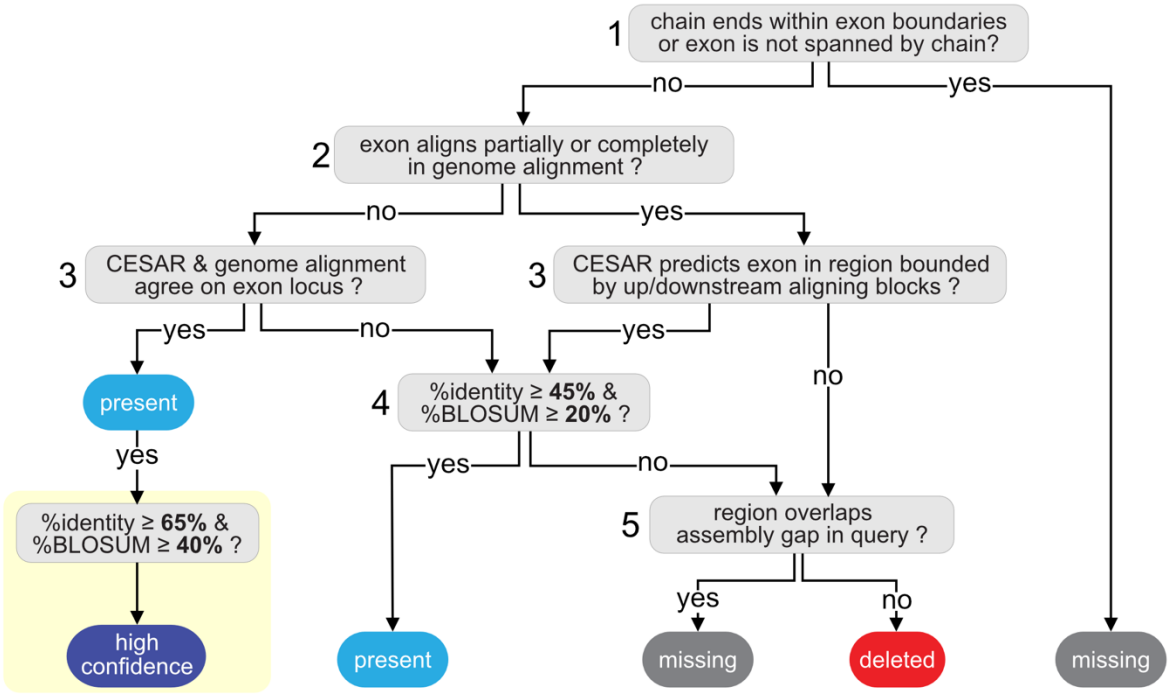

Figure 2.12 Exon classification decision tree

Applying the illustrated procedure, TOGA classifies exons as Present, Deleted, or Missing. For a given exon, TOGA assesses whether the orthologous chain ends inside this exon or whether it is not spanned by an orthologous chain (if yes – it is classified as Missing) (1). For all other exons, TOGA determines an 'expected region', which is a region in the query genome that should overlap or contain the exon according to the genome alignment chain (figure 2.10). TOGA then assesses whether CESAR 2.0 in multi-exon mode finds an exon candidate in the expected region (3), which shows that genome alignment and CESAR agree on the exon location in the query. Such exons are classified as present (P), and the subset of these exons which fulfills %nucleotide identity ≥ 65% and %BLOSUM ≥ 40% are labeled as high confidence and further considered for phylogeny inference (yellow background box). Exons, for which genome alignment and CESAR disagree on the query location (3), are classified as present (P) if the CESAR exon alignment is better than randomized exon alignments, which TOGA determines using a threshold of %nucleotide identity ≥ 45% and %BLOSUM ≥ 20% (4). For exons that do not align at all in the genome alignment (2), the 'expected region' is defined as the query region bounded by the nearest up- and downstream alignment block in the chain. Exons for which the CESAR exon candidate is found in the expected region (3) and that align better than randomized exons (4) are also classified as present (P). For all other exons, TOGA determines whether the expected region overlaps an assembly gap in the query genome (5). If so, the exon is classified missing (M), otherwise as deleted (D).

2.4.3 Detecting gene-inactivating mutations

Inactivating mutations detection procedure implemented in TOGA rigorously scans the predicted reading frame of every transcript predicted by CESAR. This procedure is necessary to distinguish between intact and lost transcripts. In particular, TOGA considers the following gene-inactivating mutations:

1) frameshifting insertions and deletions,
2) in-frame (premature) stop codons,
3) mutations that disrupt the canonical donor ("GT"/"GC") or acceptor ("AG") splice site dinucleotides,
4) deletions of single or multiple consecutive exons that together have a total length not multiple of three, resulting in frameshifts.

The abovementioned mutations and their influence on the reading frame are described in detail in subsection 1.1.1.3 of the introductory chapter.

2.4.3.1 Differences with previously published gene loss pipeline

Gene loss detection pipeline integrated into TOGA is based on previously published work (Sharma et al., 2018). Despite this, the version integrated into TOGA has several changes that improve gene loss detection accuracy. These changes are described in detail in this subsection.

Contrary to our previous work, we do not consider larger frame-preserving insertions and deletions as inactivating mutations. We observed many cases where such deletions longer than 600bp inside huge exons can occur in conserved genes. These large frame-preserving deletions result in substantially shorter but likely functional proteins. Figure 2.13 (next page) provides two UCSC genome browser screenshots exemplifying these cases. In example A, the alignment chain indicates that the large *RESF1* exon exhibits a 636pb-long deletion. However, the corresponding gene in the mouse genome is intact. The human *RESF1* gene encodes a 1747 amino acid long protein, whereas the mouse ortholog is shorter and encodes a 1521 amino acid long protein. Similarly, example B illustrates 114 bp insertion in the mouse ortholog of the human *CRNN* gene. The previous gene loss pipeline implementation would consider this insertion as a loss-of-function mutation. However, the orthologous gene is intact in the mouse and encodes a longer protein.

These examples suggest that large frame-preserving insertions and deletions may not result in gene inactivation. For that reason, the gene loss detection pipeline implemented in TOGA does not consider them as such, which allows TOGA to make more careful gene inactivation predictions.

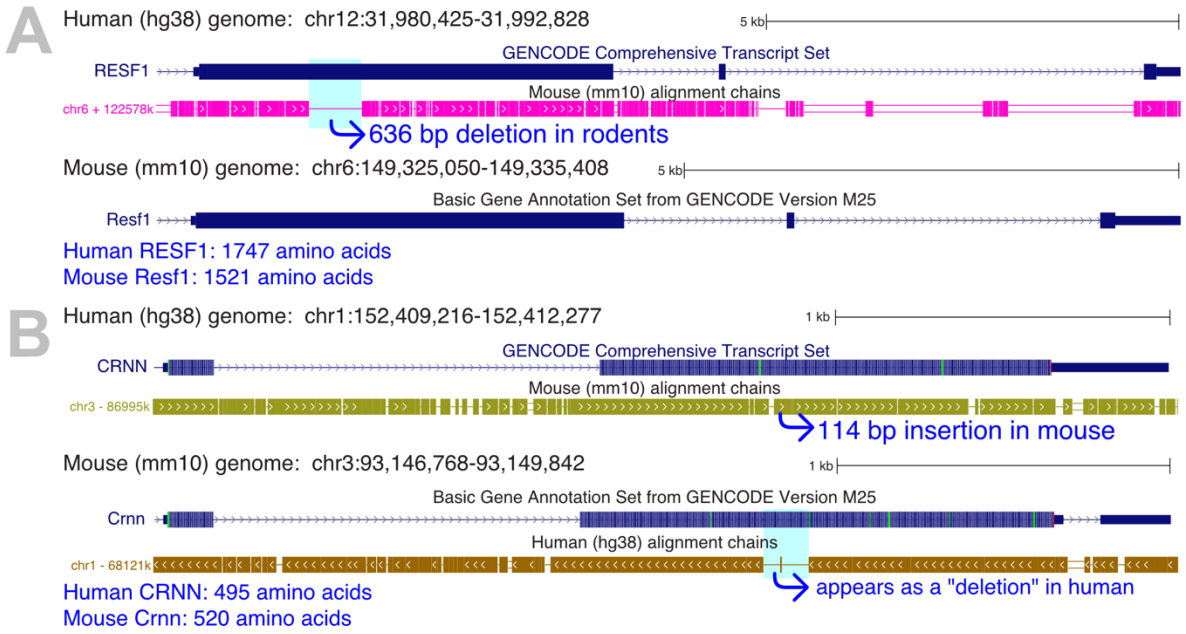

Figure 2.13 Large frame-preserving deletions in large exons may not result in gene inactivation

Panel A: UCSC browser screenshot shows that human gene RESF1 (1747 aa) is substantially longer than the mouse ortholog Resf1 (1521 aa). Despite this, both genes are intact. The alignment chain shows 636bp deletion in rodents. Panel B shows a similar example: Mouse Crnn gene (520 aa) is longer than human ortholog (CRNN, 495 aa long).

Moreover, we do not consider frame-preserving exon deletions as loss-of-functions mutations. This is based on observations that deletions of the entire exons can occur in intact protein-coding genes. Figure 2.14 illustrates this case on the example of the *CALCOCO2* gene. This gene exhibits drastic exon structure changes between the human and mouse genomes. The human gene consists of 12 coding exons conserved in placental mammals and encodes a 446 amino acid long protein. However, in the mouse genome, the last seven consecutive exons of this gene are deleted. Despite these exon deletions, the mouse gene has an intact reading frame, encodes a shorter protein of 331 amino acids, and is expressed as confirmed by the mouse mRNA track. In order to avoid the misclassification of such genes as lost, TOGA considers exon deletions as inactivating only if they disrupt the ORF.

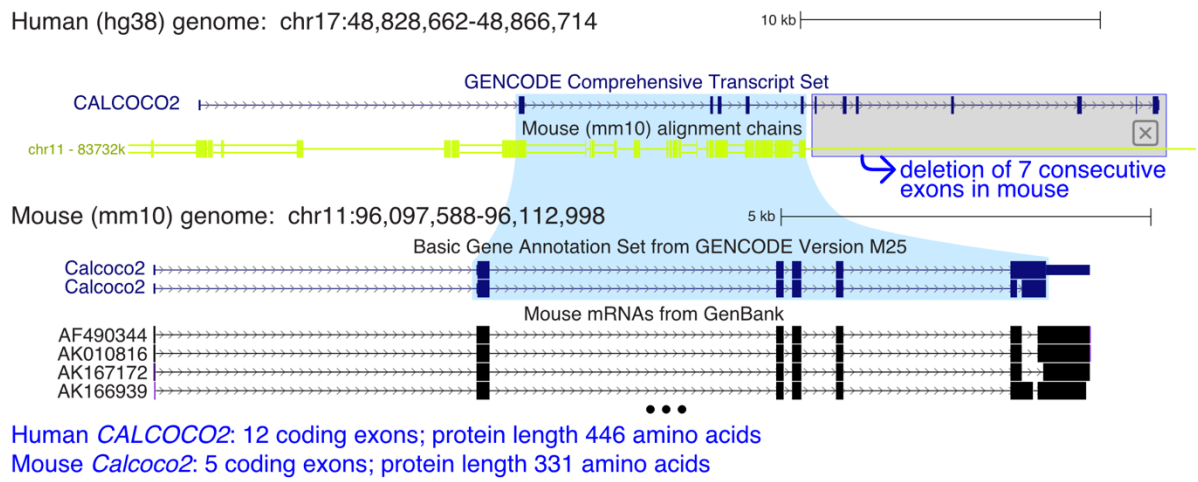

Figure 2.14 Example of radically different exon-intron structure between orthologs

UCSC browser shows CALCOCO2 genes in human and mouse genomes. Seven consecutive exons of this gene are deleted in the mouse Calcoco2 gene. However, this gene is still intact in the mouse genome, as transcriptomic data illustrates.

Importantly, frame-preserving exon deletion may result in the assembly of a stop codon at the exon-exon boundaries (figure 2.15). TOGA additionally scans the predicted reading frame for such mutations and, if detected, considers them inactivating.

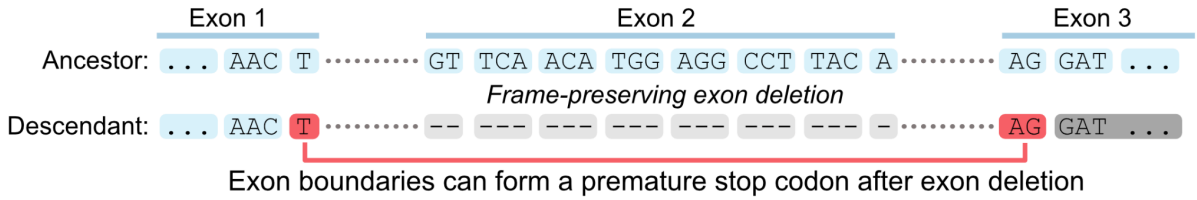

Figure 2.15 Frame-preserving exon deletion introduces a premature stop codon

In this example, the deletion of the exon two is not considered inactivating mutation because it does not disrupt the gene ORF. However, after the exon deletion, boundaries of exons 1 (T-) and 3 (-AG) can form a premature stop codon.

Gene loss detection pipeline integrated into TOGA ignores “TGA” in-frame codons that are already present in the reference sequence. This modification is important to avoid erroneously reporting the inactivation of selenocysteine-coding genes because such “TGA” codons usually do not interrupt the protein translation process.

2.4.3.2 Inactivating mutations filtering

Accurately detecting gene-inactivating mutations in predicted transcripts poses numerous challenges. For example, sequencing errors and alignment artifacts may mimic inactivating mutations in genes that are, in fact, conserved. Moreover, even factual inactivating mutations do not guarantee that a gene does not encode an intact protein. For example, two frameshifting indels may compensate each other (figure 2.16). Compensated frameshifts alter the amino acid sequence, but the resulting protein preserves some sequence similarity and most likely is intact.

Figure 2.16 Compensated frameshifts

The figure illustrates a frameshift compensation event: consecutive +1 and -1 frameshifts compensate each other. The sequence between these frameshifts is expressed in an alternative frame. If the alternative sequence does not comprise a stop codon, this mutation is not considered inactivating. Sequence after the last frameshift is translated in the ancestral frame.

Also, inactivating mutations that occur close to the N or C termini of the encoded proteins are less likely to inactivate the gene because these regions are shown to be under weaker evolutionary constraints (MacArthur et al., 2012). Figure 2.17 below illustrates this postulate on the distribution of frameshift positions discovered in the mouse coding sequence. Out of 5566 considered frameshifting mutations, 58% of them are detected in the first or last 10% of CDS. Since the majority of detected frameshifting mutations occur close to the N or C termini of the proteins, it confirms that these regions are more tolerant to inactivating mutations. In addition, to avoid misclassification of potentially conserved genes as lost, TOGA applies the following filters that were used in our previous implementation:

1. In the case of precise intron deletions that merge two neighboring exons into a single larger exon, we do not consider subsequent deletion of the splice sites an inactivating mutation.
2. We do not consider splice site mutations for U12 splice sites labeled as such in the reference or inferred from non-canonical reference splice site dinucleotides.
3. If two or more frameshifts compensate each other (e.g., a -1 and -2 bp deletion, or three -1 bp deletions) and do not result in a stop codon in the new reading frame are not considered as inactivating mutations (figure 2.16).

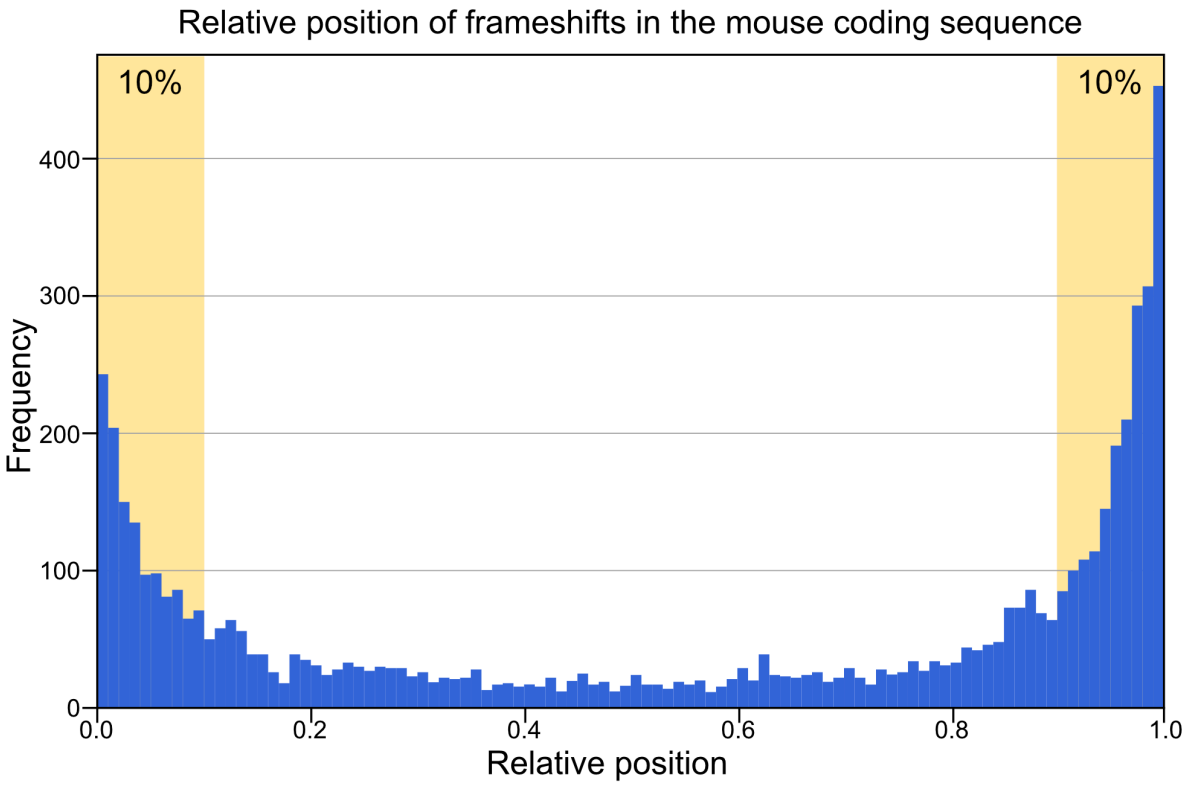

Figure 2.17 Relative position of frameshifts in the mouse coding sequence predicted by TOGA

Figure shows that the vast majority of the frameshifts in the mouse coding sequence are located close to 3' or 5' termini. These regions evolve under weaker evolutionary constraints; therefore, TOGA ignores inactivating mutations detected in the first or last 10% of the CDS.

The series of filters integrated into TOGA drastically reduced the number of falsely reported inactivating mutations. The detailed evaluation of gene loss detection specificity is presented in part 3.3 of this work.

2.4.4 Transcript annotation and classification

To annotate transcripts in the query genome, TOGA utilizes the coordinates of CESAR exons predictions that were classified as present in the previous step. Gene orthology must be inferred based on the number of (co-)orthologs in the query genome that encode a functional protein. For example, even if TOGA detects a single orthologous locus for the given gene with high confidence, the predicted gene could be inactivated in the query, resulting in a one-to-zero orthology relationship. Similarly, TOGA may detect multiple orthologous loci for a single gene and infer one-to-one orthology due to the inactivation of redundant copies. Figure 2.18 below provides a specific example of this principle in action.

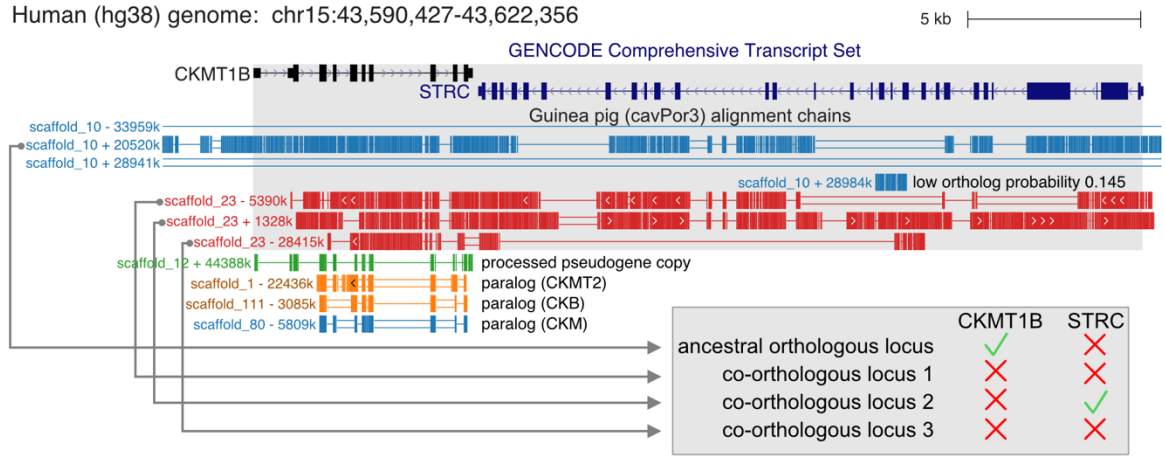

Figure 2.18 Loss of ancestral copy

As the UCSC genome browser screenshot shows, TOGA reveals four orthologous loci for the STRC gene in the guinea pig genome. However, the gene is inactivated in three of those orthologous loci, including the ancestral one. Instead of classifying this gene as lost, TOGA detects one intact copy and reveals the correct one-to-one orthology relationship between human and guinea pig STRC. It is worth mentioning that the CKMT1B gene is intact in the ancestral locus but is inactivated in others.

To determine whether an annotated transcript is expected to encode a functional protein, TOGA implements a transcript classification step. Transcript classification is not a straightforward problem since assembly gaps result in missing parts of the CDS, and individual exons can get lost in otherwise clearly conserved genes, as shown in previous work (Sharma et al., 2018). To take this complexity into account, we decided to classify annotated transcripts into five different classes (figure 2.19):

1) "Intact" transcripts, which are expected to encode functional proteins. In these transcripts, the middle 80% of the CDS is present (no missing sequence detected) and exhibits no gene-inactivating mutation.
2) "Partial Intact" transcripts, for which ≥50% of the CDS is present, and the middle 80% of the CDS exhibits no inactivating mutation. These transcripts are also very likely to

encode functional proteins, but the evidence is weaker as more CDS is missing due to assembly gaps.

3) "Missing" transcripts, for which less than 50% of the CDS is present, and the middle 80% of the CDS exhibits no inactivating mutation. These transcripts are undecided as more than half of the CDS is missing, but no strong evidence for loss was detected. Additionally, we distinguish the "partial missing" subclass for which the orthologous chain spans less than 35% of CDS. This subclass is even more questionable because it is hard to determine whether this copy exists in the query.
4) "Uncertain Loss" transcripts exhibit at least one inactivating mutation in the middle 80% of the CDS. The evidence is not strong enough to classify the transcript as lost; hence, the chances of whether it encodes a functional transcript or not are barely equal.
5) "Lost" transcripts, for which evidence for loss is sufficiently strong, are unlikely to encode a functional protein. Gene loss criteria are explained in detail in the following section.

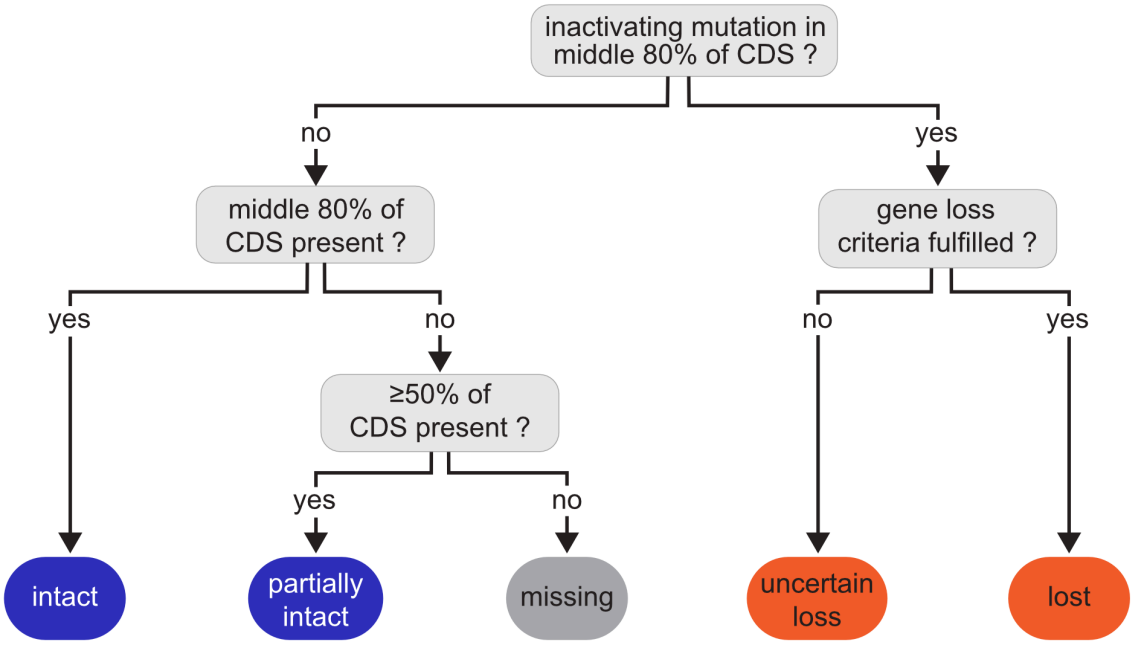

Figure 2.19 Decision tree of gene classification

TOGA begins the transcript classification by first determining whether the transcript exhibits no or at least one gene-inactivating mutation in the middle 80% of the CDS. This fundamental distinction is motivated by earlier observations that inactivating mutations in conserved genes mainly occur in the first or last 10% of the CDS. Transcripts that exhibit no inactivating mutations in the 80% of the CDS are further classified as "Intact," "Partial Intact," or "Missing," depending on the amount of missing sequence. Transcripts that have inactivating mutations in this region are classified as "lost" or "uncertain loss," depending on the satisfaction of gene loss criteria.

2.4.4.1 Transcript loss criteria

Analyzing the list of detected inactivating mutations, TOGA quantifies the maximum percent of the reading frame that remains intact in the query. To distinguish between intact, partially intact, and missing transcripts, TOGA computes this value ignoring missing sequences and pretending that this sequence was never included in the CDS. Alternatively, to distinguish between uncertain loss and lost transcripts, TOGA considers the missing sequence as aligning codons, making the careful assumption that missing codons correspond to sense codons in the unknown query sequence. To compute these values, TOGA applies a procedure explained in figure 2.20. This procedure results in a consistent classification of transcripts with the same inactivating mutations and only differing in the amount of missing sequence.

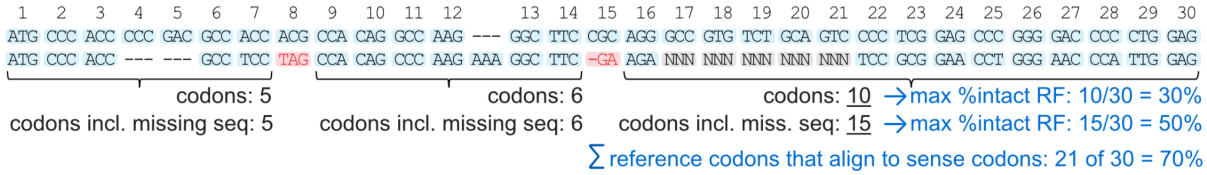

Figure 2.20 %intact calculation procedure

In this particular example, the examined transcript has 30 codons. "NNN" triplets represent a missing sequence occurring due to assembly gaps. Two inactivating mutations (stop codon and frameshift, highlighted in red) determine the boundaries of three individual parts of the reading frame that remain intact in the query species. We determine the number of codons that align, ignoring deleted and inserted codons for each consecutively intact part. We compute two versions of the %intact value: (i) ignoring the missing sequence and (ii) pretending that the missing sequence is intact. In case (i), the maximum percent of the reading frame that remains intact is the third part in this example, with 10 of 30 codons aligning (30%). In case (ii), the maximum percent of the reading frame that remains intact covers 15 of 30 total codons (50%).

Following previous work (Sharma et al., 2018), we use the following criteria to classify transcripts that exhibit at least one inactivating mutation in the middle 80% of the CDS as "uncertain loss" or "lost." Lost transcripts have a maximum percent intact reading frame <60% and exhibit inactivating mutations in at least two coding exons. The latter requirement is motivated by previous observations that loss-of-function mutations in a single exon of an otherwise-conserved gene do not provide strong evidence to infer gene loss. For genes consisting of more than ten exons, we require inactivating mutations in at least 20% of the coding exons.

By definition, this requirement cannot be satisfied in single-exon genes; therefore, we require two inactivating mutations in such genes to infer gene loss. Moreover, since the sizes of individual exons can be relatively large, we make an exception for multi-exon transcripts,

where a single large exon represents a significant part of the CDS (threshold 40% of CDS). Such transcripts are also classified as "lost" if at least two inactivating mutations occurred in this biggest exon. All other transcripts that have at least one inactivating mutation in the middle 80% of the CDS are classified as "uncertain loss," indicating that evidence for loss is not strong enough as a more significant part of the CDS remains potentially intact (>60%), or not enough exons exhibit inactivating mutations. Figure 2.21 summarizes all possible transcript classifications, illustrating different cases and the maximum percent of the intact reading frame in the query species.

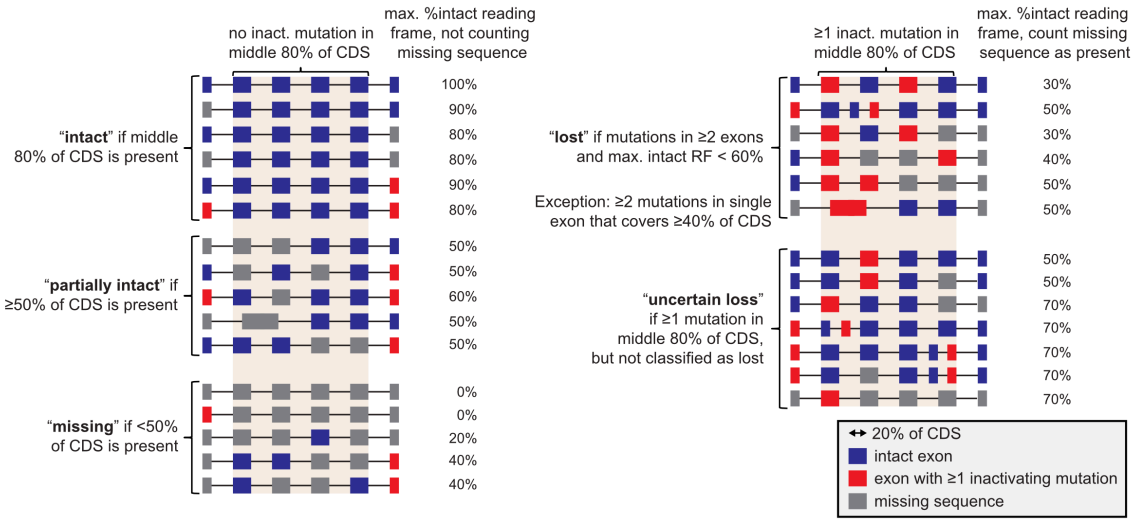

Figure 2.21 Examples of different transcript classes

Exemplified transcripts consist of 5-7 coding exons that make up 10%, 20%, or 40% of the total CDS (boxes of different sizes). Blue and red boxes represent coding exons with and without inactivating mutations, respectively. Also, grey boxes represent coding exons missing in the query (usually due to assembly gaps). For transcripts exhibiting mutations in the middle 80% of the CDS, we conservatively count missing sequence parts as present sequences lacking inactivating mutations when determining the maximum percent of the intact reading frame.

Since TOGA does not consider frame-preserving deletions as inactivating mutations in the current implementation, this could lead to the classification of wholly and nearly deleted transcripts as intact. To avoid this misclassification, we added an additional step to re-classify likely non-functional genes where the significant fraction of CDS is lost due to frame-preserving deletions. For this purpose, we compute the percentage of reference codons that align to sense codons in the query and classify a transcript as “uncertain loss” if this value is less than 50% and as “lost” if it is less than 35%. Please note that this value is equal to 0% by definition if and only if a gene is entirely deleted.

2.5 Orthology inference

TOGA performs the final orthology inference, considering that:

1. each transcript might have several orthologous loci,
2. each of these loci could be orthologous for several reference transcripts, and
3. some of these predicted genes could be classified as lost at the final step of the pipeline.

To accomplish this, TOGA aggregates the obtained data across both reference and query and classifies the predicted orthologs as one-to-one, one-to-many, many-to-one, or many-to-many.

2.5.1 Classifying genes based on the classification of all transcripts and all orthologous loci

A gene in the reference genome may have several isoforms and several inferred orthologous loci in the query. TOGA uses all orthologous loci and the classification of all predicted transcripts to determine whether the gene has at least one functional ortholog in the query and, if so, what the orthology type is (one-to-one, one-to-many, many-to-one, or many-to-many). For a given gene and one orthologous locus, TOGA considers the classification of all transcripts that were annotated for this locus and applies the following order of precedence: "intact," "partially intact," "uncertain loss," "missing," "lost," "partial missing," and "paralogous projection."

Therefore, if at least a single transcript is classified as intact, TOGA concludes that this orthologous locus contains at least one functional gene ortholog. An orthologous locus is inferred to contain a lost gene if and only if (i) all annotated transcripts of the given gene are classified as lost or (ii) annotated transcripts classification exhibits a mixture of lost and "partial missing" classes.

The higher rank order of "missing" class compared to "lost" reflects our reasoning that a missing transcript, for which the query sequence of some exons is unknown, might actually encode a functional transcript, making TOGA's gene loss inferences conservative. However, the "partial missing" class has a lower rank than "lost" because such transcripts are unlikely to represent a functional gene.

To determine orthology type, TOGA considers classifications of each reference gene in all respective orthologous loci, and for each of these query loci, which reference genes were annotated. This principle of hierarchical classification is illustrated in figure 2.22. In this example, the hypothetical gene “*XYZ123*” consists of two isoforms called A and B. Furthermore, this gene has two orthologous loci in the query represented by chains 1 and 2.

Thus, each isoform is projected twice to the query genome, resulting in four individual projections (predicted transcripts) in the query genome. Further, each of those projections gets classified.

In this example, isoform A is classified as "intact" in the first locus and "lost" in the second locus. Simultaneously, in the same loci, isoform B is classified as "uncertainty lost" and "missing," respectively. Following the transcript classification principle, isoform A is classified as intact because one of the respective projections is classified as such. For isoform B, the highest rank of projection is "uncertain loss." The entire "*XYZ123*" gene obtains the "intact" class, because it is the highest-ranking class of all orthologous transcripts.

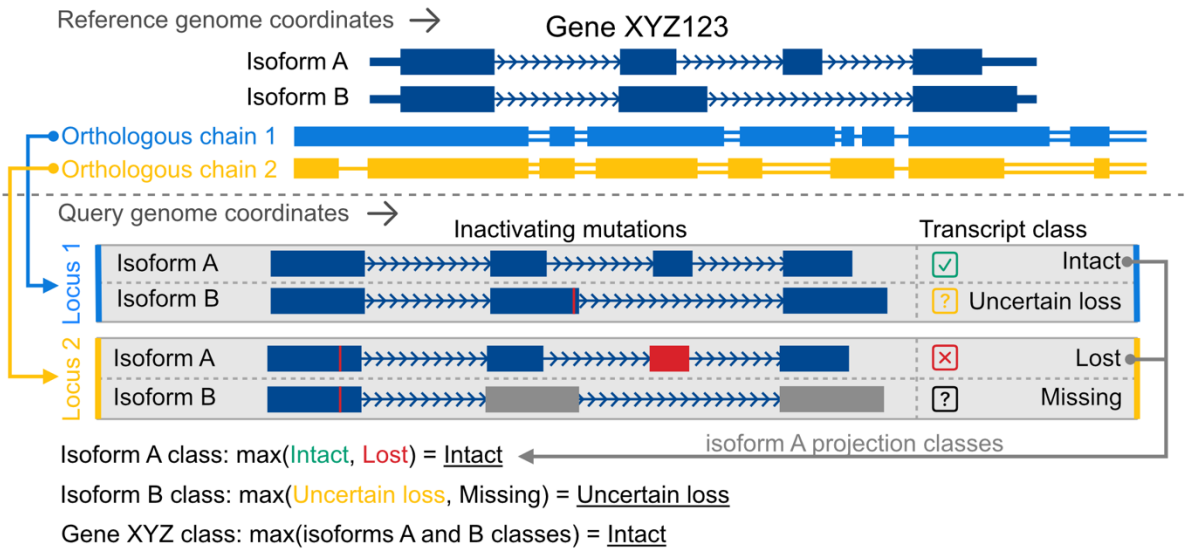

Figure 2.22 Transcripts classification order

Reference gene XYZ123 comprises two alternative splicing isoforms: A and B. Also, two orthologous chains align to this region: 1 and 2. Therefore, TOGA annotates four different XYZ123 transcripts in the query genome: A.1, A.2, B.1, and B.2. To classify the XYZ123 gene in the query, TOGA selects the highest rank of the predicted transcripts. Since the projection A.1 was classified as intact, TOGA classifies the entire gene XYZ123 as intact in the query species.

2.5.2 Association of predicted transcripts into genes

Noticeably, during the preceding steps, TOGA predicts individual transcripts in the query genome but not the entire genes. However, consolidation of individual transcripts into genes is necessary for the final orthology inference. Specifically, it is essential to infer many-to-one and many-to-many orthologs because these orthology classes imply projections of several reference genes to the same locus in the query genome.

To associate predicted transcripts into genes, TOGA groups transcripts located in the same locus. Within each group, it identifies pairs of transcripts that (i) are located on the same strand and (ii) have overlap between predicted CDS. Pairs that satisfy these two criteria are assigned to the same query gene (figure 2.23, panel A). Also, panels B and C in figure 2.23 exemplify cases where intersecting transcripts are not assigned to the same gene.

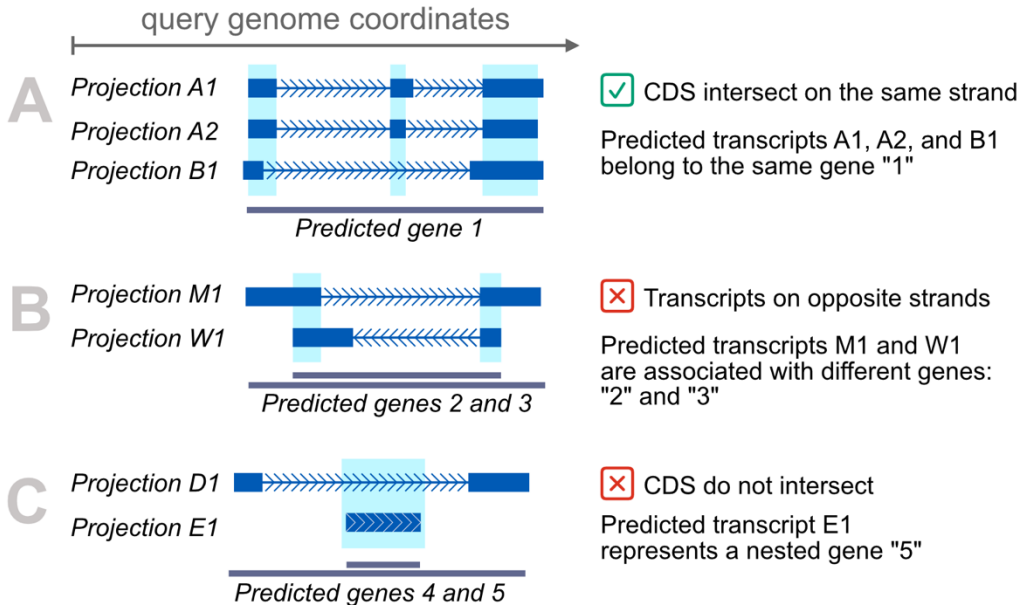

Figure 2.23 Association of predicted transcripts into genes

Case A: three TOGA projections are located on the same strand, and their CDS intersect. Therefore, TOGA associates them with a single query gene. Cabe B: CDS of the TOGA projections W1 and M1 intersect. However, they are located on the opposite strands. Thus, TOGA annotates them as representatives of two separate genes. Case C: E1 is a nested transcript, and its CDS does not intersect with the CDS of the projection D1. Therefore, TOGA does not associate D1 and E1 with the same gene in the query.

2.5.3 Building an orthology graph

In the last step, TOGA builds a bipartite graph where nodes represent reference and query genes and edges symbolize the inferred orthology relationships weighted by the gradient boosting orthology score of the respective chain. Then, it splits the graph into connected components, classifying each subgraph as one-2-one, one-2-many, many-2-one, or many-2-many orthology connections. This principle is illustrated in figure 2.24.

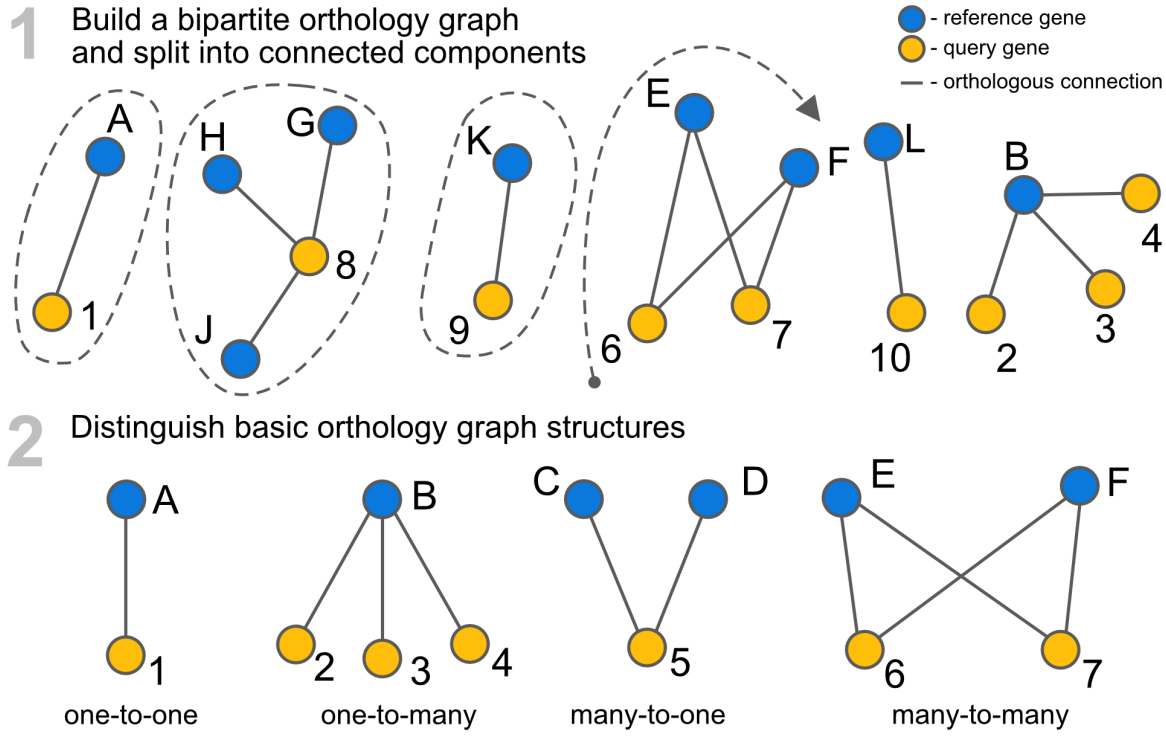

Figure 2.24 Final orthology inference

1: TOGA builds a bipartite graph where nodes are reference or query genes, and edges represent identified orthology connections between these genes. Then, it splits the graph into connected components. 2: TOGA classifies individual subgraphs as one-to-one, one-to-many, many-to-one, or many-to-many.

However, some many-to-many orthology subgraphs may stay incomplete, which means that some reference and query gene nodes are not connected within the subgraph. These graphs are subject to more detailed analysis because they may contain edges that have considerably weaker support. To remove individual orthology relationships within a set of many-to-many orthologous genes that have substantially weaker support, TOGA uses the chain orthology scores. For genes with a putative many-to-many orthology relationship, where 'cross-gene' orthology is supported only by alignment chains with weak orthology scores, this procedure typically results in correct one-to-one orthology relationships.

In detail, TOGA analyzes whether edges with substantially weaker orthology scores can be removed from the incomplete many-to-many orthology graph (figure 2.25). To this end, TOGA subdivides all edges into two sets: The first set contains all edges that connect a leaf node (reference or query gene that has only one inferred ortholog). Set 2 contains all other edges. Let S_{min} be the minimum orthology score of edges in set 1. Branches in set 2 with a score < S_{min} * 0.9 will be removed unless this would result in either (i) an isolated node that loses all its orthology connections or (ii) a split of reference genes that have > 1 mutual connections between different subgraphs. These exceptions make the TOGA orthology refinement procedure conservative.

Panel A illustrates the most trivial case, where a weak 1-B edge (score 0.54) that connects A-1 and B-2-3 orthologs can be removed because the remaining edges show a higher (>0.89) support. This procedure results in a separation of weakly supported many-to-many orthologs into strongly supported components representing one-to-one and one-to-many orthologs.

In panel B, reference genes A and B have two mutual orthologous connections to query genes 1 and 2 and are considered indivisible. Edges 1-B and 2-B show significantly weaker connection support than the rest of the edges. However, removal of these edges separates A and B into different components, which TOGA does not permit. Therefore, TOGA does not perform this operation, annotating connections between nodes A, B, 1, 2, and 3 as many-to-many orthologs.

Panel C provides an example where deletion of weaker nodes results in isolated node C, which is not permitted. Finally, a complete many-to-many graph is illustrated in panel D. TOGA takes no action in this case because all orthologous connections have substantially strong support.

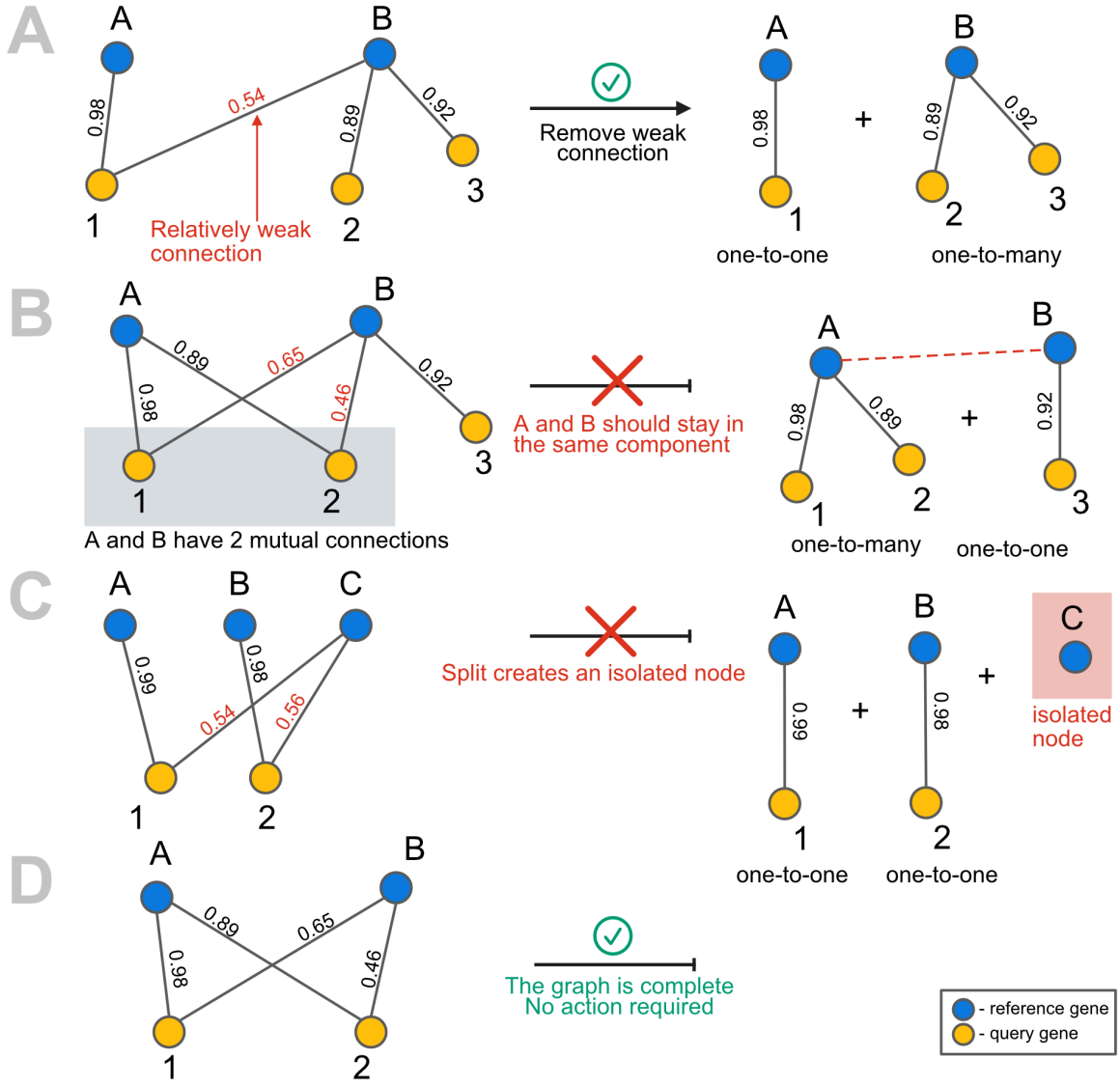

Figure 2.25 Graph pruning to resolve complex many-to-many orthology cases

Case A illustrates a trivial case where a weak B-1 edge can be deleted. Therefore, putative many-to-many orthology is split into two orthology groups with more substantial support: one-to-one (A-1) and one-to-many (2-B-3). Case B: since reference genes A and B have >1 mutual orthologous connections, TOGA does not permit the separation of them into different orthology groups. Case C: removal of weakly supported connections produces a separate node C, which is not permitted. Case D illustrates a complete bipartite graph: in this case, TOGA does not take any action.

2.6 TOGA output

As for the pipeline output, TOGA produces the following files:

1. Query genome annotation in bed-12 format. Each predicted transcript is named following the convention described in part 2.3. Bed-12 can be further converted into other formats.
2. Isoforms table for the query annotation. This file establishes the correspondence between predicted transcripts and genes in the query genome.
3. Orthology classifications table that establishes the correspondence between orthologous genes in reference in query. For each connection between reference and query genes, the table also shows the orthology connection class such as one-to-one, one-to-many, etc.
4. A table containing gene loss pipeline output. For each predicted transcript and gene, the table assigns a class such as "intact", "lost", etc.
5. Fasta file containing nucleotide alignments per exon.
6. Fasta file containing pairwise protein alignments. All protein alignments are corrected for frame-disrupting mutations such that frameshifting indels do not change the amino acid sequence downstream of the mutation.
7. Fasta file containing codon alignments. These alignments are also corrected for frameshifting indels; therefore, alignment downstream of the frameshifting mutations is not affected and remains intact.
8. A table containing a list of detected inactivating mutations in each predicted transcript.

Output files are designed in a user-friendly manner such that they are easy to parse with basic command-line tools such as grep. Therefore, TOGA output files can be used in subsequent analysis tools with minimal effort.

2.7 Filtering reference genome annotation

Many controversial TOGA results we observed during the TOGA development phases come from incorrect transcript annotations in the reference. For instance, the same reference gene could be annotated twice under different identifiers, resulting in erroneous reports of a two-to-two orthology. As another example, if a gene is represented only by reference-specific isoforms, it could lead to a false gene loss report since the factual gene in the query has a different exon-intron composition. Even the state of art collections of representative transcripts like APPRIS (Rodriguez et al., 2013) might contain transcripts that could potentially lead to obstacles, such as NMD transcripts, incomplete annotations, or reference-specific ones.

Additionally, many sequenced genomes do not have a high-quality gene annotation. In such genome annotations, inaccurate gene mappings that could confuse the TOGA pipeline are usually abundant. Reference-based annotation approaches such as toga strongly depend on the reference annotation quality; therefore, filtering procedure is crucial for such assemblies. To avoid reference annotation-related issues, I implemented a filter to thoroughly check the reference annotation and remove all potentially problematic transcripts. In particular, the filtering step is capable of recognizing the following cases:

- incomplete transcripts with CDS length not multiple of 3, which CESAR cannot correctly process because it requires an intact protein-coding transcript
- transcripts with annotated micro introns (shorter than 20bp) because they are usually false introns introduced to mask frameshifts
- transcripts that are shorter than 80% of the longest isoform
- potential NMD targets
- transcripts that do not start with ATG codon
- transcripts that do not end with a stop codon
- transcripts containing in-frame stop codons different from TGA, which may encode selenocysteine.

Furthermore, the filter handles transcripts with identical annotated CDS leaving one with the longest annotated UTR to avoid duplicates. However, the filter is unable to identify fused transcripts in the reference genome annotation. The reason is that actual fusion transcripts are indistinguishable from proper transcripts covering multiple reference-specific annotations using the data at our disposal. The proper distinction of these cases requires external sources of information. Notwithstanding, these filters are sufficient to prevent the vast majority of reference annotation-related issues and apply moderate-quality genome annotations as the reference for the TOGA pipeline.

3. TOGA results

In the third chapter of the present work, I evaluate the quality of different TOGA methodology aspects as follows: part 3.1 covers the creation, training, and evaluation of machine learning models for alignment chain classification. Then, part 3.2 shows the assessment of annotating and assembling fragmented genes in low-quality genomes. Subsequent part 3.3. contains the evaluation of gene loss detection accuracy. After this, part 3.4 shows the evaluation of overall TOGA annotation quality compared to state-of-the-art genome annotation methods. Part 3.5 describes how genome alignment sensitivity affects TOGA results.

Additionally, this chapter mentions projects that successfully used TOGA (part 3.6): (i) producing comprehensive genome annotations of 6 bat species (section 3.6.3) and (ii) a study of *BAAT* gene evolutionary history (section 3.6.2). To demonstrate that TOGA scales to hundreds of genomes, I describe the application of TOGA to 500 mammalian genomes in part 3.7. Furthermore, part 3.8 describes the extension to UCSC browser that I have developed in supplement to the main TOGA pipeline.

3.1 Orthology classification accuracy

In the following part, I describe the procedure of building the training dataset for chain classification (section 3.1.1). Afterward, I explain the machine learning model selection (section 3.1.2) and provide classification quality evaluation (section 3.1.3).

3.1.1 Creating training dataset to distinguish orthologous chains

To train a machine learning model for orthologs classification, this is essential to obtain a comprehensive and precise dataset of orthologs and paralogs. Albeit, at the time of writing, there is no openly available dataset suitable for our purpose - even the most advanced datasets may include mistakes and controversial cases. These mistakes, such as an erroneous identification of paralogs and orthologs, may confuse the training procedure and therefore decrease the overall prediction accuracy.

In order to generate the high-quality training dataset for our classification model, we first downloaded the most recent dataset from Ensembl BIOMART (version 99) containing one-to-one orthologs between the human and mouse with high orthology confidence. Second, to exclude potentially erroneous and controversial data points, we implemented a series of strict filters (subsection 3.1.1.1). To avoid duplicates, we considered only the longest isoform

for each gene such that each reference locus is represented only once. Furthermore, we included data points simulating translocation events to compensate for the low frequency of this event in the one-to-one orthologs data (subsection 3.1.1.2).

In detail, we created a dataset of the following structure: for each incorporated gene-chain pair that uniquely identifies a locus in the query genome, we computed a set of characteristic features (subsection 2.2.2). Then we split the resulting data points into two classes: positives, which presumably represent orthologous connections, and negatives, which are likely to represent a non-orthologous alignment.

3.1.1.1 Filters and positive - negative chains

As for positives (orthologous chains), we selected those chain-gene pairs, where (i) the chain is the top-level (highest-scoring) chain covering the gene, and (ii) the chain represents a factual orthologous alignment of the gene. To satisfy the latter condition, we require that the Ensembl-annotated one-to-one ortholog in the mouse is located at the query coordinates provided by this chain. To obtain negatives (non-orthologous chains that typically represent alignments to paralogs or processed pseudogenes), we reasoned that other chains overlapping exons of one-to-one orthologous genes by definition could not represent co-orthologs. Consequently, we added such gene-chain pairs to the negative set.

To avoid including negative chains that cover only an insignificant fraction of a gene, we only considered non-orthologous chains, where aligning blocks overlap at least 35% of the CDS. Furthermore, we only considered chains with an alignment score of at least 7500 and genes whose coding exons overlap less than 75 chains for the positive and negative sets. The latter requirement is implemented to exclude genes belonging to big gene families because the proper determination of homology class could be non-trivial for such genes.

UCSC genome browser screenshots shown in figures 3.1 and 3.2 demonstrate this principle. In example A (figure 3.1), as Ensembl reads, the illustrated human transcripts are one-to-one orthologs between human and mouse genomes. The orthologs of these genes annotated by Ensembl are located on chromosome 3 in the mouse genome at the coordinates corresponding to the top-level chain. Since the above-defined criteria are satisfied, we used the gene-chain pairs, including this chain, as positive instances in the training data. Since one-to-one orthology relationship implies that other exon-overlapping chains cannot represent co-ortholog, we used remaining chains as negative (non-orthologous) training data. Indeed, these chains represent alignments to paralogous or processed pseudogenes.

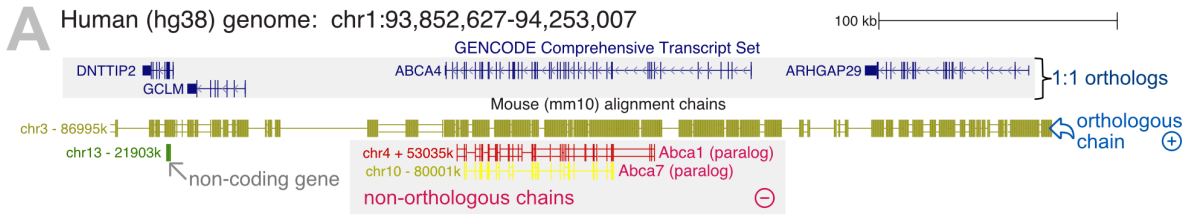

Figure 3.1 Producing training dataset, example A

The genes shown on this UCSC browser screenshot are classified as one-to-one orthologs between human and mouse by Ensembl. Therefore, we used the top-level chain as positive training data. Other chains fulfil the negative training data, since they represent paralogs.

The second example (B, figure 3.2) illustrates a comparable situation: the top-level chain provides coordinates of Ensembl-annotated orthologs in the mouse genome located on chromosome 6. Accordingly, we included this chain into our training data as the orthologous set representative. Other chains visualized in this screenshot actually represent alignments to paralogs and processed pseudogenes; thus, we used them as negative (non-orthologous) data points.

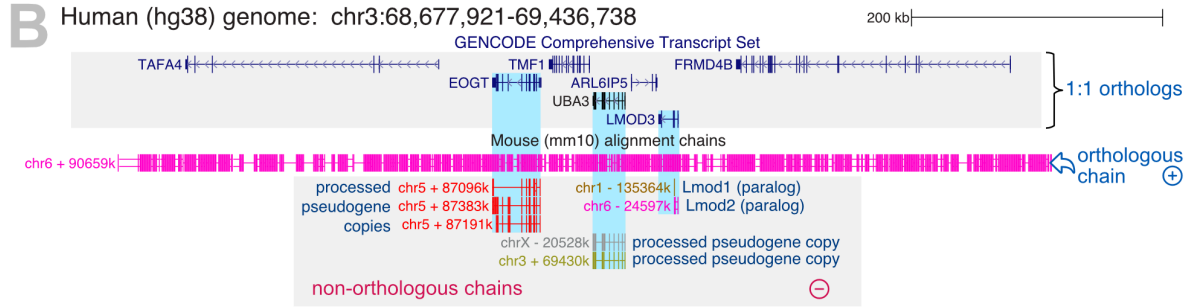

Figure 3.2 Producing training dataset, example B

Similar to figure 3.1: top-level chain represents an Ensembl one-to-one ortholog and was used as positive training data, the rest of chains were added to the negatives set.

3.1.1.2 Augmenting translocations

After creating the human-to-mouse orthologous gene set, we noticed that one-to-one orthologs aroused by inversions or translocations are underrepresented in the resulting data. This underrepresentation could result in the model overfitting towards high synteny values and following lack of accuracy in detecting translocated orthologs. In order to avoid this, we enriched the positive training dataset with artificially rearranged chain-gene pairs.

To produce the set imitating translocation events, we trimmed long syntenic chains to shorter single gene-covering chains. In particular, we considered all one-to-one orthologs whose orthologous chain is among the set of top 100 scoring orthologous chains already appeared in the positive training set. To determine the breakpoints of an artificial

rearrangement for each of these genes, we selected gene start and end coordinates and shifted them by a random number ranging from -3000 to 10000. As a result, the artificial rearrangement may even lack some parts of the gene's beginning or end. However, to avoid cases where the artificial rearrangement lacks most of the coding sequence, we only considered artificial rearrangements that include at least 80% of the CDS (figure 3.3).

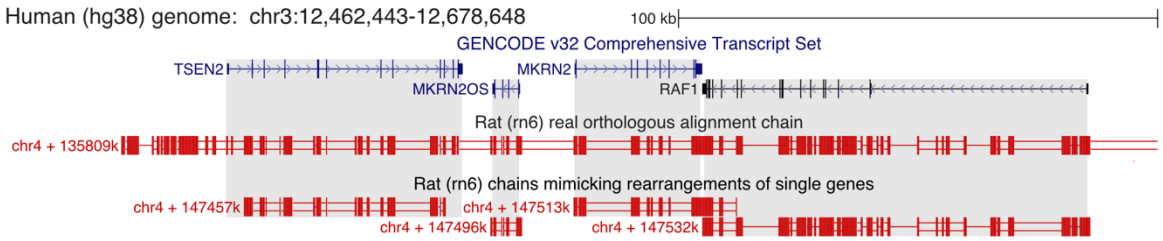

Figure 3.3 Augmenting the training dataset with artificial translocation events

The UCSC genome browser screenshot shows that top-level alignment chain to the rat genome covers all four human genes located in this 216kb locus. The entire chain spans 12.5 Mb locus of the human genome, which includes dozens of genes. Therefore, this chain exhibits a very high synteny value and provides excellent material to fulfill the augmented dataset. The resulting gene-covering chains obtained by syntenic chain split are visualized on the track below. Since the shorter chains lack synteny and intergenic alignments, they perfectly mimic the actual translocation events.

To produce the final training dataset with balanced proportions, we combined 14376 real orthologous and 5844 artificially rearranged gene-chain pairs as the positive set and considered 20220 randomly chosen gene-chain pairs as the negative set. To create independent test datasets, we applied the same procedure to genome alignments of different query species, such as human-to-rat, human-to-dog, and human-to-armadillo.

3.1.2 Machine learning model selection and optimization

Since we aim to separate a given set of gene-chain pairs (determining genomic loci in the query genome) into orthologs and paralogs, we encounter the ordinary binary classification problem. Binary classification methods have a long history and found applications in a great variety of fields such as medical testing (Esteva et al., 2017), quality control in industry (Aminzadeh and Kurfess, 2019), anti-spam filters (Dada et al., 2019), and many others. Since the binary classification problem is remarkably widespread, numerous algorithms, including those based on machine learning concepts, were developed to solve it. The set of machine learning algorithms that solve classification problems include but is not limited to approaches such as decision trees, support vector machines (Cortes and Vapnik, 1995), linear regression (Stigler, 1986), ensemble algorithms (Polikar, 2006), and neural networks (NNs) (Krizhevsky

et al., 2017). These methods vary in terms of complexity, accuracy, and applicability to different situations.

We classify gene-chain pairs using a fixed number of features - hence, our data is structured. On structured data, ensemble approaches such as gradient boosting and random forest typically outperform resource-demanding neural networks (Hu et al., 2020), whereas NNs perform better on unstructured data such as images, sound, or natural text (Sutskever et al., 2014; Krizhevsky et al., 2017). This observation designates that ensemble-based machine learning algorithms provide a reasonable solution for our type of challenge.

Ensemble algorithms are based on the concept that multiple weak classifiers such as decision trees obtain significantly better performance than any constituent learning classifier alone (Opitz and Maclin, 1999). The most extensively applied ensemble algorithms are Random Forest (Tin Kam Ho, 1995) or Gradient Boosting (Mason et al., n.d.).

Briefly, a random forest classifier consists of multiple decision trees where each of them is trained individually on a random subset of the training data. Distinct random forest algorithm implementations offer various procedures of random dataset splitting and individual tree learning. Random forest is a voting method: to produce the prediction, it aggregates predictions from each separate tree. The result supported by the highest number of votes represents the prediction of the entire model. Within this classifier, individual trees compensate for each other's errors, providing an outstanding prediction quality. Furthermore, since each tree can be evaluated individually, this approach provides substantial parallelization abilities. Therefore, random forests have a plethora of practical applications.

Gradient boosting on trees depicts a more advanced class of ensemble prediction algorithms. Gradient boosting is also based on multiple weak decision trees; however, it works in a forward stage-wise manner, compensating the shortcomings of previously generated trees before including a new one into the sequence of estimators. The exact procedure to evaluate the shortcomings of already added trees varies between different gradient boosting implementations, although the general idea is to minimize the loss function using the gradient descent. In contrast to random forests, gradient boosting algorithms aggregate ensemble results sequentially: the outcome of one model is the input to the next one. Therefore, gradient boosting algorithms cannot be parallelized as efficiently as random forests; however, some heuristics to achieve the parallelization exist. Typically, gradient boosting algorithms demonstrate higher classification performance over random forests. However, for some situations, this statement is somewhat disputed.

Out of numerous ensemble machine learning classification methods, we selected the cutting-edge XGBoost (Chen and Guestrin, 2016) gradient boosting algorithm - a method that combines parallelization, optimizations, tree pruning, and many other features. This method provides the best in the class performance for various tasks, including those comparable to

our chain classification challenge. In the following subsection, I cover the XGBoost model hyperparameters optimization process and learning procedure. After, I review individual feature impact on the prediction process.

3.1.2.1 XGBoost hyperparameters optimization

To achieve the optimal classification accuracy, we adjusted the principal model parameters such as the number of trees, the learning rate, and tree depth using the conventional cross-validation procedure. Cross-validation is a resampling procedure that randomly splits the whole dataset into training and testing subsets given the number of times. In particular, we used 80% of the dataset for training and 20% for validation. Using this procedure, I evaluated various combinations of model hyperparameters on random subsets of our data. As a result, I selected the following combination of model parameters as the most optimal for our task: (i) 50 decision trees with (ii) a maximal depth of 3 and (iii) a learning rate of 0.05 for both multi- and single-exon classifiers.

3.1.2.2 Feature importance

To identify what features contribute to the predictive performance of the entire model, I computed the "gain" value (Chen and Guestrin, 2016). This value aggregates the contribution of each feature for each tree in the model. The higher value of this metric indicates the higher relative importance of a given feature for generating a prediction. The plot showing the "gain" value for both multi- and single- exon models is shown in figure 3.4.

Global CDS fraction exhibits a substantially higher “gain” value that other features. It implies that this feature has the most significant impact on the prediction process in both multi- and single-exon modes than any other feature participating in the classification. Basically, this feature indicates how well a given chain aligns to a neutrally evolving sequence, and for the majority of classification cases, this is conclusive. Also, the data clearly shows that synteny is an auxiliary but not determining feature. Figures 3.8, 3.9, and 3.10 in the manual results evaluation subsection (3.1.3.2) show several examples of actual single gene rearrangements that were correctly classified as orthologous by TOGA, confirming the lower importance of synteny for the decision-making procedure.

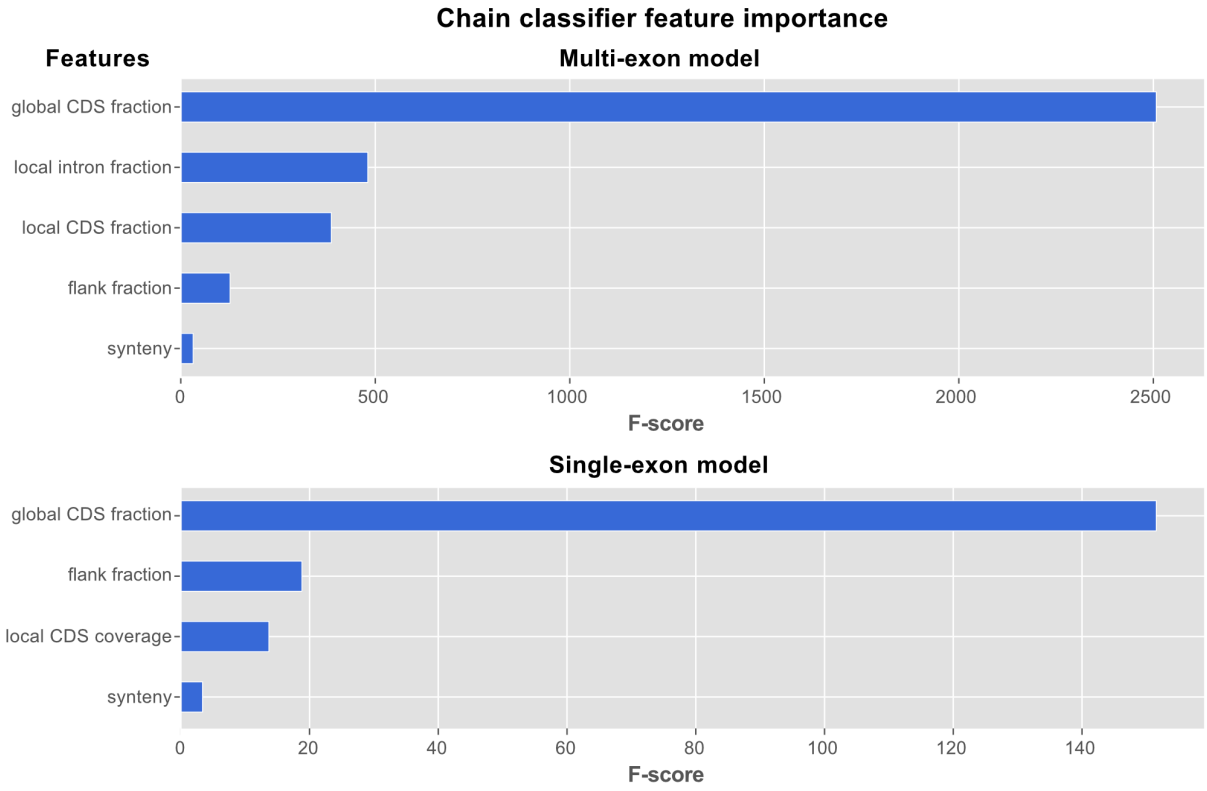

Figure 3.4 Feature importance

The plot shows the "gain" value computed for each feature applied in multi- and single-exon models. The higher value indicates the higher feature importance in generating a prediction. The plot explicitly shows that the "global CDS fraction" feature is the most critical feature for both models. It also illustrates that synteny is not a determining feature.

3.1.3 Evaluation of prediction accuracy

To evaluate the classification model's performance, we built a ROC curve and calculated the area under it, which is the standard approach for binary classifiers evaluation. Additionally, we manually analyzed the cases of detected misclassification (subsections 3.1.3.3 and 3.1.3.4). Overall, we are completely satisfied with the quality provided by our model.

3.1.3.1 ROC curve

The ROC curve is created by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various discrimination thresholds for a considered binary classifier. It applies to binary classifiers providing the evaluations as a float number from 0 to 1. For example, suppose the discrimination threshold for such a classifier is set to 0.9. In that case, all items that gained a value ≥ 0.9 are classified as positives, and the rest are assigned to the negatives class. True Positive Rate or recall is defined as TP / (TP +FN), and False Positive Rate is defined as FP / (FP + TN). Here, TN stands for a quantity of true negative classifications, FP - false positives, and FN - false negatives. The ROC curve is built as follows. For a discrimination threshold of 0, all items are classified as positive; hence both TPR and FPR equal to 1. In contrast, the threshold of 1 results in TPR and FPR both equal to 0. Then, one may compute TPR and FPR values for each threshold from 0 to 1 and, as a result, obtain the ROC curve.

Area Under ROC Curve (ROC-AUC) provides an aggregate benchmark of binary classifier performance across all possible discrimination thresholds. The ROC-AUC value ranges from 0 to 1 and could be interpreted as the probability that the model ranks a random positive item higher than a random negative item. The absolutely precise model that makes no mistakes has a ROC-AUC value of 1, and a model based on the flipping coin would have the value of 0.5. A model that always makes mistakes would demonstrate the ROC-AUC of 0. However, the inversion of model predictions could transform it into an ideal model.

To evaluate the model quality (trained on mouse), we built ROC curves based on TOGA annotation of three mammalian genomes: rat, dog, and armadillo. For each assembly, we evaluated both multi- and single- exon models. Consequently, for each model evaluation, we computed AUC for the entire gene set and artificially translocated genes only. This results in 3 * 2 * 2 = 12 ROC curves illustrated in figure 3.5.

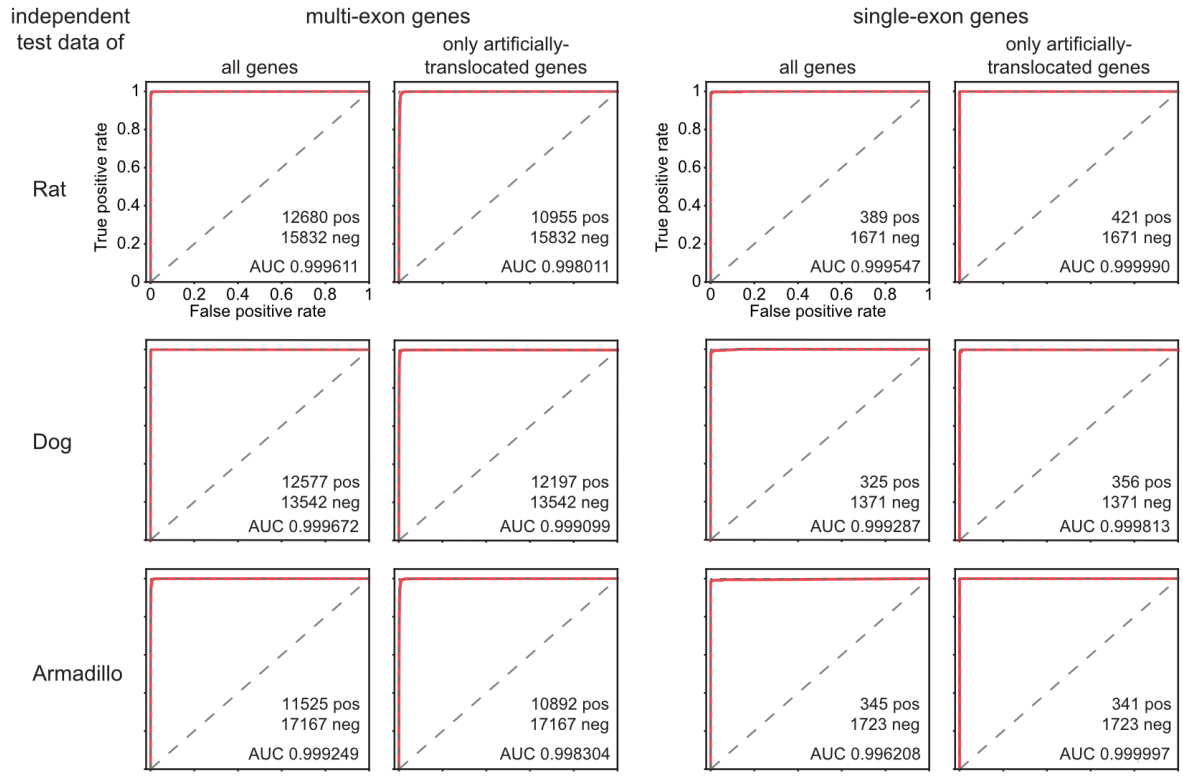

Figure 3.5 ROC curves to evaluate classification quality

The model was trained on the alignment between human and mouse genomes and validated on fully independent datasets. AUC value is >0.998 in all tests, suggesting that both models provide nearly ideal classification quality.

In fact, the selected features for chain classification are already very distinctive. To demonstrate this, I built a primitive classifier based on a single feature and applied a threshold to separate positives from negatives. In particular, I utilized the most important "global CDS fraction" feature and evaluated its predictive power for both single- and multi-exon genes. For multi-exon genes, the primitive classifier showed an accuracy of 96.8% with a separating threshold of 0.3 (figure 3.6). Further, I applied the same procedure for the single-exon genes classifier. This classifier also demonstrated a great predictive power with accuracy of 95.9% (figure 3.7).

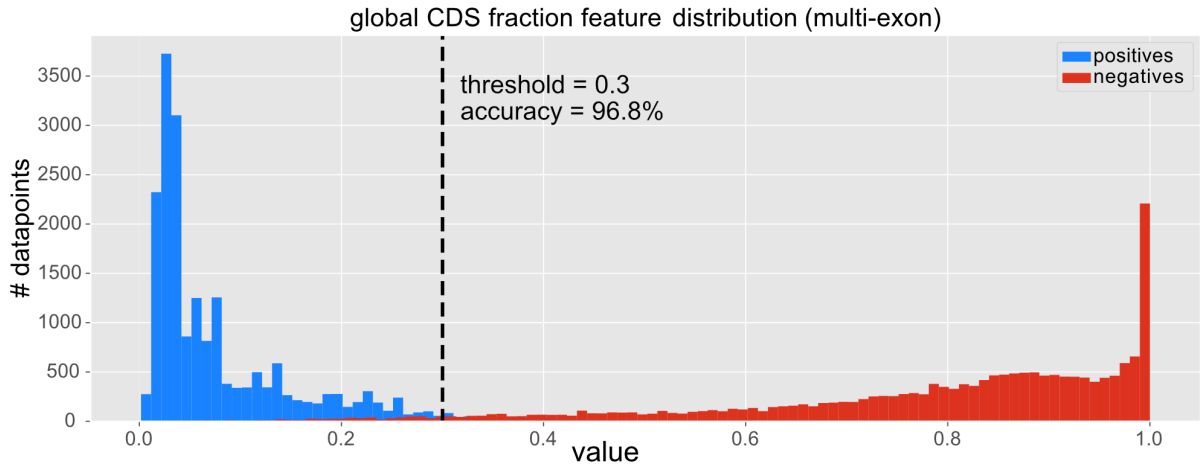

Figure 3.6 The most distinctive feature of the multi-exon model

The plot illustrates the "global CDS fraction feature" distribution in positives (blue) and negatives (red) in the multi-exon dataset. A primitive classifier based only on this feature with the threshold of 0.3 provides a classification accuracy of 96.8%.

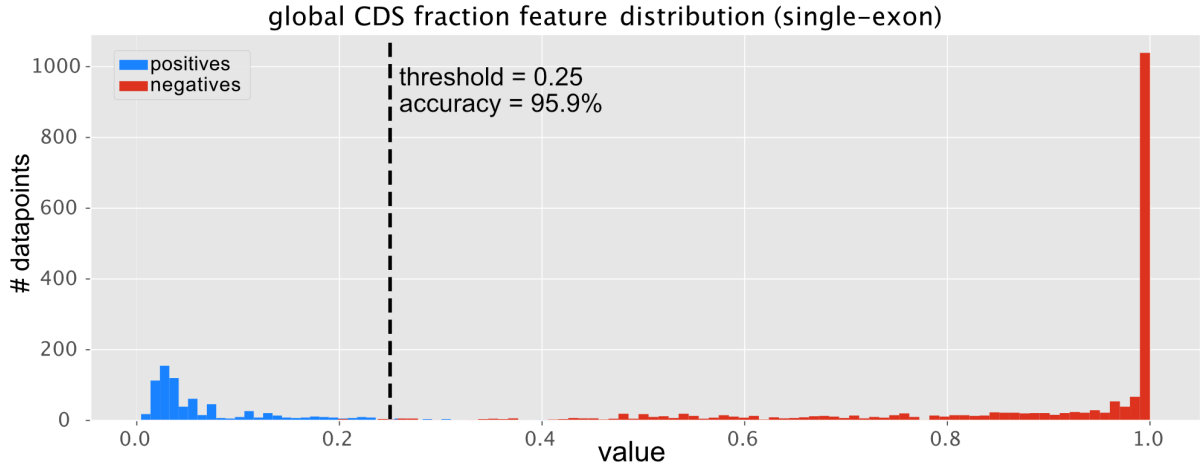

Figure 3.7 The most distinctive feature of the single-exon model

Similar to 3.6, but for the single-exon dataset. With a threshold of 0.25, this feature solely provides the classification quality of 95.9%.

Indeed, if a single feature provides high predictive power, it would be fair to expect that a combination of distinctive features has the potential to perform even better. The best gradient boosting algorithm in the class realized this potential, which emerged in the presented classification quality. Notwithstanding, we conducted a manual analysis of the results to confirm that this incredible accuracy is not a consequence of evaluation mistakes. For manual quality evaluation, we arbitrarily selected cases falling into the following classes:

1. True positive cases where the synteny feature value was very low (subsection 3.1.3.2)
2. Cases resulting in a false-positive outcome (subsection 3.1.3.3)
3. Cases of false-negative results (subsection 3.1.3.4)

3.1.3.2 Low synteny true positive examples

To confirm that TOGA is able to identify orthologous genes that underwent actual rearrangements correctly, we manually inspected genes with low synteny feature values. Figures below show UCSC browser screenshots illustrating three selected examples.

Example A (figure 3.8) illustrates the locus in the human genome containing genes *FI44L* and *IFI44*. In contrast to the mouse and other mammals, the *IFI44L* gene is inverted in rats. The second-level chr2+ chain representing the ortholog in the rat genome exposes the inversion because it is located on the opposite strand to the top-level chain (chr2-). This local inversion breaks co-linearity with the surrounding alignments, resulting in a chain that covers only this gene. Consequently, the synteny feature of this chain has value 1. Nevertheless, due to training the classifier on a dataset augmented with artificial single-gene rearrangements and using other features in addition to synteny, TOGA correctly classifies this chain as the ortholog with the probability of 0.98. The third-level chain represents the alignment between human *IFI44L* and the rat paralog IFI44, and the model correctly classified this chain as non-orthologous.

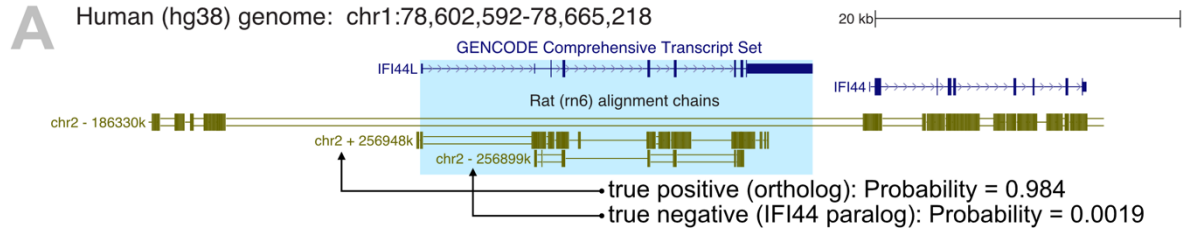

Figure 3.8 Correct classification of translocated ortholog, example A

UCSC browser screenshot shows locus in the human genome containing genes IFI44L and IFI44. The second-level alignment corresponding to inverted ortholog in the rat genome was correctly classified as such by TOGA.

In example B (figure 3.9), we consider the RRP7A gene locus in the human genome. The second level alignment chain represents the ortholog and indicates that it is translocated in the rat genome. Importantly, inspecting chains of other mammals (for clarity not shown here) revealed that this rearrangement happened along the primate lineage before the great apes split. Thus, the rearrangement occurred in the lineage leading to the reference species. As in example A, the chain covers only *RRP7A*, and its synteny feature has the value of 1. Nevertheless, due to intronic and gene-flanking alignments, the model can correctly identify this chain as the ortholog with a high probability of 0.99.

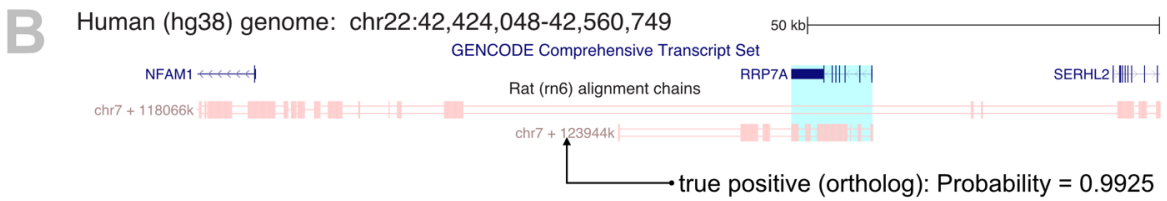

Figure 3.9 Correct classification of translocated ortholog, example B

Similar to 3.8. TOGA correctly identified the second-level chain as orthologous.

Additionally, Example C (figure 3.10) considers a single-exon gene *TACSTD2*. The chain representing the ortholog in the rat genome covers only this gene, together with upstream and downstream flanks. The reason is that *TACSTD2* was translocated to a different locus in the rat (this translocation is shared with the mouse, not shown here). Features quantifying the amount of gene-flanking alignments enable TOGA to correctly classify this chain as orthologous with a probability of 0.94, despite the lack of conserved gene order.

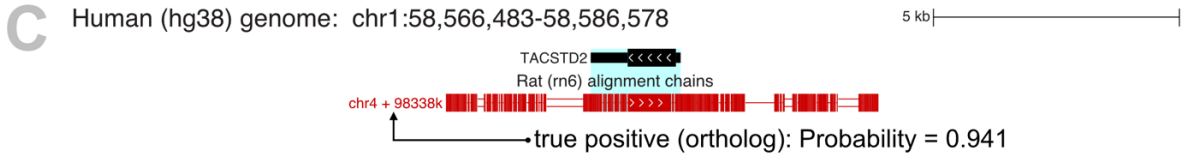

Figure 3.10 Correct classification of translocated ortholog, example C

Similar to 3.8 and 3.9, TOGA correctly identified the translocated TACSTD2 copy in the rat genome.

These results suggest that TOGA is able to detect actual translocation events in the absence of synteny. Specifically, low synteny values do not lead to misclassification of co-orthologs as paralogs. In the following subsections, I consider examples where TOGA prediction contradicts the evaluation dataset.

3.1.3.3 Analysis of false positive misclassifications

Following, we inspected the set of false-positive misclassifications - here, TOGA classified representatives of the negative group as positives. In this subsection, we present two selected cases. Example A (figure 3.11) considers the locus in the human genome that comprises the *KNOP1* gene. Ensembl lists this gene as a one-to-one ortholog between human and rat genomes. Alignment chains indicate that the rat genome contains at least four homologous sequences. Here, TOGA correctly classified the top-level chain representing the Ensembl-annotated ortholog of KNOP1 as such (probability >0.99).

Two other chains (aligned to chrX and chr2) represent processed pseudogene copies of KNOP1 and are also correctly classified as non-orthologous ones (probabilities < 0.004). Notably, the fourth chain (chr12) shows that parts of *KNOP1* and the neighboring *IQCK* gene were duplicated in the rat genome. Nearly identical chain block structure in the first and fourth

chains supports this statement. The duplicated region is located ~200 kb upstream of the original *KNOP1* locus in the rat genome. Since Ensembl annotated *KNOP1* as a one-to-one ortholog between human and rat, we labeled all exon-overlapping chains except the first one as negatives in the test dataset (including this fourth chain).

However, this chain factually represents a co-ortholog (lineage-specific duplication), and TOGA indeed estimates a high probability of 0.996 that this chain represents an orthologous locus. Supporting this, Ensembl (*ENSRNOT00000075020*.1) does annotate a shorter 380 amino acid-comprising *KNOP1* gene at the duplicated locus. Nevertheless, Ensembl does not classify this gene as a co-ortholog to human *KNOP1*. This leads us to the conclusion that the fourth chain was mislabeled as a negative in the testing data. Even though TOGA correctly classified this chain as co-orthologous, we conservatively classified this case as false positive.

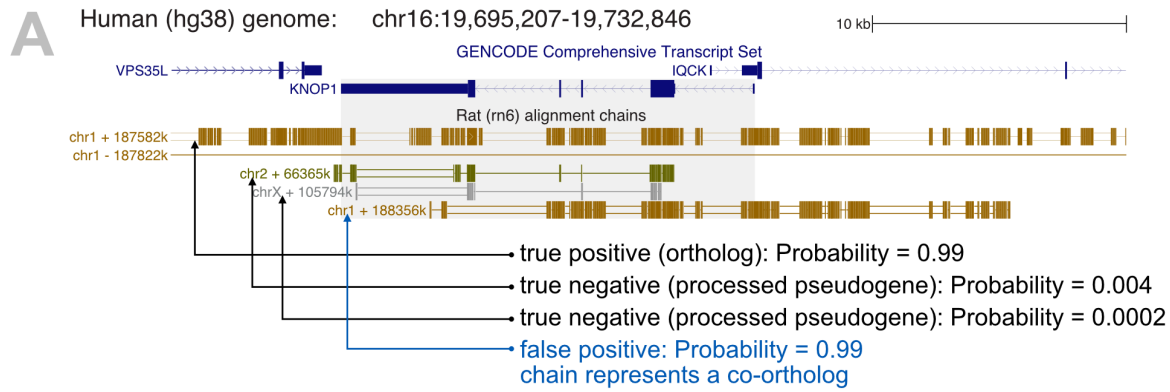

Figure 3.11 False-positive chain misclassification, example A

UCSC genome browser screenshot shows human genome locus containing KNOP1 gene. TOGA correctly classifies the top-level chain as orthologous - Ensembl also annotates the KNOP1 ortholog in the corresponding locus. However, TOGA also classifies another chr1 chain as co-orthologous alignment. Ensembl annotates a gene in the corresponding locus but does not classify this as a co-ortholog. Being conservative, we classified this case as a false-positive prediction.

In example B (figure 3.12), we consider the *PDIA5* gene locus. As previously, Ensembl reports a one-to-one orthology relationship between human and rat *PDIA5* genes, and TOGA correctly classifies the respective top chr11 chain as positive (probability >0.999). However, the second chr11 chain indicates that the upstream part of the PDIA5 locus is duplicated in the rat genome. As in example A, we mislabeled the second-level chain as negative in our test data, relying on the Ensembl ortholog annotation of *PDIA5* even though this chain indeed represents a co-orthologous locus. Thus, TOGA correctly identifies this chain as an orthologous sequence alignment with a high probability of 0.999.

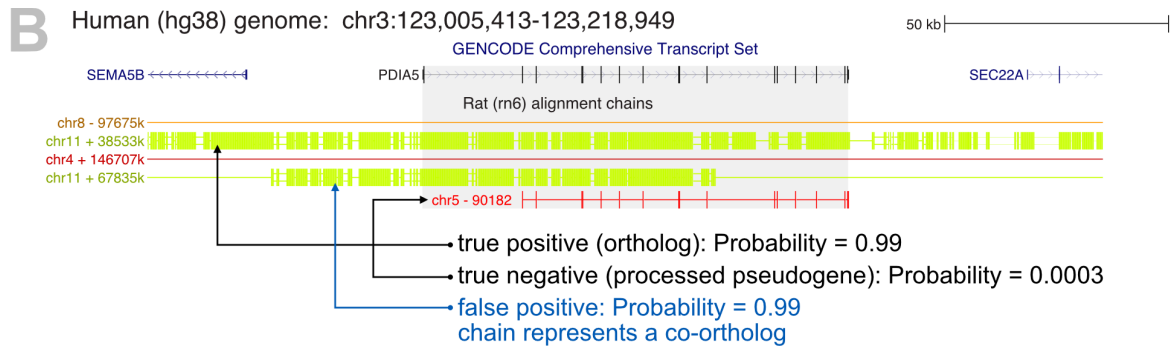

Figure 3.12 False-positive chain misclassification, example B

Similar to 3.11. TOGA identifies two orthologous chains for the human PDIA5 gene. According to Ensembl, PDIA5 has only one ortholog in the rat genome. Therefore, we conservatively identified this case as false-positive. However, the TOGA-identified co-orthologous locus indeed comprises a gene, implying that TOGA is likely correct.

In general, manual inspections show that false-positive chain misclassifications often represent partial co-orthologs arising from partial gene duplications and that these chains are actually mislabeled as paralogs in the test dataset. This indicates that TOGA correctly classified these chains as aligning to a query locus, potentially containing an ortholog. Since the orthologous loci identification precedes the transcript annotation, TOGA evaluates the actual presence of protein-coding genes within these loci in the following pipeline steps.

3.1.3.4 False negative misclassifications

In this subsection, we consider the opposite class of misclassifications: false-negatives. It implies that TOGA could not identify orthologous loci proposed by Ensembl. Here, I review two selected examples of false-negative misclassifications in detail.

Example A (figure 3.13) shows the case of the *PNPLA4* gene. Ensembl correctly annotates this gene as a one-to-one ortholog between human and rat genomes. Since chromosome X lacks intronic and intergenic alignments, this chain appears to be a typical paralogous chain. However, this chain, in fact, represents the orthologous locus of *PNPLA4*. TOGA incorrectly classifies the chain as a non-orthologous locus. The reason for complete sequence divergence of all intronic sequences is unknown; however, it could be connected to faster X chromosome evolution (Vicoso and Charlesworth, 2006; Charlesworth et al., 2018). Notably, most of the ~100 orthologs misclassified by our model are located on the X chromosome.

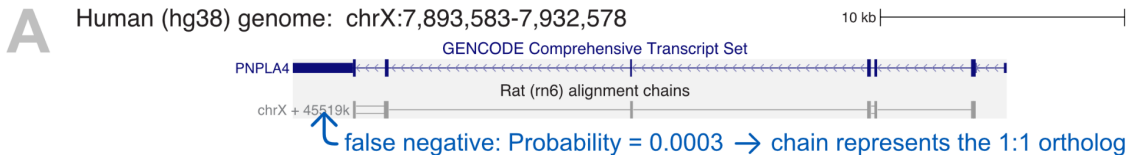

Figure 3.13 False-negative chain misclassification, example A

UCSC genome browser screenshot shows human PNPLA4 gene and rat alignment chains. Due to extreme neutral sequence divergence on chromosome X, TOGA misclassified the orthologous chain as paralogous.

Example B (figure 3.14) provides a closer look at the case of the *NURP2* gene. According to Ensembl, this single-exon gene has a one-to-one ortholog in the rat genome, and as in the previous example, the chr12 chain represents the actual orthologous locus. However, this chain shows exceptional intron divergence and lack of gene order. Consequently, TOGA incorrectly classified it as non-orthologous.

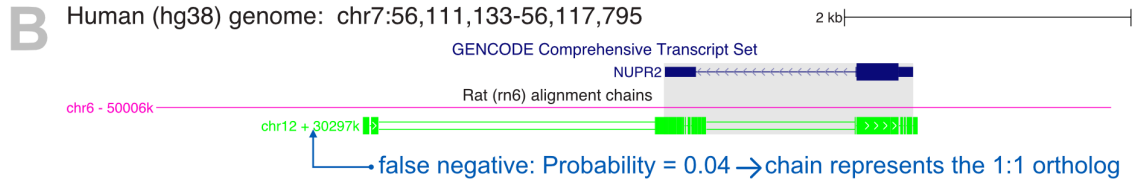

Figure 3.14 False-negative chain misclassification, example B

Similar to 3.13, TOGA misclassified a truly orthologous chain due to extreme divergence of neutrally evolving regions.

It is perhaps worth mentioning that TOGA does annotate a gene at the chrX locus in (A) and the chr12 locus in (B) since the respective loci in the human genome receive no annotation through an orthologous chain. However, the annotation is labeled as a paralogous projection as the chains used for annotation were classified as non-orthologous. As a result of the analysis of false-negative classifications, we found that the extraordinary divergence of neutral sequences often characterizes false-negative chain misclassifications. These examples highlight limitations of the genome alignment-based orthology inference method implemented in TOGA.

3.1.3.5 Manual evaluation summary

TOGA demonstrates a remarkably high accuracy in the detection of orthologous loci. Many of the reported mistakes are explained by inaccuracies in the testing data. However, the approach has specific weaknesses - with the TOGA approach, orthologs that exhibit unusual divergence of neutrally evolving sequences are indistinguishable from paralogs. Besides, the TOGA pipeline includes additional steps following the orthologous loci identification. In the following parts, I provide the quality assessment of these subsequent steps.

3.2 Fragmented Genomes Handling Accuracy

As stated in section 2.2.6, the TOGA pipeline includes the functionality to detect fragmented genes split between different scaffolds in the query genome and join the respective orthologous loci to assemble the gene from these pieces. Here I evaluate the accuracy of this approach by applying TOGA to a highly fragmented Kogia (pygmy sperm whale) genome draft assembly (contig/scaffold N50 of 26kb/28kb). A close Kogia relative belonging to the same family (Physeteridae), the sperm whale, has an assembly of much longer contiguity (contig/scaffold N50 of 42kb/122Mb), providing an excellent reference for approach evaluation.

We compared how similar are assembled Kogia and undivided Physeter genes to evaluate the accuracy of this step. Kogia genes located on a single scaffold have a median nucleotide sequence identity of 98.76% to Physeter genes (figure 3.15). If the assembled gene exhibits a similar nucleotide identity, we could conclude that the procedure is accurate. Otherwise, if the procedure makes mistakes and assembles genes from non-homologous fragments, we expect sequence similarity to drop. However, I found that genes assembled from 2, 3, or 4+ fragments exhibited essentially the same sequence similarity as genes that are already contained on one scaffold, indicating that this TOGA step performs accurately.

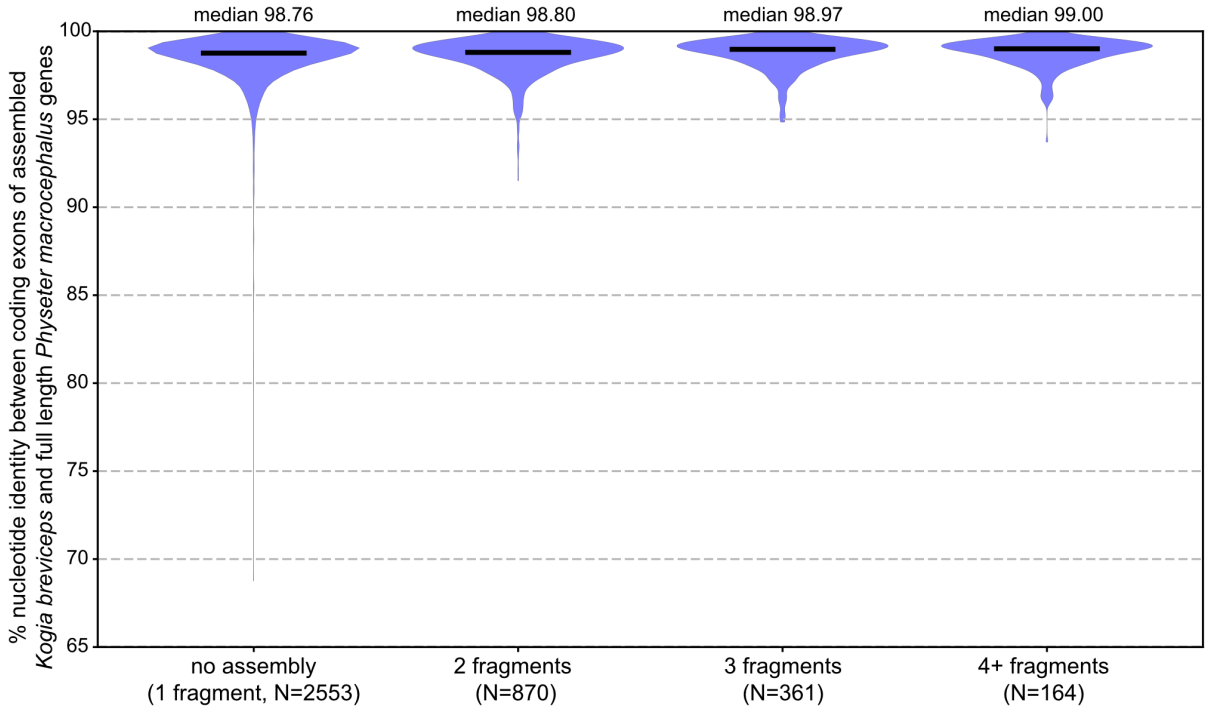

Figure 3.15 Fragmented genes assembly quality evaluation

The plot shows nucleotide sequence identities between human orthologs detected in Kogia and the sperm whale. The first subplot (on the left) shows sequence identity distribution for genes located on a single scaffold. The following subplots show cases where genes in Kogia are assembled from 2, 3, and 4+ pieces. The results suggest that genes assembled from fragments exhibited the same sequence identity as already residing genes on one scaffold.

3.3 Gene loss detection quality

Gene loss detection functionality implemented in TOGA was inspired by the methodology previously developed in our lab (Sharma et al., 2018). It follows the same pattern involving the extensive search of loss-of-function mutations and computing %intact reading frame features for making a prediction. However, this work was pioneering in this field, and in the newer implementation, we introduced some changes to increase the method performance (section 2.4.3). A crucial difference is that TOGA accounts for multiple orthologous loci while the previous pipeline utilized only the top-scoring chains. Henceforth, TOGA can detect an intact co-ortholog in the query genome even if it is inactivated in the ancestral locus, whereas the previous implementation could misclassify these cases as gene loss.

The previously implemented pipeline already demonstrated high accuracy. To confirm that TOGA implementation of gene loss detection functionality is still accurate, we separately evaluated the quality of this step as follows. First, we evaluated the pipeline specificity, following the methodology applied in the previous work, which implies quantifying genes classified as lost in the conserved gene set. Second, we checked whether the newer implementation could replicate the findings revealed by the previous version of the method - these findings underwent manual curation and most likely represent actual gene inactivation events. Moreover, we evaluated genome assembly quality impact on the classification quality.

3.3.1 Specificity evaluation

We performed the method specificity evaluation on a large set of 11,182 human genes that are conserved in the mouse (mm10), rat (rn6), cow (bosTau8), and dog (canFam3) genomes. To obtain this set, we extracted identifiers of genes that have one-to-one ortholog between the human and each of the four analyzed species according to Ensembl database version 101. Then, we excluded genes that contain very short introns (<50bp) in any of the four considered species from this dataset. This filter is necessary because such introns usually mask assembly artifacts, such as frameshifting events and nonsense codons, or real inactivating mutations in lost genes. After all, we obtained a set of presumably conserved genes, which implies that those genes should encode a functional protein across the four mammals. Therefore, any reported gene loss in this set will be conservatively considered as a false positive.

As a result, we detected that gene loss is incorrectly inferred for 17 genes in the mouse (specificity=99.84%), 14 genes in the rat (specificity=99.87%), 13 genes in the dog (specificity=99.88%), and 10 genes in the cow (specificity=99.1%).

Manual inspection showed that most of these rare misclassifications are related to (i) incorrect reference transcript selection leading to inaccurate exon boundaries identification, (ii) extreme gene divergence in the query, (iii) Ensembl errors, and (iv) orthologs represented by a processed pseudogene. Examples of these four types are provided in the next subsections.

3.3.1.1 Importance of reference isoforms selection

The example of *CD276* gene misclassification illustrates the importance of proper reference isoform selection (figure 3.16). TOGA correctly identified the orthologous locus of this gene in the mouse genome; however, it identified the projected gene as lost. Indeed, the ortholog's reading frame is disrupted by numerous inactivating mutations, occupying a significant fraction of the CDS in the middle of this gene. The localization and number of inactivating mutations suggest that this gene is clearly inactivated in the mouse genome, which raises the question of why Ensembl annotated it as intact. However, we noticed that the detected inactivating mutations occurred only in exons 3 and 4 and that there is an alternative non-APPRIS isoform, which was excluded from the reference input annotation, that does not contain exons 3 and 4.

Since we pursued the highest quality of the reference annotations to avoid ambiguous results, we utilized only the APPRIS isoforms. For an unclear reason, the alternative isoform was not included in the APPRIS database; therefore, we neglected it in our analysis. Afterward, we inspected the alternative isoform ortholog in the mouse genome and did not detect any gene-inactivating mutations with TOGA, which indicates that it is clearly intact. Summarizing that, we may conclude that this misclassification was induced by incorrect reference isoforms selection.

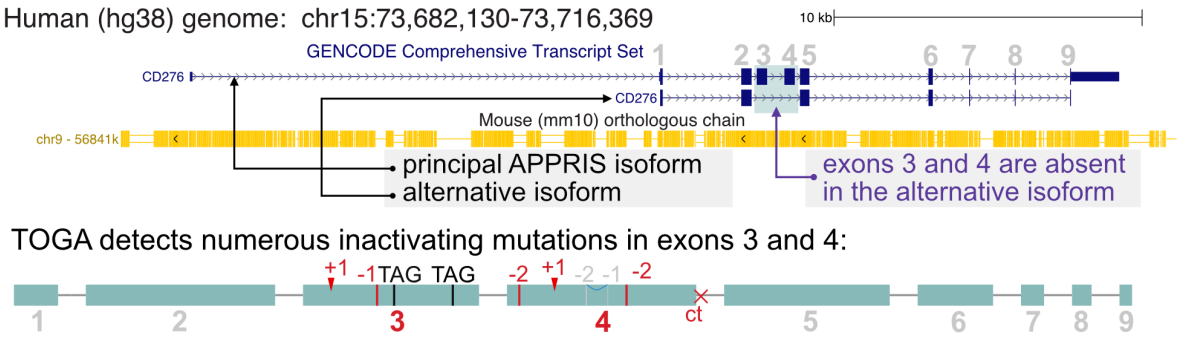

Figure 3.16 Incorrect isoform selection leads to gene misclassification

UCSC genome browser screenshot shows two isoforms of the human CD276 gene. The longest isoform comprises exons 3 and 4, which carry inactivating mutations in the orthologous locus, and is classified as lost. However, the alternative isoform does not include these exons, and as a result, orthologous projection exhibits no loss-of-function mutations and is classified as "intact."

3.3.1.2 Extreme gene sequence divergence may confuse gene loss detection quality

Another reason why a representative of our conserved gene set could be misclassified as lost is extraordinary sequence divergence. In this work, I exemplify this category with the *ESX1* gene, located on chromosome X. According to Ensembl, the nucleotide sequence identity of the human and mouse orthologs is 37%, slightly higher than the expected identity of two random sequences (figure 3.17).

A lack of sequence similarity suggests that detected inactivating mutations are possibly located in the non-homologous regions that actually do not encode the protein sequence in the mouse genome. Indeed, the 5' terminus of the gene underwent significant structural rearrangements. Two non-homologous exons compensate for the reported deletion of the first exon in the mouse genome. Also, CESAR could not identify the translation end in the 3' exon due to extreme sequence divergence in this region. Instead, it predicted a slightly longer exon introducing a +1 frameshifting insertion. In general, this case clearly illustrates the limitations of reference-based methods in annotating genes with low sequence similarity.

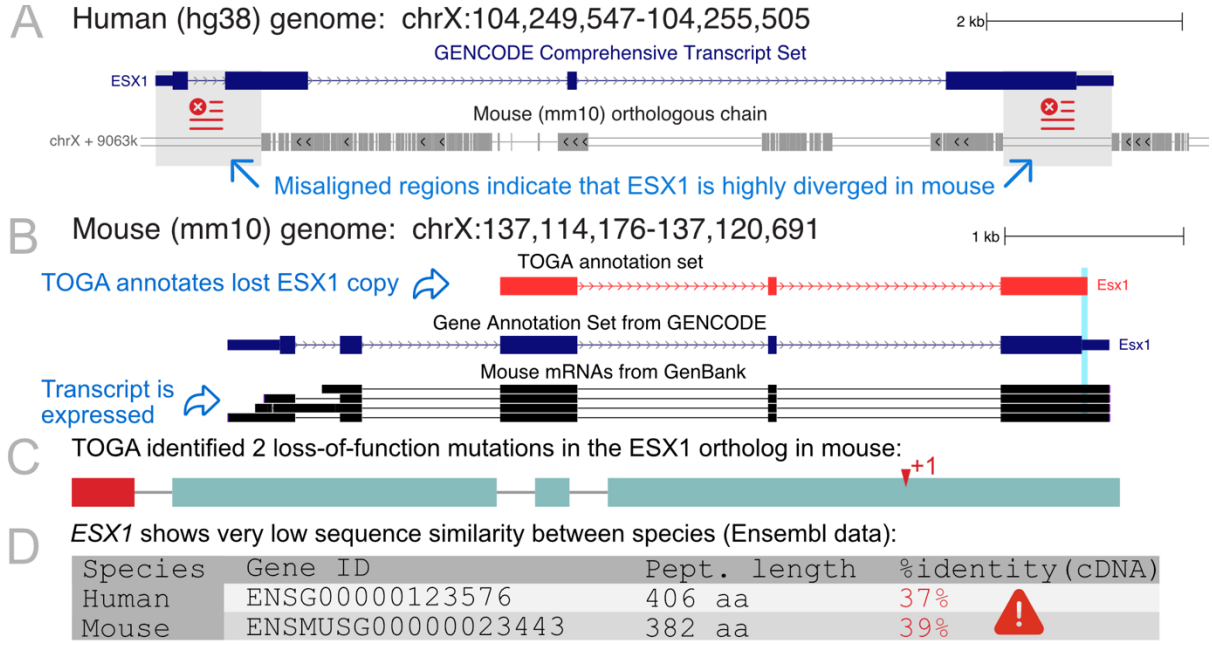

Figure 3.17 Gene misclassification induced by extreme sequence divergence

UCSC genome browser screenshot shows the ESX1 gene in human and mouse genomes. Panel A: mouse alignment chains to the human genome show that 3' and 5' termini of this gene are misaligned. B: TOGA annotates the ESX1 ortholog as lost in the mouse; however, Ensembl annotates an intact transcript. Mouse mRNA track shows that this gene is actually expressed in the mouse. Different exon composition suggests that inactivating mutations detected by TOGA are located in non-coding regions. C: Plot showing inactivating mutations detected in the TOGA projection of ESX1 gene in the mouse genome. D: screenshot from Ensembl showing that nucleotide sequence identity between ESX1 orthologs is human to mouse is lower than 40%, suggesting that this gene has highly diverged.

3.3.1.3 Errors in the conserved gene dataset

Despite our efforts to obtain an error-free conserved gene set, we included some non-orthologous genes in our analysis. For example, TOGA classified the *ZNF239* ortholog in the mouse genome as lost, and our subsequent analysis of this case suggested that it is a likely Ensembl error. This gene is a member of an abundant family of zinc finger motif containing genes represented by hundreds of genes in mammalian genomes. This gene family is challenging for orthology inference methods, and only a fraction of the total family representatives could be correctly classified with modern techniques. TOGA is not an exception, as discussed in subsection 5.1.1.3.

As shown in figure 3.18, hundreds of alignment chains cover this gene. However, in this particular case, only the top-level chain represents an orthologous alignment, and TOGA correctly recognizes it. Meanwhile, other chains primarily represent paralogous alignments or align only to the zinc finger domain. Analysis of the projected gene revealed abundant inactivating mutations, suggesting that the only one orthology candidate gene is lost.

For this gene, Ensembl mistakenly identifies another *ZNF* gene (mouse ENSMUSG*00000042097*) as the putative ortholog. This transcript prediction is located in the downstream region of the human *ZNF239* ortholog predicted by TOGA and is clearly paralogous. Presumably, gene trees could not correctly resolve the homology relationships between different *ZNF* genes.

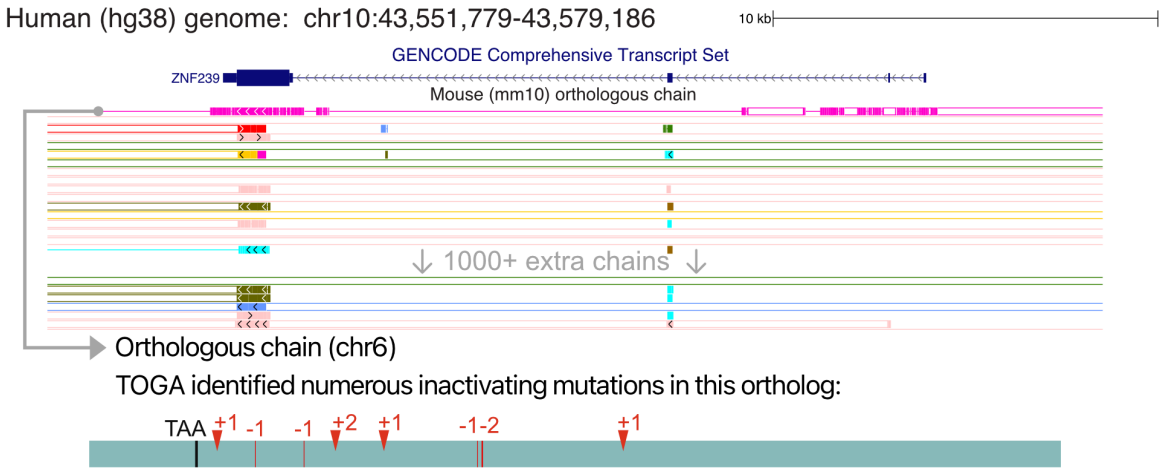

Figure 3.18 Inaccuracy in the test data

UCSC genome browser screenshot shows the ZNF239 gene in human and mouse alignment chains. TOGA identified the top-level chain as orthologous. However, the predicted transcript exhibits numerous inactivating mutations. Thus, according to TOGA, the mouse does not have a ZNF239 ortholog, which contradicts the Ensembl data. Numerous alignment chains suggest that Ensembl likely misclassified some of the paralogs as the ortholog.

3.3.1.4 Edge case: Cyclin-Q processed ortholog in mouse

In this subsection, I discuss a *CCNQ* gene case that revealed a potential TOGA weakness. In fact, TOGA classifies chains aligning to multi-exon genes that (i) are classified as non-orthologous, (ii) indicate deletions in intronic regions, and (iii) show absence of gene flanking alignment, as processed pseudogene alignments and annotates corresponding regions accordingly. These requirements imply that the aligned sequence is a cDNA copy of processed mRNAs, and usually, these regions are non-coding. Nevertheless, our implementation does not acknowledge that a minor fraction of PPs could retain activity if inserted close to an RNA polymerase II promoter (Kabza et al., 2014). However, in general processed pseudogene copies can adequately substitute the ancestral gene with a similar expression pattern only in very rare cases.

TOGA correctly identifies the top-level chain that covers the *CCNQ* gene as orthologous. Additionally, it also classifies remaining chains as alignments to processed pseudogene copies because the intronic regions in these copies are wholly deleted. Since this gene appears in our conserved gene set, Ensembl suggests that there exists an intact CCNQ one-to-one ortholog in the mouse. Presumably, the top-level chain aligns to the locus containing the ortholog predicted by Ensembl (figure 3.19).

However, TOGA detects numerous inactivating mutations in the ancestral *CCNQ* gene copy, suggesting that this gene is likely lost. Interestingly, Ensembl also annotates this copy as a pseudogene, so TOGA and Ensembl agree here. However, Ensembl annotates an ortholog of *CCNQ* in a different locus, where TOGA annotated a processed pseudogene copy (chromosome 11). Surprisingly, EST data reveals that this processed copy still encodes the RNA. Whether this processed pseudogene copy has a similar expression pattern and function as the ancestral gene is not known. Nevertheless, *CCNQ* could comprise a rare case where a processed pseudogene could substitute for a lost ancestral gene. Nevertheless, TOGA provides an annotation track for processed pseudogene copies together with identifiers of corresponding reference transcripts. In case of necessity, this data could be analyzed separately.

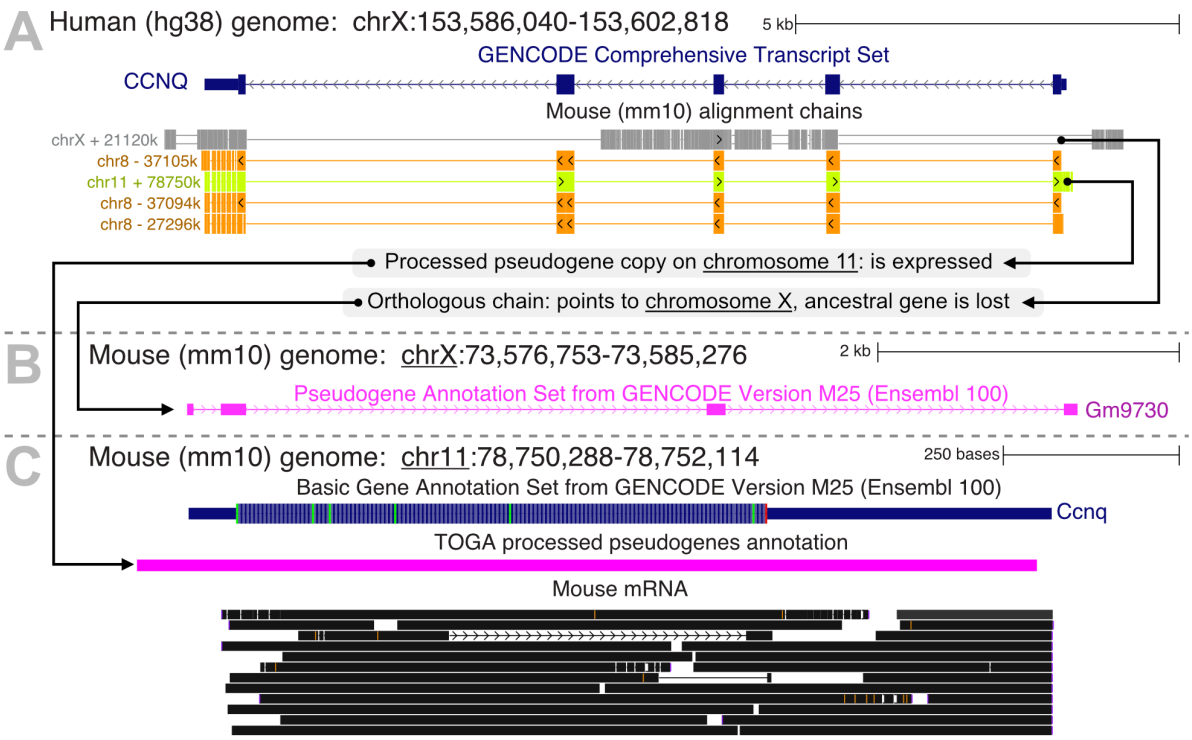

Figure 3.19 Expression of a processed copy

UCSC genome browser screenshots show different CCNQ-homologs containing loci in the human and mouse genomes. A: screenshot shows human locus containing CCNQ gene and mouse alignment chains. The top-level chain points to the ancestral orthologous locus in the mouse. Chain pointing to mouse genome locus on chromosome 11 was classified as processed pseudogene alignment since it indicates the deletion of the intronic fraction. Panel B shows the ancestral CCNQ locus in the mouse. TOGA and Ensembl annotate a pseudogene suggesting that this gene is lost. Panel C shows the processed pseudogene locus. TOGA annotates a processed pseudogene copy, whereas Ensembl annotates this copy as an intact single-exon gene. Surprisingly, transcriptomic data indicates that this copy is actually expressed and potentially substitutes the lost ancestral copy.

3.3.2 TOGA implementation of GLP reproduces previous findings

We additionally checked whether TOGA could reproduce gene loss events previously reported by our lab. Published gene losses underwent rigorous manual review, including mapping of raw sequencing reads to inactivated sequences and analysis of transcriptomic data. For instance, the gene inactivation events reported in our previous work were identified by TOGA (Sharma et al., 2018). Moreover, TOGA could reproduce findings published in our paper about gene losses associated with aquatic adaptations in Cetaceans (Huelsmann et al., 2019).

Additionally, it confirmed the inactivation of the Toll-like receptor 5 (*TLR5*) gene in four independent mammalian lineages (Sharma et al., 2020). Accordingly, this evaluation suggests that TOGA can sufficiently substitute the previous gene loss detection pipeline implementation since it is able to reproduce the earlier findings with a high degree of specificity.

3.3.3 Genome assembly quality influences gene loss classification accuracy

As I noted before, assembly artifacts could mimic inactivating mutations. Conservative gene loss criteria were introduced to minimize this issue, such that a single inactivating mutation detected due to genome assembly issues does not provide enough evidence of the gene loss. However, even with these preventive measures, truly intact genes could be misclassified as lost. To demonstrate explicit examples of this issue, we compared gene classifications for two different cow genome assemblies, GCA_000003055.5 (The Bovine Genome Sequencing and Analysis Consortium, 2009) and GCA_002263795.2 (Rosen et al., 2020). We discovered that eight genes classified as “intact” in the 2014 assembly changed their class to "lost" in the 2018 version.

Figure 3.20 illustrates a selected example for this case. The newer assembly introduces two +1 insertions in the *RRP8* gene, resulting in the classification of this gene as "Lost". However, the earlier genome assembly does not exhibit any loss-of-function mutations in this gene.

Figure 3.20 Assembly artifacts mimic inactivating mutations

The plot shows the coding exon of the RRP8 gene in the human and two different cow genome assemblies. RRP8 annotated in the newer cow genome assembly (bottom) exhibits two frameshifting mutations absent in the previous assembly (middle). Most likely, there are no actual inactivating mutations but assembly artifacts introduced in the newer cow genome assembly.

Most likely, the differences are explained by sequencing technologies used to produce these assemblies. The older one was initially produced using Sanger sequencing technology, providing data of high base accuracy, while the newer genome was sequenced with more error-prone PacBio technology. PacBio reads, being relatively long (about 15kb), are

characterized by frequent +1/-1 insertions, as the provided example demonstrates. Therefore, it is necessary to refine the PacBio-based assemblies with more precise short-reads technologies (Watson and Warr, 2019). It appears that a minor fraction of these frameshifting mutations ended up uncaptured in the newer assembly.

3.4 Overall TOGA annotation accuracy

In the preceding parts, I provided the quality evaluation of different separate aspects of the TOGA pipeline (chain classification, gene loss detection). This part focuses on the performance of the entire TOGA pipeline. To evaluate this, I compared TOGA to the gold-standard annotation dataset provided by Ensembl (section 3.4.1). Later, I evaluated the completeness of conserved gene annotations from the BUSCO gene set (section 3.4.2).

3.4.1 Comparing vs. Ensembl

To demonstrate that the TOGA provides high-quality, reliable annotations, we compared TOGA results to a gold standard dataset. The Ensembl database contains highly reliable annotations that are publicly available for numerous species. To achieve this annotation quality, Ensembl applies various techniques, including gene trees and alignment of biological sequences such as cDNAs, proteins, and RNA-seq reads (Curwen, 2004; Aken et al., 2016). Because of that, the Ensembl data is widely applied in numerous studies. It is perhaps worth mentioning that we applied Ensembl data to establish and validate different aspects of the TOGA pipeline, but to avoid circularity, training (human-mouse) and test datasets are independent.

At first, we compared the number of human orthologs that TOGA and Ensembl detected in the rat genome. Surprisingly, despite the entirely different genome annotation approach, two methods demonstrated a high degree of agreement - both methods detected orthologs of 16617 human genes (figure 3.21, panel A). This evidence suggests that the TOGA approach is entirely viable. However, there are 415 human genes for which Ensembl identified an ortholog in the rat genome, but TOGA did not. On the other hand, TOGA exclusively annotated 1336 orthologs that did not appear in the Ensembl orthology database. In total, two methods could annotate 18368 human genes in the rat genome.

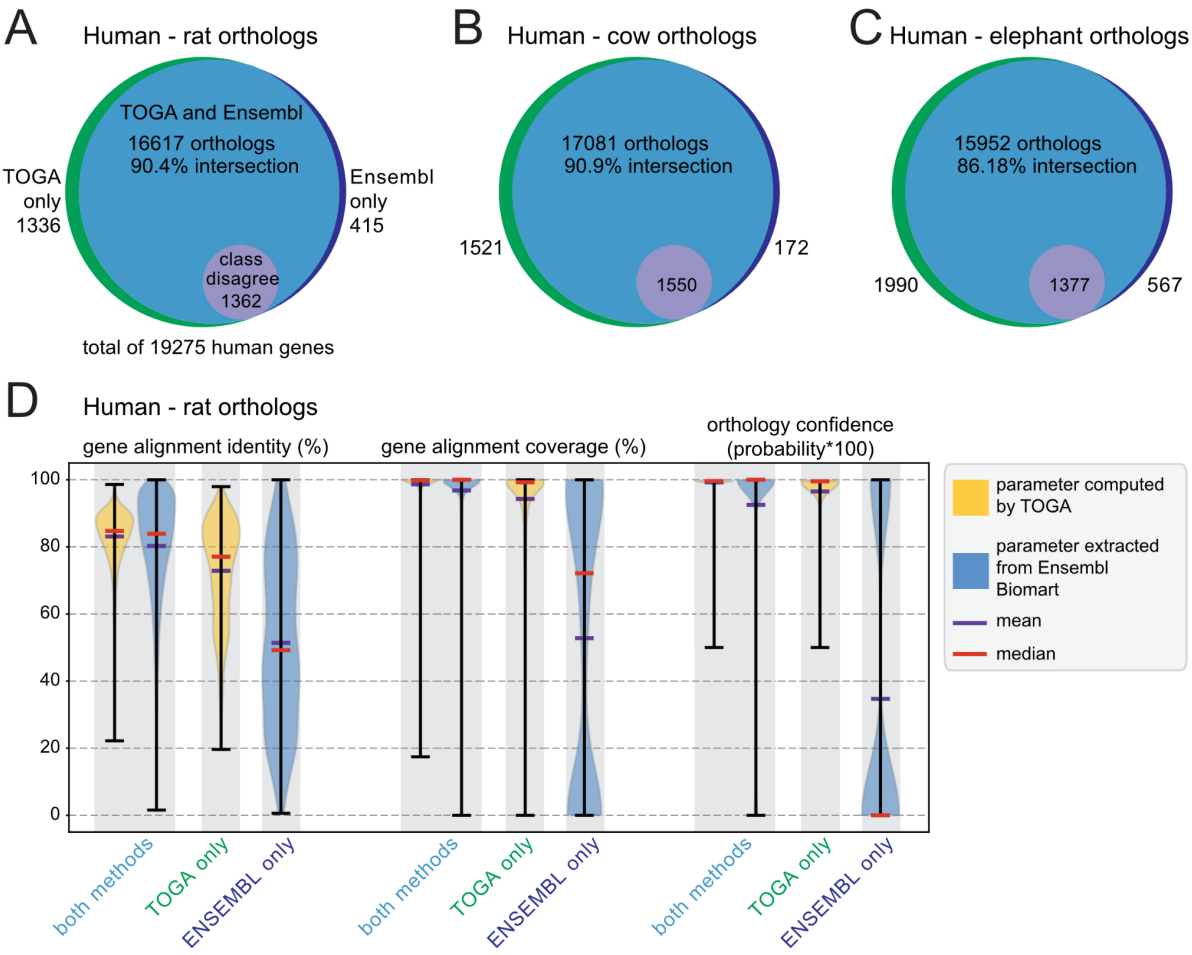

Figure 3.21 Detailed comparison between TOGA and Ensembl orthology predictions

Panel A: Venn diagram showing sets of human orthologs detected in the rat genome by Ensembl and TOGA. Mainly, two methods agree; both methods detected orthologs for 16617 human genes. Panels B and C: similar to A, but for the cow and elephant genomes, respectively. Panel D: statistics for orthologs identified in the rat (i) only by TOGA, (ii) by both methods, and (iii) by Ensembl only. Details in the main text.

We additionally checked whether the sets of orthologs predicted only by TOGA (TOGA-only) and exclusively by Ensembl (Ensembl-only) exhibit statistical differences with the genes on the intersection. In particular, for each group, we evaluated the following features: nucleotide %identity, gene coverage, and level of orthology confidence, that were either computed by TOGA or extracted from Ensembl Biomart. The combination of data sources is necessary because the methods cannot provide these values for orthologs they could not predict. Therefore, for orthologs discovered exclusively by TOGA, we do not have values from Ensembl, and vice versa. However, for the genes annotated by both, we could evaluate the feature distribution from two sources.

Noticeably, TOGA and Ensembl provide different distributions of the nucleotide sequence identity for the same genes. This difference is explained by different techniques to

evaluate the nucleotide identity in both methods: TOGA computes the number of identical bases in the pairwise nucleotide alignment, while Ensembl computes the percentage of query sequences matching the reference sequence.

Nevertheless, as shown in figure 3.21 (panel D), the data suggest that orthologs, predicted by both methods, exhibit high confidence and sequence identity values. TOGA-specific predictions show slightly lower sequence identity, but still, the median is close to 80%. The results imply that TOGA-specific predictions are likely mostly actual orthologs. However, the Ensembl-specific predictions have significantly lower values than orthologs predicted by both methods; for example, the median %identity is close to 50%. This suggests that the Ensembl-specific prediction set is likely to contain many non-orthologous gene annotations.

To check whether these results are consistent for a wider set of species, we performed a similar comparison for 14 additional mammals - the results are illustrated in figure 3.22.

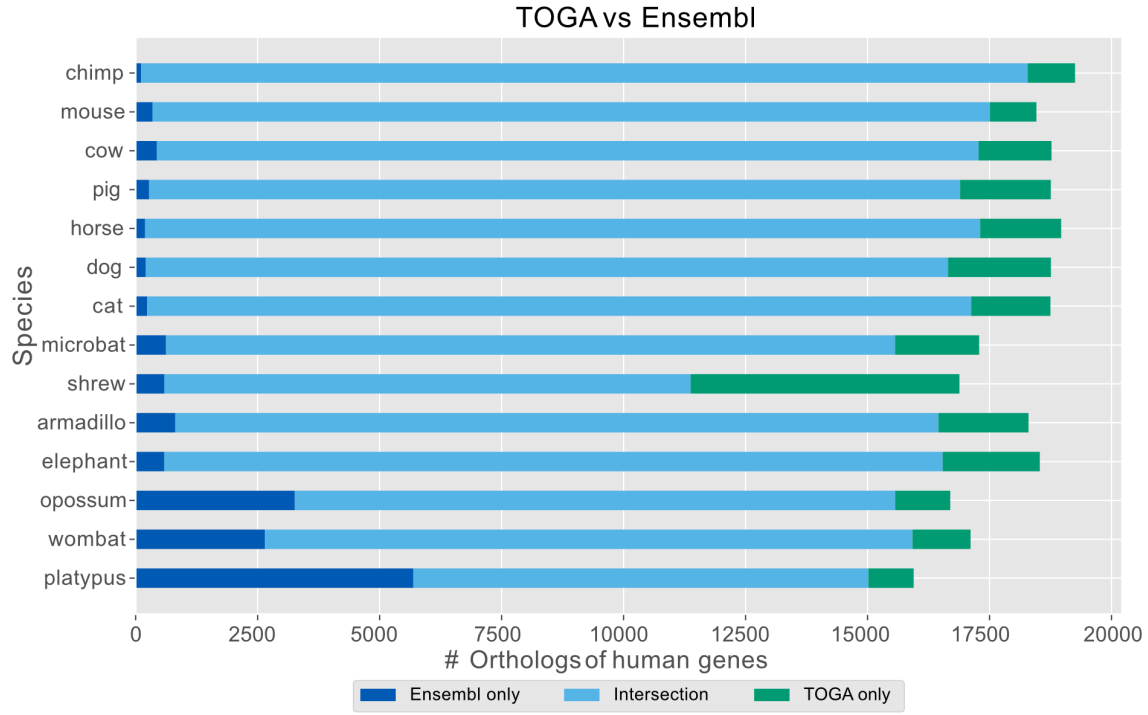

Figure 3.22 Comparison between TOGA and Ensembl on multiple species

The plot shows sets of human orthologs detected (i) only by Ensembl, (ii) by both methods, and (iii) by TOGA only in 14 different mammals. The data suggest that methods mostly agree, but TOGA detects slightly more orthologs for placental mammals. For marsupials and monotremes, Ensembl outperforms TOGA because these species are more distant from the reference (human).

Indeed, for most placental mammals, 90% of human gene orthologs are detected by both methods. Sizes of the gene set predicted by only one method repeat the same pattern: Ensembl continuously detects a few hundreds of orthologs undetected by TOGA, and TOGA detects about a thousand additional orthologs. This implies that two approaches can nicely

complement each other to provide an extensive genome annotation. However, for more distant species such as Marsupials and Monotremes, Ensembl clearly outperforms TOGA. On these evolutionary distances, neutral regions are almost randomized, even in orthologous regions. This aspect and ways to overcome these limitations are discussed in detail in section 5.3.1.

Additionally, we manually inspected the sets of genes predicted by only one of the considered methods. The results revealed insights into the methodological advantages and weaknesses of TOGA compared to the classic gene annotation approach. Subsection 3.4.1.1 provides the results of TOGA-only gene evaluations, while subsection 3.4.1.2 is devoted to an Ensembl-only set of genes.

3.4.1.1 Characterize cases where Ensembl didn't find any orthologs

This subsection gives an overview of TOGA-specific orthology predictions that do not appear in the Ensembl database. This set of genes highlights the advantages of TOGA methodology in inferring orthology where gene trees may lack predictive power.

An example of the Y-box-binding protein 1 (*YBX1*) gene illustrates the weaknesses of tree-based methods in inferring orthology for extremely conserved genes. *YBX1* is an essential gene encoding a DNA- and RNA-binding protein involved in various fundamental cellular processes (Chen et al., 2000; Capowski et al., 2001; Gaudreault et al., 2004; Chattopadhyay et al., 2008). TOGA correctly identifies the orthologous alignment chain and annotates the corresponding gene in the rat genome. Ensembl also annotates the same gene precisely in the same locus, also called *Ybx1*, implying that it is the human *YBX1* gene ortholog (figure 3.23). Surprisingly, the Ensembl states that the human does not have a *YBX1* ortholog in the rat genome.

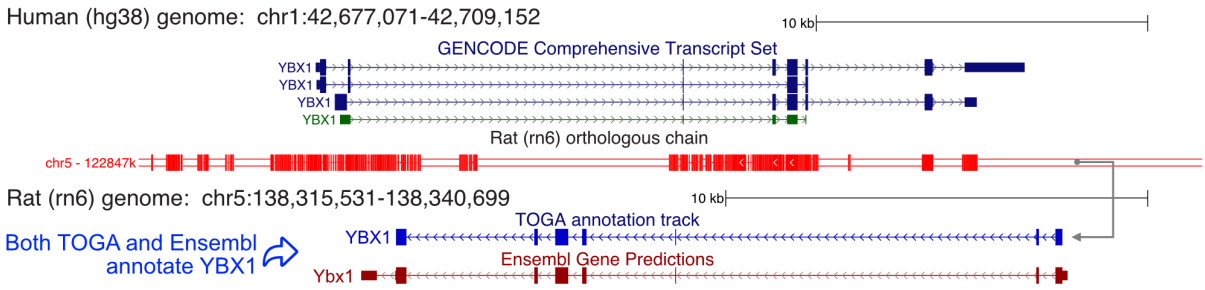

Figure 3.23 Ensembl did not infer ortholog for the YBX1 gene

UCSC genome browser shows the human YBX1 gene and orthologous rat alignment chain. Both Ensembl and TOGA annotate a YBX1 ortholog in the respective locus in the rat genome. However, Ensembl does not list these genes as orthologs.

The fact that Ensembl did not connect the annotated *YBX1* orthologs suggests that the gene tree could not resolve the homology relationships between these genes with a high degree of confidence. A potential reason for that could be that the gene exhibits a very high sequence similarity, so the number of changes between sequences does not provide enough evidence to resolve the homology. Indeed, the protein alignment of the *YBX1* gene exhibits nearly 100% protein sequence identity (figure 3.24). There are only four amino acids that are not identical.

Figure 3.24 YBX1 gene protein sequence alignment

The alignment shows that the YBX1 gene is highly conservative and has only four amino acid changes between the human and rat genomes.

This case illustrates the advantages of not only using coding sequences to infer orthologs. By using additional evidence, such as intronic and intergenic alignments, TOGA can infer orthologs where gene trees may have limitations.

Another representative of this group is the cyclin B2 (*CCNB2*) gene (figure 3.25). This gene is essential for controlling the cell cycle at the mitosis transition; therefore, the loss of this gene is unexpected in highly complex mammalian species. Alignment chains explicitly point to a single orthologous locus in the rat genome, and TOGA classifies it accordingly. Furthermore, TOGA detected several loss-of-function mutations in exon six. Since the loss of a single exon does not provide enough evidence to classify this transcript as "Lost," TOGA assigned the "Uncertain Loss" class to this gene. Subsequently, TOGA included this gene in the final orthology set. However, in this locus, the Ensembl annotation is absent, which is likely a mistake. Presumably, Ensembl did not annotate *CCNB2* because transcriptomic data for the rat genome does not reveal the expression of the entire gene. Instead, aligned RNA sequences correspond to one or another human non-coding RNAs, with a bit of overlap.

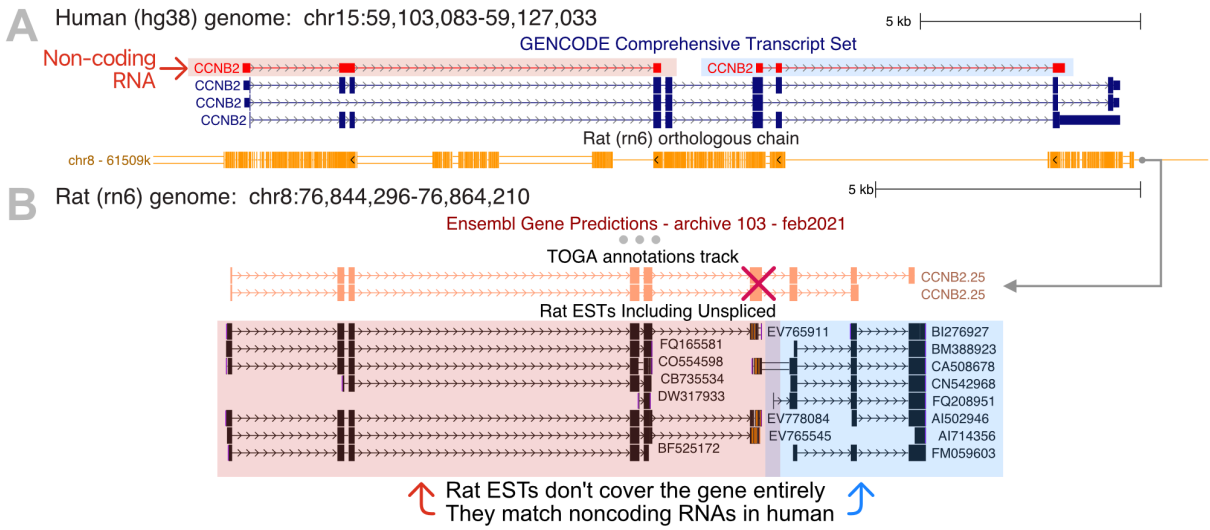

Figure 3.25 Ensembl did not infer ortholog of CCNB2 gene

A: UCSC genome browser shows CCNB2 gene locus in the human genome and rat alignment chain. In addition, two non-coding RNAs are expressed from the same locus in the human. B: respective locus in the rat genome. For some reason, Ensembl does not annotate any transcript in this region. Potentially, this is because there is no EST that cover the gene entirely.

As the next example of TOGA-specific orthology prediction, I examine the Keratin Associated Protein 5-9 (*KRTAP5-9*) encoding gene (figure 3.26). The top-level alignment chain appears to be orthologous, and *TOGA* classifies it accordingly. However, other chains indicate that alignment blocks covering this gene are collinear with alignments up and downstream of the demonstrated locus. This observation reflects that *KRTAP* genes appear in clusters.

TOGA annotates two different genes in the locus corresponding to the top-level alignment chain: *KRTAP5-9* and *KRTAP5-2*. However, Ensembl also annotates a gene in this locus under the *KRTAP5-5* name. It also points out that the *KRTAP5-9* ortholog in the mouse genome does not exist. In contrast, TOGA annotates *KRTAP5-5* ortholog in a different locus. In this case, TOGA mainly relies on flanking alignments and synteny to make the decision. However, it cannot be excluded that its prediction is also incorrect considering the evolutionary process of *KRTAP* genes (Wu et al., 2008; Khan et al., 2014).

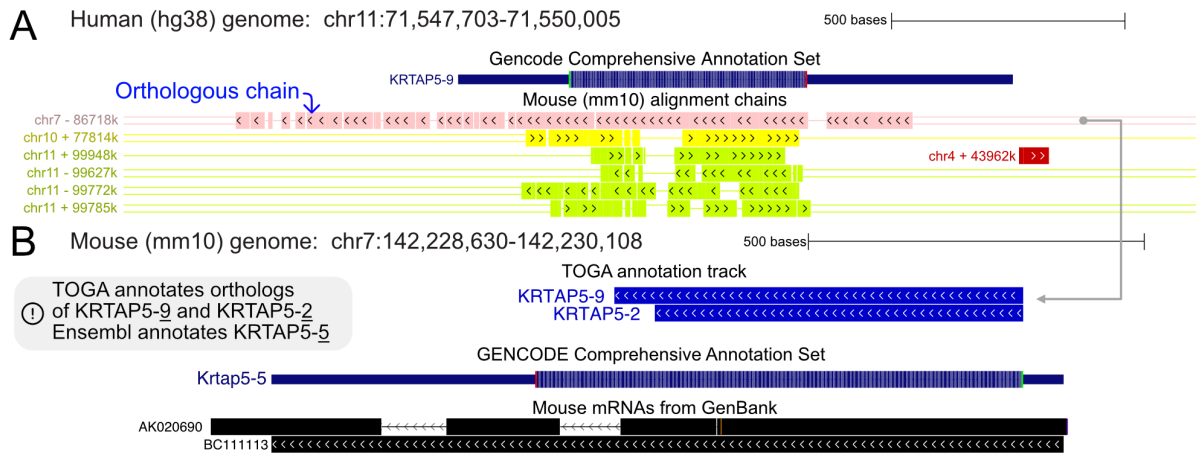

Figure 3.26 Orthology inference within the KRTAP gene family

Panel A: UCSC genome browser screenshot shows human locus containing KRTAP5-9 gene and mouse alignment chains. TOGA classified the top-level chain as orthologous. Panel B: the respective region in the mouse genome. TOGA annotates the respective transcript twice: as KRTAP5-9 and KRTAP5-2 ortholog, implying many-to-one orthology. However, Ensembl annotates the same transcript as KRTAP5-5.

Since the set of TOGA-specific orthology predictions provides crucial importance in understanding the methodology advantages and capabilities, I performed a more profound analysis to identify statistically supported characteristics of these genes. As a result, I extracted characteristic features of orthologs that are more likely to be identified relying on intron divergence than gene-tree-based methods. The details of this deeper analysis are presented and discussed in part 5.1.

3.4.1.2 TOGA classified Ensembl-annotated ortholog as Lost or Missing

Moreover, we manually inspected orthologous that TOGA could not identify despite the presence of corresponding data in the Ensembl dataset. Actually, TOGA classified 366 out of 415 of these genes as missing or lost, which implies that it correctly classified an orthologous chain for these genes. However, for these genes Ensembl and TOGA disagree whether the mapped gene actually encodes an intact protein.

Here, Ensembl has an advantage because it actively uses the transcriptomic data; therefore, it could avoid gene misclassification due to assembly artifacts that mimic loss-of-function mutations. In some of these cases, Ensembl misclassifies a paralog as the ortholog, while the actual ortholog is pseudogenized. However, TOGA could miss a real orthologous locus, as shown in the *CCNQ* gene example (subsection 3.3.1.4). For the remaining 51 genes, TOGA could not find an orthologous chain; therefore, actual TOGA limitations could explain these cases.

TOGA results

The vast majority of orthologs inferred exclusively by Ensembl are classified as Lost or Missing by TOGA. This implies that TOGA finds orthologous loci for nearly all human genes. Since Ensembl uses external evidence to annotate genes, it could distinguish true inactivating mutations from assembly errors. In contrast, if assembly errors are abundant, TOGA could conclude that a conserved gene is lost.

The case of the "Creatine Kinase, Mitochondrial 2" (*CKMT2*) gene clearly illustrates this principle (figure 3.27). In Panel A, a UCSC browser screenshot shows *CKMT2*-containing locus in the human genome. As we can see, it is covered by an orthologous chain, indicating that the orthologous locus is located on chromosome 2 in the rat genome. Panel B shows the respective locus in the rat genome where TOGA and Ensembl both annotate the same *CKMT2* ortholog. However, TOGA classifies this gene as inactivated since it exhibits inactivating mutations in exons 5 and 6 (Panel C). Additionally, exon 9 contains a frameshifting insertion, but since it is located in the last 10% of the CDS, TOGA does not consider this mutation inactivating. Moreover, exon two is missing because the corresponding locus in the rat genome contains an assembly gap.

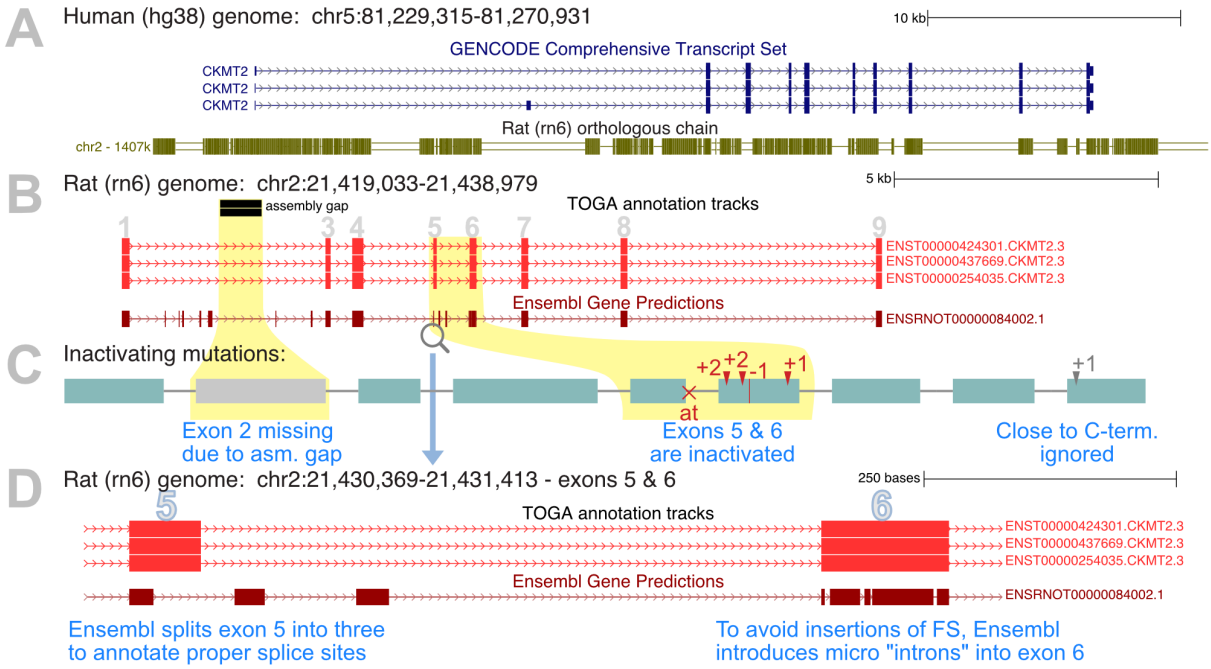

Figure 3.27 Ensembl introduces false micro introns and splits exons to produce "Intact" annotations

Panel A: UCSC genome browser screenshot shows CKMT2 locus in the human genome and rat orthologous chain. Panel B: respective locus in the rat genome, where TOGA annotated CKMT2 orthologs as pseudogenes. It contradicts Ensembl data: it annotates CKMT2 ortholog as intact. Panel C: TOGA detected inactivating mutations in exons 5 and 6, which is why this gene is classified as Lost. Panel D: a closer look at rat locus containing exons 5 and 6. Ensembl introduces artificial introns to compensate for the inactivating mutations and provide a technically intact transcript.

Ensembl annotated this highly-mutated gene as follows: To produce technically correct gene annotations, Ensembl introduces tiny artificial introns to avoid the inclusion of disrupted regions into the ORF. In fact, such introns are a computational workaround because the spliceosome is unable to splicing such short introns in practice. As a result, this Ensembl annotation constitutes a reading frame without inactivating mutations, but it does not reflect the reality. In contrast, TOGA cannot recognize these inactivating mutations as assembly artifacts and concludes that this gene is definitely lost in the query species. However, in the absence of external evidence such as RNA data, this is justified.

3.4.1.3 Actual TOGA limitations

This subsection is focused on cases that actual TOGA approach limitations could explain. Since TOGA heavily relies on intronic and intergenic regions divergence, retroposed single-exon genes could be invisible for the method. Moreover, previously shown cases related to exceptionally diverged chromosome X also belong to this class (subsection 3.1.3.3).

Figure 3.28 provides examples of two genes that actually have an ortholog in the rat genome, but TOGA classified them as wholly deleted. Indeed, orthologs of the *CSNK2A3* (A) and *AC011005.1* (B) genes undergo translocations and therefore are absent in the ancestral locus. Since the flanking alignment and synteny are entirely absent, and genes have no intronic fractions, the actual orthologous chains are indistinguishable from paralogous copies for TOGA.

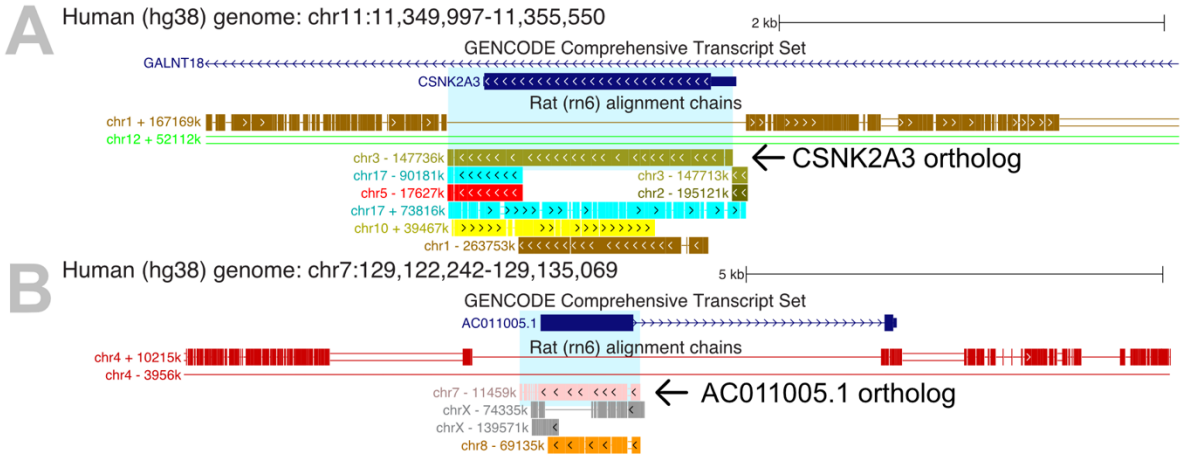

Figure 3.28 Translocated single-exon co-orthologs

A: UCSC browser screenshot shows CSNK2A3 gene in human and rat alignment chains. The gene underwent translocation in the rat genome. However, the orthologous chain lacks any flanking alignment and therefore was misclassified by TOGA as paralogous. Panel B: similar to panel A, but for AC011005.1 gene - orthologous chain can be indistinguishable from paralogous in case of single-exon gene translocation.

3.4.2 BUSCO completeness scores

To provide additional evidence for TOGA annotation sufficiency, we measured the completeness of Benchmarking Universal Single-Copy Orthologs (BUSCO) gene set annotations. The BUSCO set contains universal single-copy genes that are expected to be conserved in the given clade, in our particular case for the Vertebrata. The proportion of BUSCO genes that a given annotation method is able to recognize is a reliable metric to assess gene prediction quality. However, the BUSCO set is also applied to benchmark genome assembly quality. For instance, the more considerable number of missing BUSCO genes in a given genome is a hallmark of poor assembly quality. Therefore, it is necessary to consider that even a flawless annotation method would demonstrate lower prediction completeness on such genomes. In such genomes, some loci harboring particular genes might be absent.

To perform the quality evaluation, I downloaded the BUSCO genes set for Vertebrata clade (odb9), containing 3306 highly conserved genes that are expected to be present in each vertebrate species. Then, I selected several species representing different vertebrate clades and annotated them with TOGA using the human genome as the reference. To produce the results, I quantified the number of genes representing each TOGA class in each analyzed genome, such as Intact, Uncertain, Lost, or Missing. Since TOGA yields a large number of gene classes, they are grouped for results interpretability as follows:

1. The first group unites the following annotation classes: "Intact," "Partial Intact," and "Uncertain Loss". For genes classified as such, TOGA establishes the orthologous connections between reference and the query.
2. The second group comprises the classes of "Missing," "Lost," and "Partial missing". Association with one of these categories implies that TOGA identified an orthologous chain but failed to identify an intact transcript in the query genome.
3. The third group comprises a single annotation class, "Paralogous projection". For these genes, TOGA could not identify an orthologous chain
4. And the last group, "no chain", implies the absence of any alignment chains intersecting a given gene.

On the next page, figure 3.29 demonstrates the results of BUSCO completeness evaluation.

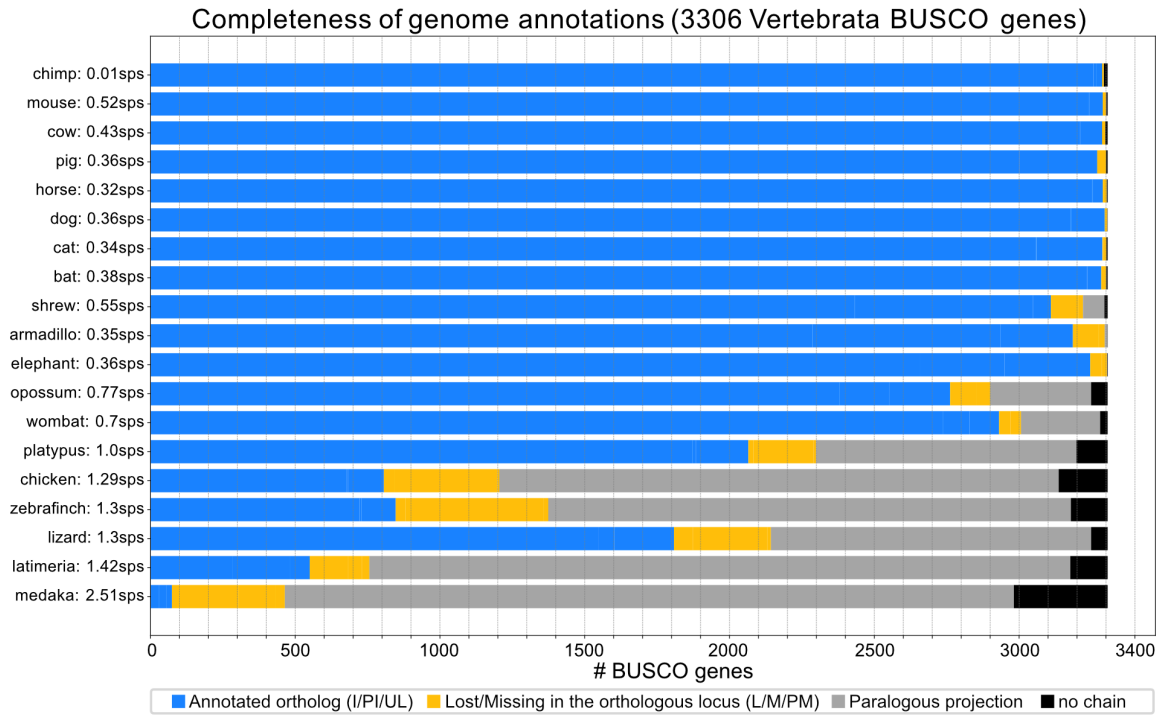

Figure 3.29 BUSCO completeness of TOGA annotations

The plot shows the completeness of BUSCO gene annotations for 19 different vertebrate species. Sps means "substitutions per neutral site," which is the measure of molecular distance between species (in this case - to the human). The higher value means that the lower fraction of the neutrally evolving sequence will align. According to the plot, TOGA provides nearly ideal BUSCO completeness with molecular distances below 0.7 sps, which corresponds to the placental clade. If the distance exceeds 1, then TOGA cannot correctly distinguish orthologs from paralogs for many genes because of the lack of neutral sequence alignment.

According to results, TOGA could identify nearly all BUSCO genes in the Boreoeutheria species, close relatives to the reference. However, it performed slightly worse on the shrew lineage, which is highly diverged and represented by a low-quality genome assembly. Armadillo and Elephant represent more phylogenetically distant Xenarthra and Afrotheria clades and demonstrate slightly lower but still satisfactory annotation quality.

Notably, the prediction quality decreases with the molecular distance between species (number of substitutions per neutral site, sps). For example, the annotation quality of Marsupialia genomes (~0.7 sps) is significantly lower than for Placental species (<0.55 sps). In Marsupialia annotations, the number of paralogous projections became noticeable. In such cases, TOGA was unable to detect an orthologous chain due to elevated neutral sequence divergence. The annotation of the platypus genome (1 sps) demonstrates even lower completeness because, at this distance, the neutrally evolving regions are expected to be entirely randomized.

For molecular distances above 1 sps, orthologs and paralogs are expected to be largely indistinguishable for TOGA since intronic and intergenic regions are highly diverged, which is illustrated by very poor annotation completeness in species outside the mammalian clade. Notwithstanding, it still annotates most of these genes without reaching any conclusions regarding orthology relationships (paralogous projections). The influence of molecular distance on TOGA annotation quality is further discussed in section 5.3.1 devoted to TOGA methodology limitations analysis. Also, the section provides ideas of how to increase TOGA annotation completeness for molecular distances closer to 1sps.

As demonstrated in this section, the TOGA approach provides a high annotation completeness in comparison to the gold standard dataset. It outperforms traditional methods on relatively close evolutionary distances, manifesting the advantages of the proposed concept. The practical applications of this concept are illustrated in the following part.

3.5 Genome alignment chaining procedure affects TOGA prediction quality

To infer orthology, TOGA utilizes the alignment chains, and ergo the overall pipeline performance rigorously depends on the genome alignment quality. Genome alignment inaccuracies typically lead to wrong mapping of the coding sequence to the query genome, and lack of alignment sensitivity results in orthology underprediction. In this part, I discuss potential issues that chain obstructions could cause. Conceptually, chains are a form of genome alignment representation, which suggests they could be produced from various sources. This section briefly compares different genome aligners to illustrate why alignment sensitivity is crucial for TOGA classification quality. In particular, I compared three different sources of genome alignment chains:

1. The LASTZ-based procedure that we applied to produce previously published 120-way genome alignment (Hecker and Hiller, 2020) which includes post-processing procedures to increase alignment sensitivity (Suarez et al., 2017; Osipova et al., 2019).
2. The LASTZ-based method used for standard UCSC-browser annotation track (Kent et al., 2003)
3. CACTUS 1000 multiple genome alignments (Armstrong et al., 2020)

Basically, both sets 1 and 2 are produced using the same LASTZ aligner. However, to generate our chains, we used an optimized parameter set and applied post-processing procedures to increase method sensitivity in poorly aligned regions. Figure 3.30 shows a UCSC browser screenshot providing a side-by-side comparison of three genome alignments.

Notably, chains produced under the first methodology provide greater sensitivity in neutrally evolving regions. Statistically, this implies that more sensitive chains would increase the TOGA outcome because the chain classification step of the pipeline heavily relies on neutrally evolving sequence alignment.

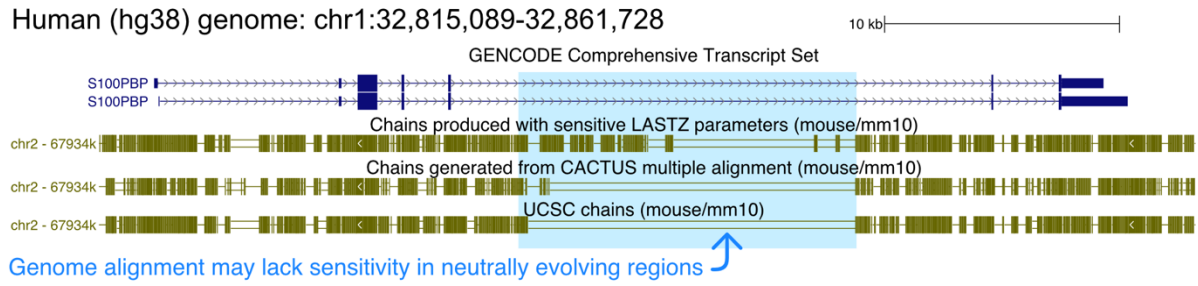

Figure 3.30 Comparison between alignment chains

UCSC browser screenshot shows a side-by-side comparison of mouse alignment chains produced using different approaches: (i) sensitive LASTZ approach (Hillerlab), (ii) chains extracted from CACTUS alignment, and (iii) default LASTZ chains provided by UCSC. Chains align to the S100PBP gene locus in the human genome. A highly diverged intronic region is highlighted with grey.

To evaluate the dependency of TOGA pipeline outcome on the alignment chains sensitivity, I annotated the mouse (mm10), cow (bosTau8), and dog (canFam3) genomes using (i) our and (ii) UCSC chains with the human genome as the reference. Then, I quantified orthologous genes detected by TOGA using different chains. The results of this analysis are presented in table 1.

Species	Hillerlab chains	UCSC chains	Difference
Mouse (mm10)	18117	17926	191
Cow (bosTau8)	18334	18263	71
Dog (canFam3)	18193	18113	80

Table 1 Number of orthologs predicted by TOGA using Hillerlab and UCSC chains

These results suggest that TOGA yields slightly more orthologous connections if the genome alignment sensitivity is increased. Below, I consider some examples of orthologs missing due to a lack of alignment sensitivity. The first example demonstrates how a lack of chain block connectivity affects gene classification (figure 3.31). Our chain post-processing procedure could detect the collinearity of the separate aligning blocks and connect them, resulting in a complete chain. Further, TOGA identifies the corresponding region in the query genome as orthologous for the *TMLHE* gene. Besides, UCSC alignment recognizes almost the same blocks of sequence similarity between the human and mouse genomes; however, it

did not connect these blocks together. As a result, TOGA was unable to infer orthology for this gene using the UCSC chains set.

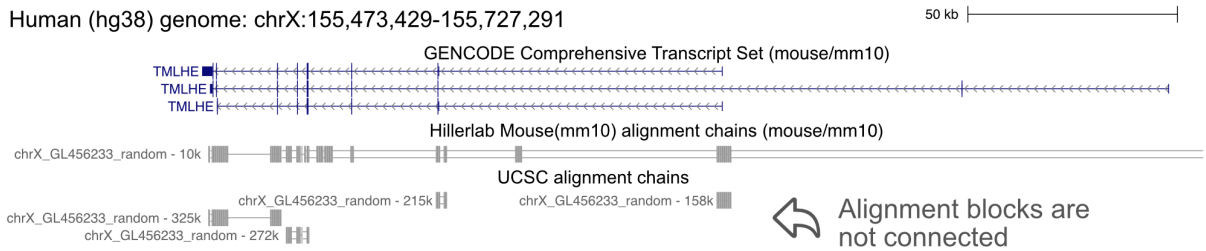

Figure 3.31 Not connected alignment blocks in UCSC chain track

UCSC genome browser screenshot shows TMLHE gene locus in the human genome and two alignment chain tracks: (i) Hillerlab and (ii) UCSC. Both tracks exhibit aligning blocks in almost the exact locations. However, the UCSC chaining procedure did not detect collinearity of these local alignments and therefore created many small chains. In contrast, the Hillerlab procedure chained these blocks together.

The following example considers the human Defensin Beta 4B (*DEFB4B*) gene involved in antimicrobial activity, which exhibits an extreme sequence divergence with its ortholog in the mouse genome (figure 3.32). Because of a lack of sensitivity, the UCSC genome alignment procedure could not reveal any block of sequence similarity between human and mouse genomes resulting in entirely absent chains. The nucleotide alignment provided by our chains (on the illustration) clearly illustrates that the coding sequence of this gene is indeed highly diverged (sequence identity ~50%). For this reason, TOGA could not recognize any *DEFB4B* ortholog in the mouse genome based on UCSC alignment. However, utilizing our chains, TOGA established the orthologous connection between human and mouse DEFB4B genes.

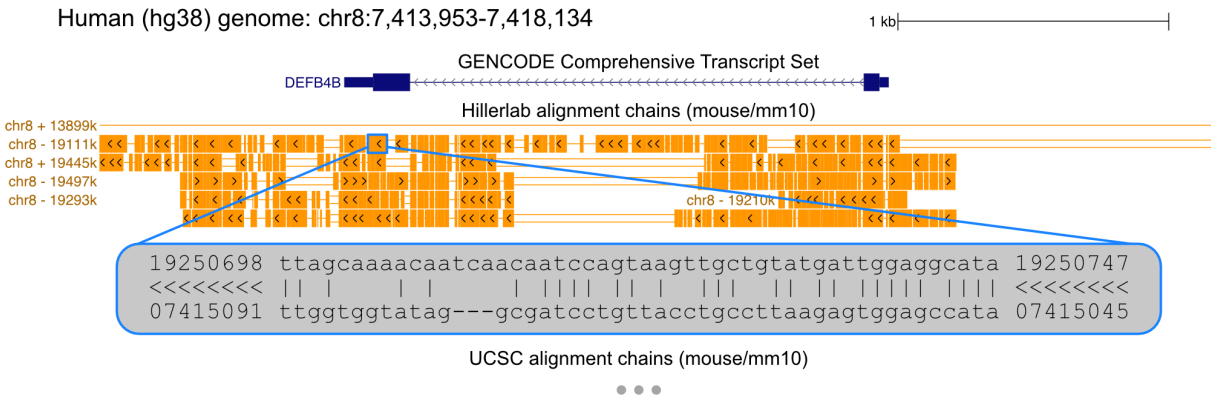

Figure 3.32 Absent UCSC alignment to DEFB4B gene

UCSC genome browser screenshot shows DEFB4B locus in the human genome with two alignment chain tracks: (i) Hillerlab and (ii) UCSC. The region has highly diverged (magnified local alignment - sequence identity is about 50%), and the UCSC procedure could not identify any sequence similarity here.

The last selected example illustrates the *WFDC12* gene alignment between the human and cow genomes (figure 3.33). Actually, this gene underwent tandem triplication in the cow genome, and two of these copies got inactivated. Our chains successfully captured all three copies of this gene, whereas the UCSC procedure could align this gene to only one of these copies. Consequently, using sensitive chains, TOGA could detect an intact copy and establish the one-to-one orthology connection between human and cow *WFDC12* genes. In contrast, using UCSC chains, TOGA examined only one of these copies, which apparently is inactivated, and consequently reported that this human gene has no ortholog in the cow genome.

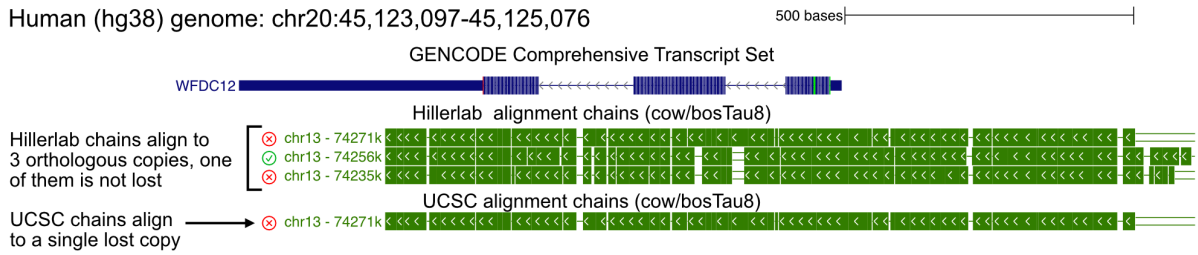

Figure 3.33 UCSC genome alignment missed two WFDC12 copies

UCSC genome browser screenshot shows that Hillerlab chains could align three copies of the WFDC12 gene in the cow genome (2 copies are inactivated), whereas UCSC chains aligned only a single copy. The copy detected by the UCSC genome alignment procedure is inactivated.

3.6 Practical TOGA applications

In this part, I give an overview of three independent projects that involved the TOGA genome annotation pipeline. Two of them: (i) annotating six reference-quality bat genomes (section 3.5.1) and (ii) analysis of the *BAAT* gene evolutionary history (section 3.5.2), are already published in the scientific literature. The third one, producing comprehensive gene annotations for 450 mammals (section 3.5.3), provides plenty of data for subsequent comparative studies and demonstrates the great method scalability.

3.6.1 Producing highly complete annotations for six reference-quality bat genomes

An earlier TOGA version was used as one key type of evidence to annotate genes in 6 reference-quality bat genome assemblies generated by the Bat1K project (Jebb et al., 2020). To create these annotations, TOGA results were integrated with other evidences such as *ab initio* approaches and transcriptome data. Using EvidenceModeller (v.1.1.1) (Haas et al., 2008) to integrate these evidences, 19,122–21,303 protein-coding genes were annotated in the six

bat genome assemblies. This resulting data reached 99.3–99.7% of the BUSCO gene set annotation completeness. Notably, the completeness of the produced annotation is higher than for available annotations of dog, cat, horse, cow, and pig genomes. Genomes of only two species surpassed our annotation: human and mouse, which have received extensive manual curation over the last decades.

In addition to annotation, I used TOGA to generate a comprehensive set of 12931 orthologous genes across the 6 bats and 42 other mammals. Using this orthology set, we performed multiple analyses for gene selection, losses, and gains that uncovered the genetic origins of fascinating bat adaptations (details in the article). This application demonstrated that the TOGA approach has the excelling potential to become a widely recognized tool for genomes annotation. We had an opportunity to examine TOGA in practice during this project and compare it to various standard annotation methods.

3.6.2 Analysis of the evolutionary history of the BAAT gene

As another case study, published in Genome Biology and Evolution (Kirilenko et al., 2019), I applied TOGA in a comparative study to explore the genetic origin of bile acid conjugation variability across 120 mammalian species. In vertebrates, bile acids are conjugated with amino acids to fulfill their biological functions. However, conjugated amino acids are variable among mammals: some species conjugate bile acids with both glycine and taurine, whereas others conjugate only taurine. In particular, our study was focused on the bile acid coenzyme A: amino acid N-acyltransferase (*BAAT*) - the enzyme that catalyzes bile acid conjugation in humans.

By applying TOGA to 120 mammalian genomes, we uncovered the complex evolutionary history of the *BAAT*-encoding gene, which included multiple gene loss and duplication events. Using the codon alignments, we observed multiple changes in the active center of the enzyme between Cysteine and Serine, which likely contribute to the observed variability of the bile acid conjugation pattern. This assumption was based on mutagenesis experiments showing that replacing Cysteine for Serine in the active center greatly diminishes the glycine-conjugating ability in the human enzyme.

Surprisingly, we found that this residue provides little power outside primates' clade in predicting the experimentally measured amino acids that are conjugated with bile acids. These results suggested that the mechanism of *BAAT's* enzymatic function is incompletely understood, despite relying on a classic catalytic triad. More generally, our evolutionary analysis indicates that results of mutagenesis experiments may not easily be extrapolated to other species.

3.7 TOGA annotation of 500 mammalian genomes

To demonstrate that TOGA scales to many genomes, we applied it to generate comprehensive annotations for 500 genomes, representing 450 mammalian species, using both the human and mouse genomes as the references (in total 1000 TOGA runs), creating the largest comparative dataset so far. The annotated genomes represent about ~10% of all known mammalian species and cover all major clades. Consequently, the generated dataset provides an unprecedented level of detail, empowering us to perform very comprehensive comparative studies. The histogram illustrated in figure 3.34 shows the number of annotated human orthologs in these genomes. In half of the analyzed species, TOGA could identify orthologs for at least 17862 human genes, or 17408 orthologs on average, which indicates that the resulting annotations are highly complete.

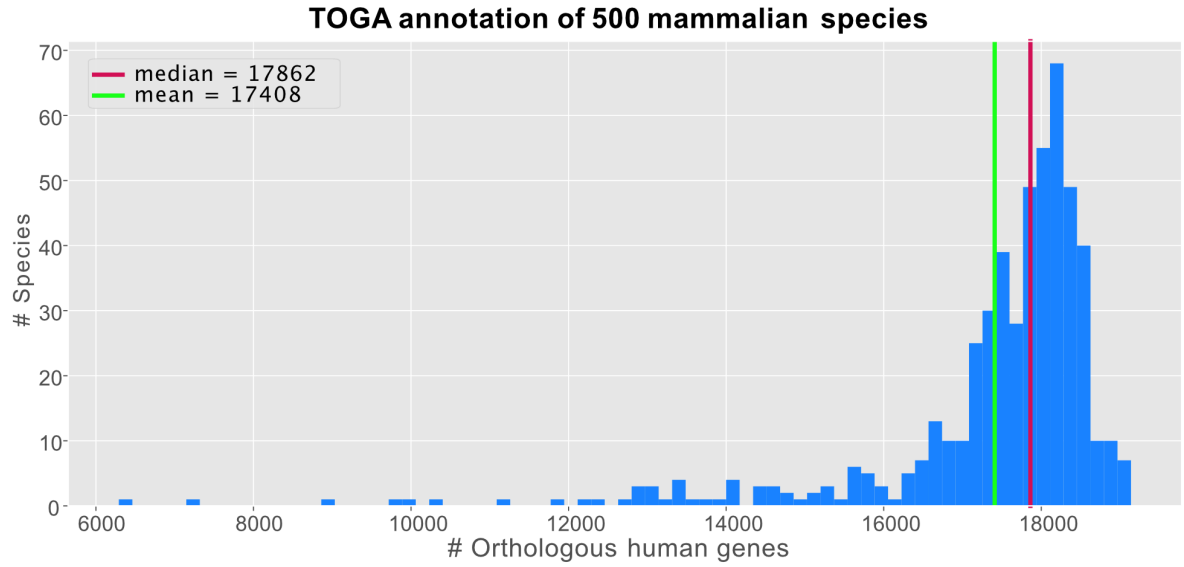

Figure 3.34 Number of human orthologs predicted in 500 assemblies of ~450 mammalian species

The histogram shows that, on average, TOGA was able to identify orthologs of 17400 human genes in 500 mammalian genomes representing 450 species.

However, the dataset includes outliers with low numbers of annotated human orthologs. The reason is that the assortment of 500 genomes comprises assemblies of variable quality. It incorporates highly fragmented draft assemblies (such as the Kogia, see section 2.2.6) as well as reference-quality mammalian genomes (Jebb et al., 2020). Since low quality assemblies do not provide all gene loci, the number of annotated orthologs is expected to be lower. Indeed, TOGA could detect only 9968 orthologs of human genes in the Alpine ibex genome, most likely because of extreme assembly incompleteness (contig/scaffold N50 of 380,983/61,905,114).

The UCSC genome browser screenshot shown in figure 3.35 illustrates the locus in the human genome comprising essential genes encoding the DNA Topoisomerase I (*TOP1*) and Chromodomain Helicase DNA Binding Protein 6 (*CHD6*). Also, it shows the alignment chains for Alpine ibex and domestic goat genomes, which are close relatives. TOP1 and CHD6 encode essential proteins necessary for DNA replication and transcription. Therefore, it is expected that any eukaryotic species including goat and ibex possess them. However, the alignment chains indicate that the corresponding region is not assembled in the ibex genome in contrast to the goat. Instead, the top-level alignment chain corresponds to the deletion of the presumably orthologous locus. This observation suggests that the Alpine Ibex genome is highly incomplete, explaining the low number of discovered orthologs.

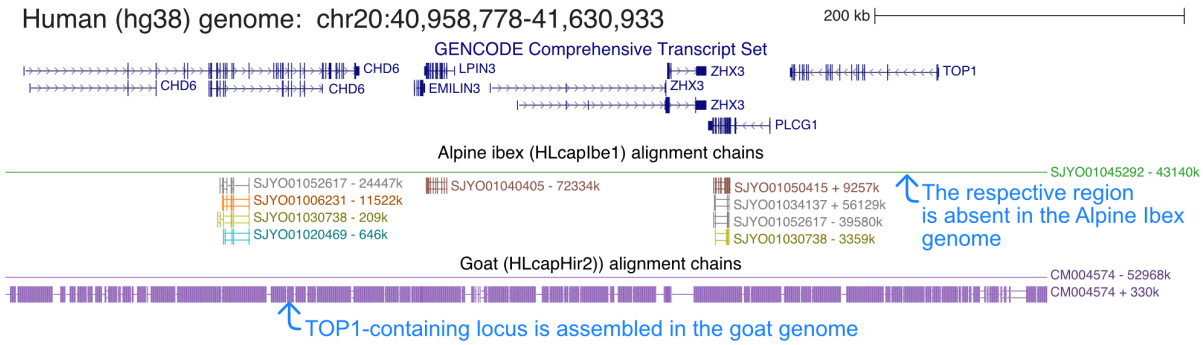

Figure 3.35 Incomplete genome assembly on the TOP1-containing locus

UCSC genome browser screenshot shows TOP1-containing locus in the human genome and alignment chain tracks to goat and Alpine ibex. The goat chain track exhibits alignment to the orthologous ancestral locus. However, the Alpine ibex chain track shows that the respective locus is deleted in the ibex genome, implying ibex genome assembly incompleteness since TOP1 is essential for DNA replication.

3.8 UCSC genome browser visualization for TOGA annotations

To make all information used by TOGA accessible and transparent to users, we adopted code from the UCSC genome browser to create a TOGA annotation track type. For each transcript, our modification provides the following data:

1. The reference transcript identifier together with a link to Ensembl (or another user-defined gene resource) and reference genome coordinates
2. The orthology score of the chain used for this projection, together with the features used for machine learning classification (section 2.2.2)
3. The transcript classification (intact, partial intact, etc.) and the features that underlie this classification (section 2.4.5)
4. A figure that visualizes all exons, including their class (present, missing, deleted), and plots all identified inactivating mutations (section 2.4.2)
5. A list of all detected inactivating mutations (section 2.4.3)

6. The pairwise protein sequence alignment between reference transcript and TOGA prediction
7. Alignments of individual exons together with coordinates, expected regions, %nucleotide identity, and %BLOSUM values (section 2.4.2).

This implementation comprises a handler function in UCSC's "hgc.c" that determines whether the user clicks on a TOGA annotation track. If that is the case, our extension fetches all data from three SQL tables containing the information described above and uses it to produce an HTML page (figure 3.36).

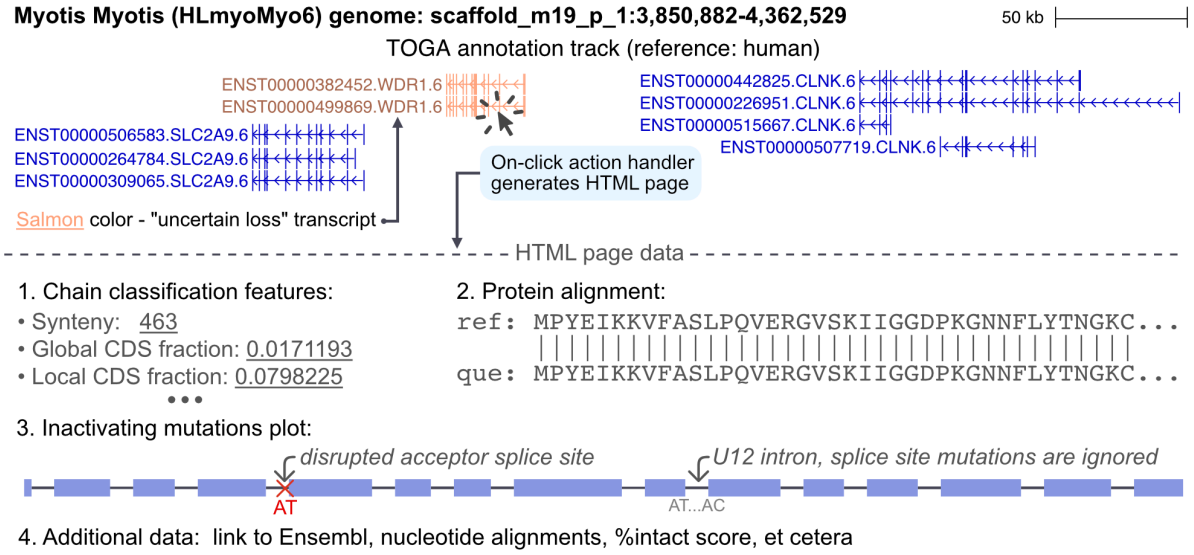

Figure 3.36 TOGA annotation track in the UCSC genome browser

The figure shows the TOGA annotation track in the UCSC genome browser. We modified the on-click action handler such that clicking on TOGA annotation generates an HTML page showing data related to the selected transcript (details in the main text).

4. TOGA extensions

High-quality orthology inferences from TOGA provide the solid basis to generate reliable alignments of orthologous exons and genes, which is invaluable for phylogenetic analysis (part 4.1) and selection tests (part 4.2).

4.1 High-quality alignments of orthologous exons for phylogeny inference

The inference of phylogenetic relationships amongst or within groups of species is one of the central goals of evolutionary biology. Phylogeny provides the evolutionary history of a group of species crucial for virtually all further evolutionary studies (Soltis and Soltis, 2003).

The variety of methods for phylogenetic inference include parsimony, maximum likelihood, or Bayesian inference. In general, to build a phylogenetic tree for a set of interest species, all these methods require multiple sequence alignments. Researchers apply sequences of various types to infer phylogeny, such as ultraconserved elements (Faircloth et al., 2012), mitochondrial DNA (Braun and Kimball, 2002), and protein-coding genes (Collins et al., 2005). These sequences vary in terms of the potential alignment length, level of noise, selective pressure on the sequence, etc.

Phylogeny projects are often negatively influenced by including non-homologous sequences due alignment artifacts and errors in orthology inference. These non-homologous inclusions may lead to wrong phylogenies. Accurate alignments are therefore critical

We propose an alternative source of data for phylogeny studies by extracting it from high-confidence TOGA annotations. To exclude non-homologous sequences, we employed the fact that TOGA considers exonic sequences and the entire gene locus in the query genome context. Thus, we can extract a set of very reliable orthologous exons. To overcome alignment artifacts, we require that such high-quality exons align well at both nucleotide and protein levels and have no insertions or deletions, which lets us eliminate the chance of alignment errors. Below I provide the details of our filtering algorithm.

4.1.1 Workflow to extract high-quality exons from TOGA annotations

To minimize the chance of including non-homologous sequences, we implemented a sequence of filters. Exons classified as “high-quality” should satisfy all of the requirements listed below. First, we extract high-confidence one-to-one orthologs from TOGA output. To do so, we select orthologs that were projected through a chain with XGBoost orthology probability above 0.95. Besides, we require that no other chain with a score higher >0.2 that covers the reference gene exists. These requirements ensure that the considered reference gene has only one orthologous copy in the query genome, excluding partial co-orthologs.

Second, we examine each candidate exon requiring high sequence similarity. In particular, we require that exons align well with %Blosum ≥ 75% and %identity ≥ 65%. With this requirement, we make sure that exon sequences are homologous, excluding highly diverged sequences. Moreover, we require that the exon flanks align, which ensures that the exon boundaries are reliably detected. To this end, we focused on the 100 bp up-and downstream exon flanks and required that at least 90 of the 100 reference bp intersect with aligning blocks from the chain and that insertions in the query do not sum to more than ten bp.

Third, we require that high-quality exons have no insertions and deletions. To achieve this we require that a single aligning block covers the whole exon and pairwise CESAR alignment reveals consensus splice sites but no indels. By applying these measures, we drastically reduce the chance of alignment ambiguity, which typically arises from an inability to precisely locate the positions of insertions or deletions.

Fourth, the transcript containing potential high-quality exons must be classified as "intact" or "partial intact." Also, the considered exon should not exhibit any inactivating mutations, including compensated or masked ones, making it likely that the exon evolves under purifying selection.

Fifth, we require that the considered exon is not duplicated in one-to-one orthologous query genome locus. This requirement avoids the possibility of including the wrong exon copy in the final alignment. To achieve this, we test whether the transcript locus comprises any additional exon copy using LASTZ. In case LASTZ detects more than one exon copy, we exclude this exon from the high-quality exon set.

As a result, we obtain a set of high confidence exon alignments that are expected to be virtually free of inclusions of non-orthologous sequence. Since the high-quality exons do not have any insertions and deletions, no additional alignment steps are required. To produce multiple alignments, it is enough to stack resulting exons on top of each other. The power of this approach is demonstrated in practice in the next section.

4.1.2 Inferring bat phylogeny

In order to check whether the high-quality exons provide sufficient material for phylogeny inference, I have built a phylogeny tree for 48 mammals applying high-quality exons exclusively. Then, I compared the resulting tree with a previously published one for the same set of species (Jebb et al., 2020). The motivation of the original phylogenetic study was that some aspects of bat evolution were still unknown (Foley et al., 2016; Doronina et al., 2017; Springer and Gatesy, 2019). This question was revisited considering the high completeness of newly sequenced bat genome annotations.

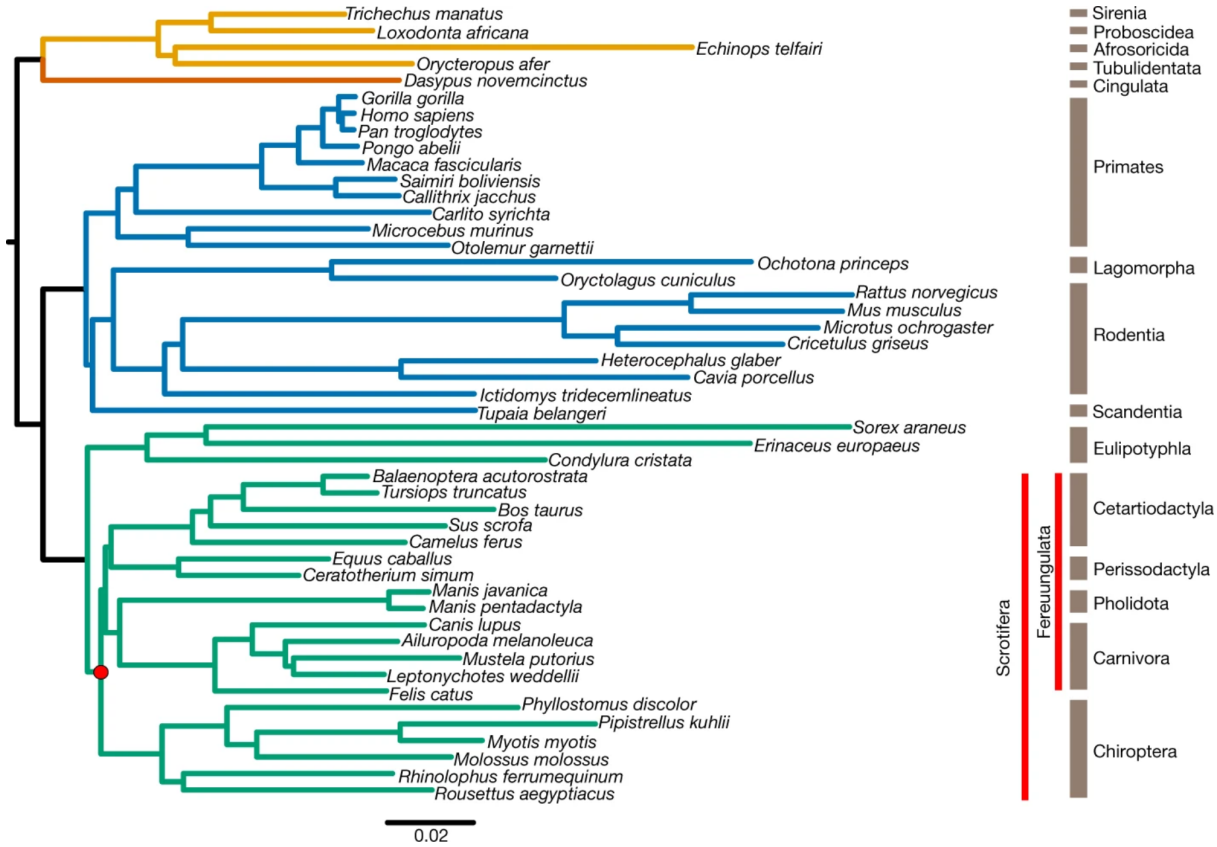

Figure 4.1 Bat's phylogeny tree on the mammalian background (Jebb. et al, 2020)

To produce the tree, we used a supermatrix of 12931 concatenated genes and the maximum likelihood tree reconstruction method. All nodes received 100% bootstrap support.

For the original study, we extracted sequences of 12.931 orthologous genes provided by an earlier TOGA version and combined them with 10.857 orthologous conserved non-coding elements collected from 48 mammalian genomes. This resulted in a total of 21.4Mp-long alignments of TOGA sequences and 5.2Mp of non-coding element sequences to reconstruct the phylogeny. However, we identified some homologous errors in the resulting alignment. Most of these cases involve a short non-homologous first or last coding exon, affecting relatively few bases in the overall gene alignment. Some of these cases were caused

by the incompleteness of utilized mammalian genome assemblies, where an assembly gap covers the real exon, and CESAR2.0 detects a sufficiently similar but non-homologous exon candidate. The original tree was produced using IQtree method (Nguyen et al., 2015) with 1000 bootstrap replicates. The analysis of concatenated protein-coding genes identified the origin of bats with 100% bootstrap support across the entire tree. The resulting tree published in the original study is demonstrated in figure 4.1.

To infer the tree based on TOGA results, I extracted a set of high-quality exons from the same set of species, which resulted in a 2MB long codon alignment, which is substantially shorter than the alignment applied in the original study. Analysis of this alignment did not reveal any ambiguous or misaligned region, confirming that the extracted sequence is indeed highly conserved. To generate the tree, we applied the same IQtree-based procedure. As a result, we obtained a tree of identical topology, which confirms that the proposed method could find an application in phylogeny studies.

4.1.3 UHQ exons summary

The approach proposed in this part has numerous advantages to generate sequences for phylogeny inference. First, produced sequences are relatively short because only a minor fraction of the annotated sequences could satisfy the strict criteria. The shortness of sequences does not compromise the quality because, by definition, they have a shallow noise level. This feature allows performing phylogeny analysis faster, without loss of quality. Another advantage is that the alignment step is no longer required. The extracted sequences, by definition, are already aligned. Consequently, we avoid any potential alignment mistakes and ambiguity. Since the proposed approach operates with exon units but not the entire genes, the results are primarily free of recombination events. Therefore, it could be utilized as input for coalescent methods that assume that no recombination events happened.

However, this method has certain limitations. It is very problematic to apply this method for hundreds of species because high-quality exons are rare. Therefore, the method's scalability is limited: with the increasing number of species, the number of exons classified as “high-quality” in the majority of them quickly shrinks. Another issue is that this method strictly depends on the assembly and alignment quality. If the assembly quality is poor and the genome alignment was not sensitive enough, then the number of extracted UHQ exons would be impractical. The implemented filters could be too strict for highly diverged species because exon flanking regions constitute neutrally evolving regions. The chance that the alignment of these regions will satisfy our criteria is relatively low.

In general, the method proves its potential in producing data for phylogeny studies involving closely related species represented by high-quality genome assemblies. In this particular niche, it could complement conventional procedures of data preparation.

4.2 Extract codon alignments from TOGA for selection analysis

To recognize particular natural selection patterns in coding sequences, methods such as aBSREL (Smith et al., 2015) and PAML (Yang, 2007) evaluate codon substitution rates. To detect molecular evolution patterns, these methods compute rates of synonymous (do not alter amino acid sequence) and nonsynonymous mutations (affect the amino acid sequence). To evaluate these rates, it is essential to have accurate codon alignments of orthologous sequences that are all aligned in the same reading frame. Any alignment artifact or genome assembly error affects mutation rates evaluation, leading to potentially incorrect conclusions (Di Franco et al., 2019).

Producing codon alignments introduces specific challenges since each aligned sequence must be translated in the same reading frame. For instance, any insertion of a non-homologous sequence or frameshifting indel could disrupt the entire alignment. Figure 4.2 provides a specific example of a codon alignment corrupted by a frameshifting insertion in the Cape golden mole sequence in the gene *CRYBA2*. The aligner mistakenly associated the disrupted codon with the reading frames of other species. These inaccuracies could further mimic nonsynonymous mutations resulting in a nonhistorical selection signal. A method such as aBSREL (Smith et al., 2015) may interpret them as a signature of positive selection that occurred in this species.

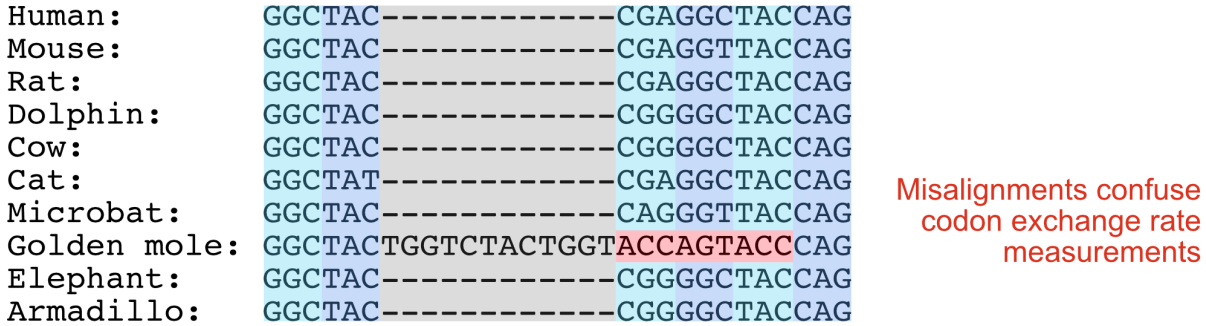

Figure 4.2 Codon alignment inaccuracies

The alignment exemplifies a misalignment induced by non-homologous insertion (in the Cape golden mole sequence). Such misalignments elevate the nonsynonymous substitution rates resulting in false selection signals.

To extend TOGA applicability, we developed an extension that extracts codon alignment qualified for selection analysis directly from TOGA output files. This extension relies on the TOGA feature that extracts pairwise codon alignments masking all frameshifting mutations. Subsequently, these masked codon sequences are aligned with MACSE2.0 (Ranwez et al., 2018) to produce codon alignments of multiple species. By generating gene alignments in an “exon-by-exon” fashion, mis-aligning some exon parts to other non-orthologous exons is avoided. As a result, the generated codon alignments are highly-reliable. The next section of this work provides implementation details.

4.2.1 Extracting masked pairwise codon alignments

Accompanying the genome annotations, TOGA also produces pairwise codon alignments that are corrected for potential inactivating mutations. Accordingly, any frameshift that occurred in the query does not influence the codon sequence alignment downstream. This implementation is based on the assumption that the reference transcript sequence comprises an intact ORF. This feature allows us to sequentially analyze each individual codon alignment and mask regions that potentially lead to alignment ambiguity.

To this end, we split the CESAR alignment segments that contain exactly three bases from the reference coding sequence. Then for each segment, we examine whether the corresponding query sequence encodes a sense codon. For instance, if the query sequence constitutes the number of nucleotides which is not multiple of three (frameshift), we mask this region by replacing it with "NNN". Furthermore, if the query sequence contains a stop codon, we also mask it with “NNN”. Figure 4.3 illustrates this principle in detail.

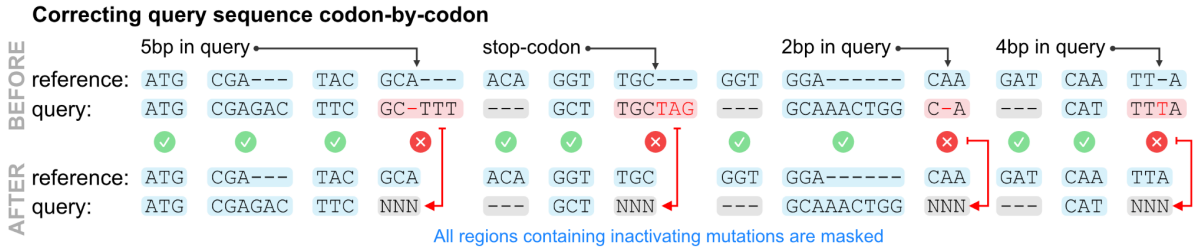

Figure 4.3 Correction of pairwise codon alignments

The figure shows pairwise codon alignments between the reference and query before (on top) and after (bottom) the sequence correction procedure. TOGA scans the codon alignment in a codon-by-codon manner. If the query sequence aligned to the reference codon comprises a frameshift or stop codon, TOGA masks this sequence replacing it with the NNN codon.

4.2.2 Using MACSE2.0 to produce multiple codon alignments

To produce multiple codon alignments, we employ the MACSE2.0 approach (Ranwez et al., 2011; Ranwez et al., 2018). This tool was specifically designed to take sequencing errors and other biological deviations from the intact reading frame into account. MACSE2.0 aligns DNA sequences at the nucleotide level with the possibility of including gap lengths that are not multiples of three while scoring the resulting alignments on their amino acid translation. This feature distinguishes this tool from other nucleotide aligners that have codon alignment functionality mode.

To additionally reduce the chance of alignment ambiguity, TOGA aligns transcript coding sequences in a “exon by exon” mode. Applying this restriction guarantees that different exons will not interfere with each other. In other terms, we ensure that codons belonging to different exons would never align with each other, minimizing the chance of generating the incorrect alignment. Of note, while TOGA has an implicit notion of exon boundaries since it works at the genomic level, methods that take protein sequences of annotated genes as input cannot implement the more accurate “exon-by-exon” mode because they are not aware of exon boundaries.

In this fashion, we created an additional workflow that enables TOGA to produce data for evolutionary selection analysis research. The combination of filters we implemented ensures that the probability of any single alignment inaccuracy is insignificant.

Figure 4.4 below demonstrates the efficiency of the proposed approach. In this example, I generated the AMPD3 gene alignment using (top) PRANK (another state-of-the art aligner, (Löytynoja, 2014)) in codon alignment mode and (bottom) the proposed MACSE2.0 based methodology. The sperm whale sequence of the *AMPD3* gene contains frameshifts that confuse the PRANK alignment procedure resulting in misalignments. Consequently, positive selection analysis reports a significant positive selection signal induced by these discrepancies. In contrast, our procedure masks all frameshifting regions, therefore it produces a proper codon alignment.

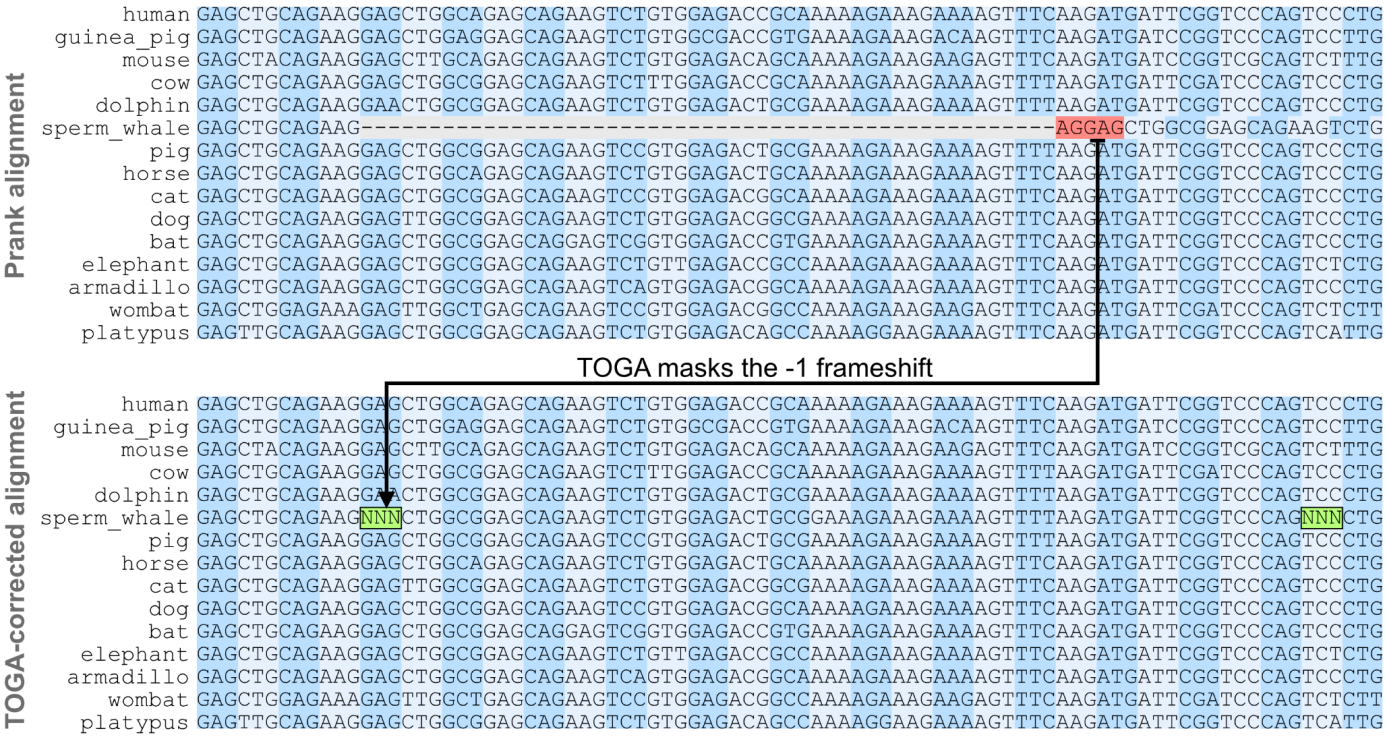

Figure 4.4 Comparison of codon alignment performed by Prank and TOGA extension

The figure shows two codon alignments of the AMPD3 gene produced by (top) PRANK in codon alignment mode and (bottom) using the proposed MACSE2.0-based methodology. The sperm whale sequence contains frameshifts (frameshifted sequence highlighted in red), confusing the prank alignment procedure resulting in misalignments. The frameshifting sequence is masked in the TOGA alignment (highlighted in green), which restores the ancestral reading frame in the sperm whale sequence.

In general, there are two primary sources of errors leading to the detection of nonhistorical selection signals: (i) codon alignment artifacts and (ii) sequencing/assembly inaccuracies. Sequentially, the last source is split into assembly errors that (i) mimic loss-of-function mutations and (ii) display false nucleotide substitutions. Unfortunately, the latter group is still an obstacle because such nucleotides are indistinguishable without additional information, such as mapped reads.

Notwithstanding, the proposed technique certainly diminishes the chance of introducing any alignment ambiguity by applying a series of filters and the MACSE approach. As regards sequencing errors, those leading to frameshifting and nonsense mutations are also covered by this method. After the filtering step, the codons comprising such artifacts are substituted with “NNN” codons. As a result, this technique provides reliable codon alignments for selection analysis that are unlikely to introduce a nonhistorical signal.

4.3 TOGA extensions summary

The set of potential applications of TOGA annotations extends beyond the examples presented in this chapter. For instance, the pipeline provides accurate predictions of gene inactivation events, which could be applied to identify clade-specific gene losses. Additionally, wide-scale annotations could be utilized to detect convergent evolutionary events, implementing the Forward Genomics concept (Hiller et al., 2012). Basically, TOGA provides the potential to create an entire ecosystem of extensions for various genomic data analyses.

5. General discussion

5.1 Summary

TOGA implements a novel paradigm of orthology inference and gene annotation based on neutral sequence divergence, contrasting to traditional methods that primarily rely on coding sequence. In this work, I evaluated different aspects of the TOGA pipeline, such as accuracy of orthology inference (part 3.1) and detection of lost genes (part 3.2). Additionally, I compared TOGA to a widely used Ensembl method (section 3.4.1) and evaluated gene annotation completeness on a BUSCO conserved gene set (section 3.4.2). Despite a completely different approach to orthology inference, the proposed method can compete with and even outperform generally accepted techniques. TOGA results can be applied in various comparative studies, such as phylogeny analysis (part 4.1) and selection screens (part 4.2). The proposed methodology easily scales to hundreds of genomes, and we applied it to produce annotations of 500 mammalian genomes, creating the largest comparative dataset so far (part 3.7). After this, we are planning to annotate 300 bird genomes. The method has already been applied in several studies, and two of them are already published in peer-reviewed journals (part 3.6). The results suggest that TOGA has a great potential to become a widely-used tool because of its scalability and reliability of the results.

5.2 TOGA limitations

Using neutral sequence divergence as a separating criterion to identify orthologous loci implies certain limitations on the molecular distance between the reference and query genomes. In this part, I discuss the application range of TOGA.

5.2.1 Annotating distant species

Orthologs and paralogs are distinguishable within the implemented paradigm until the expected nucleotide identity in the orthologous neutrally evolving sequences is higher than a specific threshold. This threshold is unquestionably less than one substitution per neutral site, because values higher than 1 implies complete randomization of neutral regions. This molecular distance corresponds to the divergence between primates and monotremes - on such evolutionary distance, the only feature that remains intact to distinguish orthologs is the synteny. However, accumulated recombinations also diminish the predictive power of this feature with time.

Furthermore, even if the expected sequence divergence exceeds 0.7 substitutions per neutral site, usually intronic sequences have already diverged enough so that the sensitive genome aligner is usually unable to align them. Such a level of divergence corresponds for example to the evolutionary distance between primates and marsupials. To illustrate the inverse relationship between the number of intronic or intergenic alignments and sequence divergence, I show orthologous chains with the human genome as the reference from various species (figure 5.1). The represented species cover molecular distances up to 2.5 substitutions per neutral site (human vs medaka).

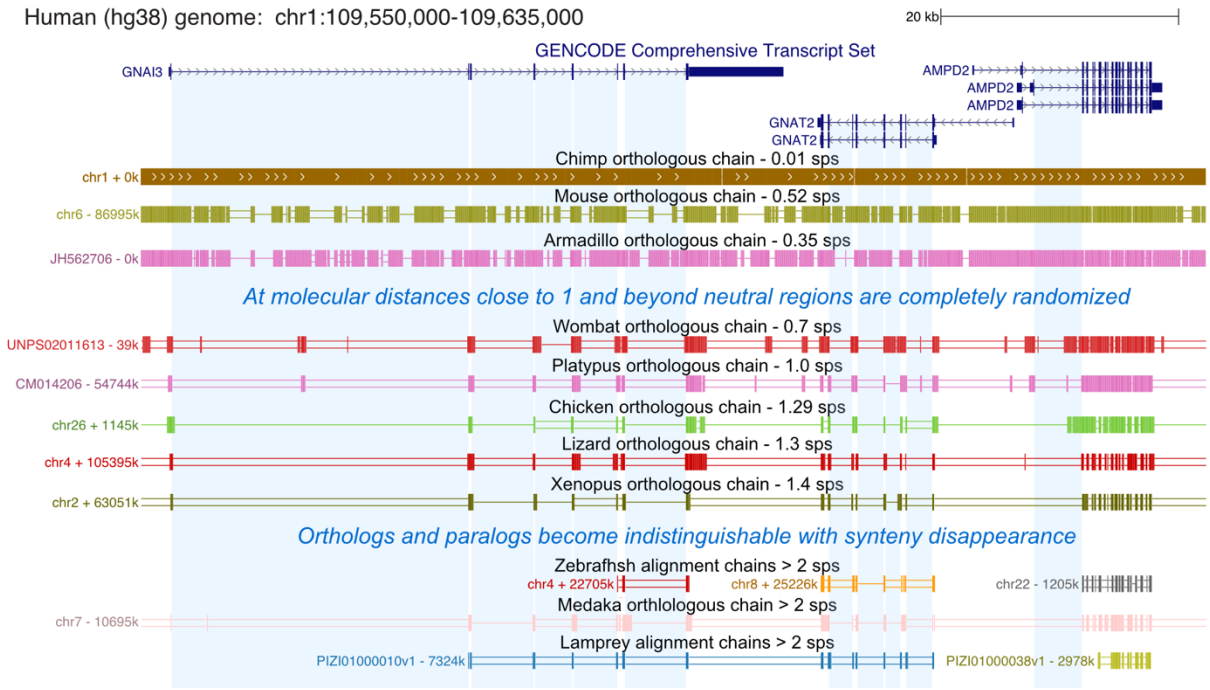

Figure 5.1 Appearance of orthologous chains on different molecular distances

UCSC genome browser screenshot shows a locus in the human genome containing genes GNAI3, GNAT2, and AMPD3. Intronic regions are highlighted with grey. It also shows orthologous alignment chains to various species, from a close relative (chimp) to a very distant lamprey indicating molecular distance to the species in substitutions per neutral site (sps). The figure shows that intronic alignment disappears when the molecular distance exceeds 0.7 sps, which affects the quality of the TOGA orthology inference (details in the main text).

The gradient boosting model was trained to infer orthologs between species with a neutral sequence divergence of up to ~0.5 substitutions per neutral site covering the entire Placentalia clade. However, it should be possible to train the orthologous loci classifier for more distant clades, potentially using different features or putting more weight on synteny. Such an alternative model could potentially extend TOGA applicability borders to more distant species.

For example, in the recent high-quality platypus genome assembly (Zhou et al., 2021), 18373 (97%) out of 18897 aligned genes are intersected by a chain that covers at least two genes. The bar plot below illustrates the distribution of platypus genes between alignment chains of various synteny (figure 5.2). The first column shows the total number of genes as a reference. According to this plot, most human orthologs in platypus appear in a context of conserved gene order, which can be exploited to better annotate more distant species, albeit at the expense of missing translocated orthologs.

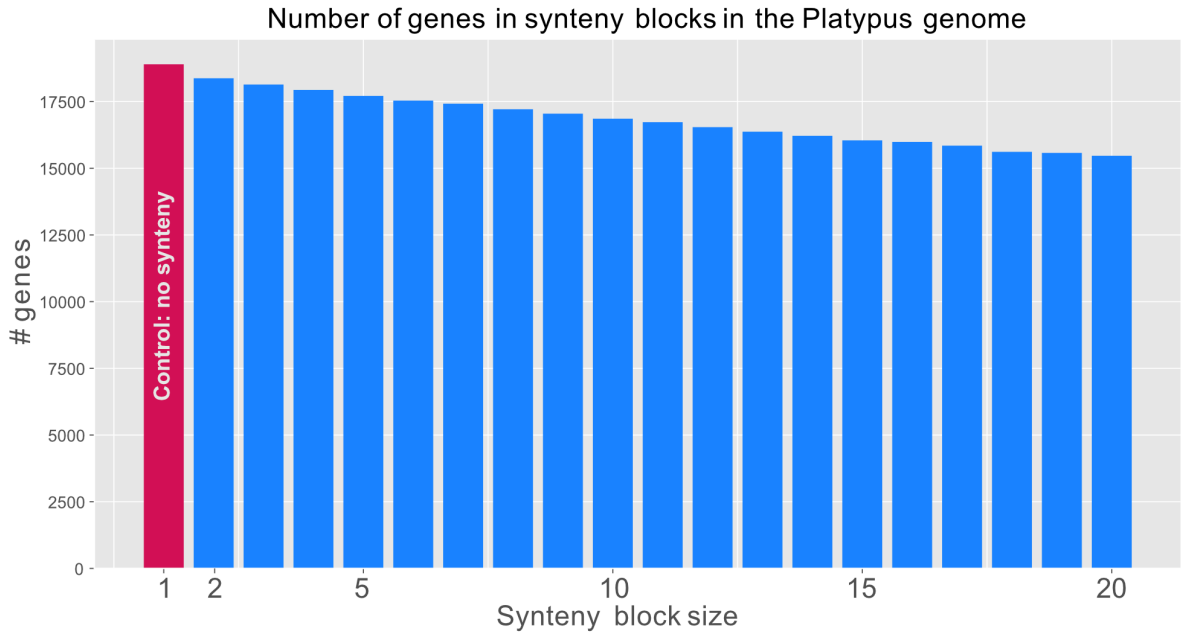

Figure 5.2 Synteny blocks in platypus

Barplot shows numbers of platypus genes that appear in synteny blocks of a given size. Here, synteny block size means the number of genes located in the conserved order.

Nevertheless, projecting genes and inferring orthologs is easier between species that are more closely related. Furthermore, by definition, reference-based approaches cannot annotate lineage-specific genes. Since more lineage-specific genes will exist between more distantly-related lineages, choosing a different reference that is more closely related is likely a better idea. For example, to comprehensively annotate marsupial genomes with TOGA, it is worth selecting a well-annotated marsupial (e.g., opossum) as the reference.

5.2.2 Clades outside mammalia

Theoretically, the TOGA approach is applicable to any pair of species that exhibit proper alignment between neutrally evolving sequences in orthologous regions. Despite the fact that the development of TOGA used primarily mammalian species, we successfully applied it to birds with chicken as the reference. To explore the applicability of TOGA to other

clades, we examined whether it could be potentially applied to infer orthologs in plant genomes, kindly provided by collaborators. The whole-genome alignment of these species revealed that the TOGA approach indeed could be applied to infer orthology in these species since the divergence of neutrally evolving sequences is surprisingly low (figure 5.3). Initially, we expected a higher degree of divergence between plant genomes.

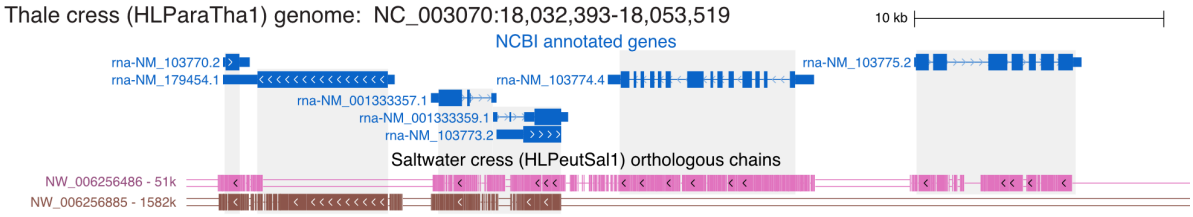

Figure 5.3 Genome alignment between two plant species

UCSC genome browser screenshot shows an arbitrarily selected locus in the thale cress genome and alignment chains to the saltwater cress. Orthologous alignment chains exhibit intronic and intergenic alignments, which indicates that TOGA methodology is potentially applicable to plant genomes.

Using TOGA, I have annotated the saltwater cress genome using the thale cress as a reference. For 48227 transcripts annotated by NCBI in the reference genome, TOGA could detect orthologs for 39233 (81%) of them, and this result could probably be improved by additional parameter optimizations. This example illustrates that the genome annotation paradigm proposed in this work has the potential to be extended to other eukaryotic species groups where the molecular distance between reference and the query allows distinguishing orthologs from paralogs.

5.2.3 Resolving big gene families

Resolving orthology relationships between big gene families is a challenging task for any orthology inference method and, up to this moment, this task is still unsolved. The most complicated cases are gene family clusters consisting of hundreds of relatively short (often single-exon) genes with a high sequence similarity, exemplified by *ZNF* and *KRTAP* gene families (subsections 3.3.1.3 and 3.4.1.1). Since the sequences are relatively short (about 100aa) and highly similar, gene tree methods are more likely to make a mistake when inferring orthology. Similarly, this is also challenging for TOGA since genome aligners are usually confused by multiple tandem duplications and cannot provide an adequate genomic context for each gene. Besides, features related to intronic alignment are obviously unavailable for single-exon genes.

The most extreme example of a challenging gene family is the odorant receptors (*OR*), representing ~1% of the whole coding sequence in mammals. In each mammalian species, they are designated by hundreds of genes. For example, the human genome comprises ~400

functional odorant receptor genes, concurrently with the same amount of inactivated copies (Matsui et al., 2010). These genes follow the "death and birth" evolutionary pattern implying constant gene family expansion and contraction through duplication and pseudogenization (Hughes et al., 2018). Some studies are explicitly focused on orthology inference within the odorant receptors family, and so far, it was feasible to split this family into several subfamilies and associate the number of copies with environmental adaptations (Lane et al., 2001; Niimura et al., 2014; Hughes et al., 2018). However, it resembles that the accurate inferring orthology for this gene family with automatic methods is impractical. However, there is a case where traditional methods outperform the TOGA approach in annotating short single-exon genes. Translocation of these genes leads to complete loss of genomic context, and therefore, TOGA cannot adequately infer orthology (figure 3.28).

5.2.4 Tandem duplications can confuse chaining procedure

Tandem gene duplications in the query genome may confuse genome aligners causing the duplication event undetected. In the chain interpretation, it appears as a single alignment chain covering the transcript in the reference. However, in the query coordinates, the chain bridges the beginning of the first copy to the end of the second one (figure 5.4). This problem arises from the inability of the chaining method to correctly align the tandem duplications.

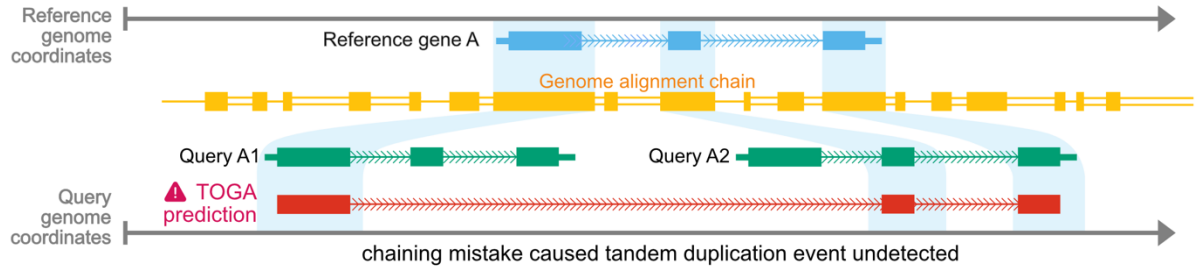

Figure 5.4 Genome alignment misinterprets tandem gene duplication in the query genome

The figure exemplifies a genome alignment chain that was unable to align tandem gene duplication correctly. Incorrect alignment caused the tandem duplication undetected from the reference genome perspective. In this case, TOGA will try to project a single gene A copy to the locus, containing two copies of this gene.

Conceptually, TOGA considers each predicted orthologous locus in the query genome for a given transcript-chain pair comprising a single corresponding transcript. Furthermore, CESAR itself is designed to retrieve an individual query transcript for the given reference coding sequence. Consequently, TOGA could produce a corrupted annotation, mixing exons from different actual copies if projecting through such a chain.

The phylogeny-grade exons filtering procedure described in part 4.1 also accounts for the potential chain corruption induced by tandem duplications, employing a lastz-based search for exon duplicates in the query locus. The motivation behind this procedure is that candidate exons could easily pass other filters, such as sequence similarity or requirement for the flanking alignment. Notwithstanding, the exon duplicate search procedure will capture this event, ensuring that non-homologous sequences do not appear in the phylogeny-grade dataset. Nevertheless, proper alignment of tandem duplications is a demanding challenge for genome alignment procedures which is still to be resolved.

5.3 TOGA-specific ortholog predictions

In section (3.4.1) devoted to comparing TOGA and Ensembl, I mentioned that TOGA consistently finds orthologs for approximately 1000 human genes that do not appear in the Ensembl dataset within the mammalian clade. In this section, I review these TOGA-specific orthology predictions in general. To analyze this set, I performed the gene set enrichment analysis and revealed statistically supported properties of TOGA-specific orthologous predictions. Additionally, I separately examined various features of these genes, such as gene length or the number of copies. These findings provide particular insights into TOGA methodology specificities and capabilities in comparison to traditional approaches.

I analyzed TOGA-specific predictions for the following query species: mouse, horse, and wombat. The latter represents a relatively distant Marsupalia clade, where TOGA performs worse than in closer Placentalia clade. I added this species to the analysis to check whether TOGA dataset particularities, if they exist, remain intact on higher evolutionary distances. My analysis showed that the following gene categories are overrepresented in the TOGA-specific orthologs set:

1. Essential and highly conserved genes, such as those encoding ribosomal proteins.
2. Short single-exon genes belonging to large gene families such as keratin-associated genes.

5.3.1 TOGA-only genes enrichment analysis

To detect statistically overrepresented gene terms within TOGA-specific predictions, I performed enrichment analysis using gProfiler (version e102_eg49_p15_7a9b4d6) (Reimand et al., 2007) on TOGA-specific orthologous genes prediction in mouse, horse, and wombat. The gProfiler enrichment test uses a tailor-made gSCS algorithm for false discovery rate corrections. The algorithm considers the hierarchical structure of gene ontology terms and should therefore give a tighter threshold to significant results. The experiment-wide significance threshold was set to 0.05, which corresponds to negative log10 of corrected P-value of 16. The tables below (2-4) introduce statistically overrepresented terms for each species starting from the mouse.

term_name	term_id	negative_log10_of_adj_p_value
keratinization	GO:0031424	34.07753986541037
Keratinization	REAC:R-HSA-6805567	32.69967362335063
keratinocyte differentiation	GO:0030216	29.651485191452146
epidermal cell differentiation	GO:0009913	25.1544653765264
intermediate filament	GO:0005882	24.238382906968774
intermediate filament cytoskeleton	GO:0045111	21.562021326978265
Ribosome, cytoplasmic	CORUM:306	20.703121972088226
skin development	GO:0043588	20.144504643051995
epidermis development	GO:0008544	19.712606490857702
Herpes simplex virus 1 infection	KEGG:05168	19.472737220409947
Developmental Biology	REAC:R-HSA-1266738	16.074061929579933

Table 2 Enrichment analysis of TOGA-specific orthologs in mouse (N = 955)

Enrichment analysis of TOGA-specific orthology predictions in the mouse genome suggests that the gene set is enriched with relatively short keratin-associated genes that appear in multiple copies. Besides, the "Ribosome, cytoplasmic" term requires particular attention since it consists of essential ribosomal proteins. Furthermore, the "Herpes simplex virus one infection" term comprises abundant zinc-finger genes, providing specific insights into TOGA orthology inference capabilities.

term_name	term_id	negative_log10_of_ad_p_value
Herpes simplex virus 1 infection	**KEGG:05168**	**37.12810062557136**
intermediate filament	**GO:0005882**	**25.87204570564712**
Keratinization	**REAC:R-HSA-6805567**	**24.001175010810233**
keratinization	**GO:0031424**	**23.62294218589523**
testis; spermatogonia cells [High]	HPA:0570813	22.738665578023447
testis; spermatogonia cells[≥Medium]	HPA:0570812	21.728572588680187
intermediate filament cytoskeleton	GO:0045111	21.254086478738834
testis; preleptotene spermatocytes[≥Medium]	HPA:0570782	19.473287745116952
testis; Sertoli cells[≥Low]	HPA:0570801	19.283264776743987
testis; preleptotene spermatocytes[≥Low]	HPA:0570781	18.17977789209747
keratinocyte differentiation	GO:0030216	17.438849806952007
testis; spermatogonia cells[≥Low]	HPA:0570811	16.000132824553216

Table 3 Enrichment analysis of TOGA-specific orthologs in horse (N = 1660)

Similar to the results for the mouse genome, the keratinization and herpes-related terms appeared again in the horse data, suggesting that these genes will definitely require additional analysis. Ribosome-related group representation, however, is not significant in this species.

term_name	term_id	negative_log10_of_ad_p_value
Herpes simplex virus 1 infection	KEGG:05168	18.888890027088735
hair	HPA:0230000	16.100817361596544

Table 4 Enrichment of TOGA-specific orthologs in wombat (N = 1190)

The wombat gene set analysis revealed only two overrepresented terms. However, the herpes-related term appeared also here, indicating that it firmly points to some TOGA distinctiveness compared to Ensembl. Another term, hair, is connected to previous results indirectly because it contains a general superset of keratin-associated genes. The ontology term generalization reflects that wombat represents a relatively distant lineage. It highlights that on such evolutionary distances TOGA specificities are smoothened.

Furthermore, I separately reviewed gene families' representation in TOGA-specific genes in all considered species (figure 5.5). As expected, Zinc-finger-containing (*ZNF*) and Keratin-associated (*KRTAP*) genes are overrepresented. The most surprising finding is that ribosomal proteins (*RPL*) occupy a prominent part of the TOGA-specific predictions in the mouse genome.

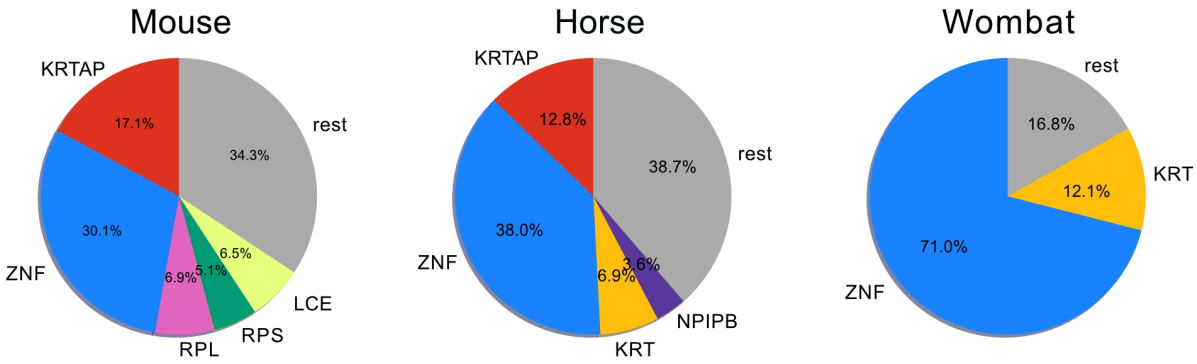

Figure 5.5 Overrepresented gene families in TOGA-specific predictions

The pie charts show overrepresented gene families in TOGA-specific orthology predictions for the mouse, horse, and wombat. According to this figure, ZNF and KRTAP genes are prevalent in these gene sets (details in the text).

In fact, the *RPL* genes must exist and be intact in any eukaryotic species, which suggests that for some reason, Ensembl could not associate human and mouse homologs as paralogs. Presumably, this is because these genes evolve under extreme purifying selection and therefore are virtually identical in mammals. Consequently, the gene tree-based methods cannot resolve the orthologous relationships between these genes because they require sequence variability for orthology inference.

KRTAP (Keratin-associated proteins) is another gene family overrepresented in the TOGA-specific orthology predictions set. These genes are essential for the formation of rigid and hair shafts. Representatives of this family are relatively short (encoding proteins of ~170aa long) single-exon genes, distributed in 5 tandemly arranged clusters in mammals. On average, this gene family constitutes ~200-300 individual genes, depending on the considered clade (Wu et al., 2008; Khan et al., 2014).

Furthermore, the Zinc-finger domain-containing genes family takes a significant part of TOGA-specific orthology predictions. These genes encode conserved DNA-binding zinc finger domains, mostly C2H2, that could be linked together and cover a great variety of possible recognized DNA sequences (Rosati et al., 1991). The zinc-finger domain is one of the most common DNA-binding motifs observed in eukaryotic transcriptional factors. In humans, zinc-finger-containing proteins occupy about ~4% of the protein-coding genes (Klug, 2010). They are often fused with various protein domains involving them in the regulation of diverse cellular

processes. Indeed, zinc-finger genes are recognized to be involved in transcriptional regulation, DNA repair, ubiquitin-mediated protein degradation, signal transduction, and numerous additional processes (Cassandri et al., 2017).

Since the *ZNF* finger domains exhibit high sequence similarity and appear in clusters, alignment chains usually cover multiple genes in a row, despite being primarily paralogous. These chains are treated as syntenic, which implies that they have a chance of being misclassified as orthologous. However, only a minor fraction of them exhibits flanking and intergenic alignment, which are important features for the TOGA decision-making process. Potentially, a fraction of TOGA-specific orthology predictions for ZNF genes indeed comprises false discoveries. However, no known method can resolve the orthology connections within this family with a high degree of accuracy given its complex evolutionary history. Therefore, statistical evaluation of TOGA quality in predicting ZNF finger genes also appears to be impractical. Figure 5.6 illustrates the extraordinary mass of chains that ordinarily cover a ZNF gene.

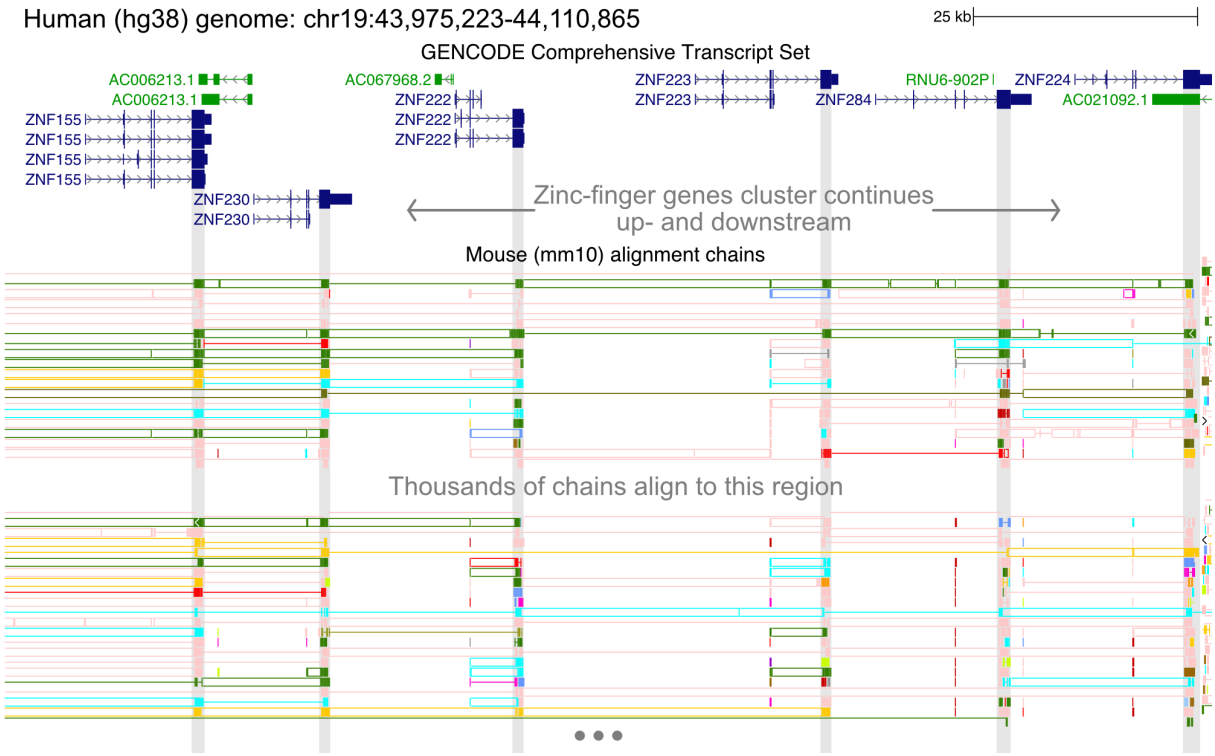

Figure 5.6 Alignment of ZNF gene cluster

UCSC genome browser screenshot shows mouse alignment chains to ZNF genes cluster. Since the zinc-finger domain is ubiquitous, and ZNF genes usually appear in clusters, each ZNF gene is covered with thousands of syntenic alignment chains. Orthology inference for such gene families is a nontrivial task (details in the main text).

5.3.2 Orthology classes representation

To confirm that TOGA-specific genes mainly belong to large gene families, I compared the relative proportion of many-to-many orthologs in all TOGA predictions against a set of TOGA-specific ones (figure 5.7). For visualization purposes, the illustrated class “one-to-many” class comprises both one-to-many and many-to-one categories.

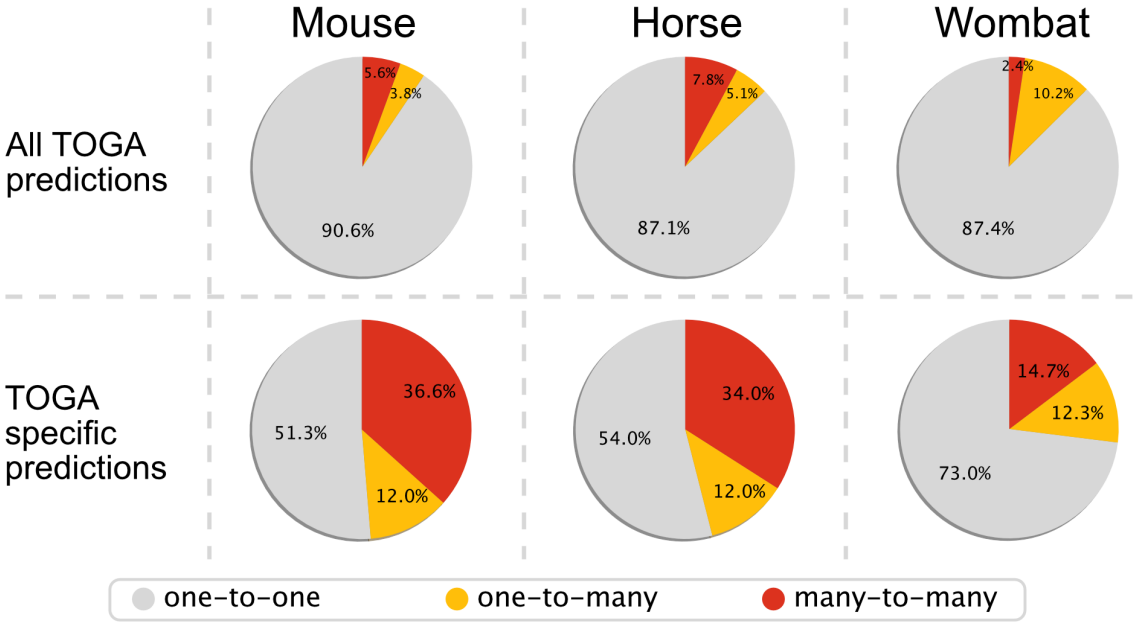

Figure 5.7 Relative orthology classes representation

The pie charts compare relative representations of different orthology classes (one-to-one, one-to-many, many-to-many) between (i) all TOGA projections excluding TOGA-specific ones (upper row) and (ii) TOGA-specific predictions. The comparison was performed for the mouse, horse, and wombat. For visualization purposes, one-to-many and many-to-one orthology classes are collapsed into a "one-to-many" group. The results show that TOGA-specific predictions have a substantially higher fraction of representatives of gene families (one-to-many and many-to-many) than the overall TOGA predictions set, where one-to-one orthology is prevalent.

According to data, one-to-one orthologs indeed occupy the majority (~90%) of the total TOGA predictions set, and only a minority belongs to the many-to-many class. However, the many-to-many category is more prevalent in the TOGA-specific orthology predictions.

It is worth noticing that many-to-many orthologs contribute only to 15% of predictions in the wombat set of TOGA-specific genes. Meanwhile, the many-to-many orthologs occupy less than 3% of the total number of genes in the total prediction set for the wombat. Therefore, the difference between the entire prediction and toga-specific sets is still significant. The most likely explanation is that in a more distant Marsupalia clade, the neutral sequence divergence is much higher; therefore, the expected number of detected orthologous connections is smaller.

The results indicate that TOGA-specific predictions are indeed enriched with genes belonging to large gene families. In the next section, I additionally examine the length of TOGA-specific predictions.

5.3.3 Analysis of the TOGA-specific orthologous gene lengths

The elevated presence of genes belonging to *ZNF*, *KRTAP*, and other large gene families suggested that TOGA-specific orthologs should be substantially shorter than average. In particular, I compared gene CDS length of two sets: (i) TOGA-specific orthologs (TOGA-only) and (ii) total set of TOGA predictions excluding genes that appear in the first set (all genes \ TOGA-only). The histograms showing distributions of gene lengths are shown in figure 5.8.

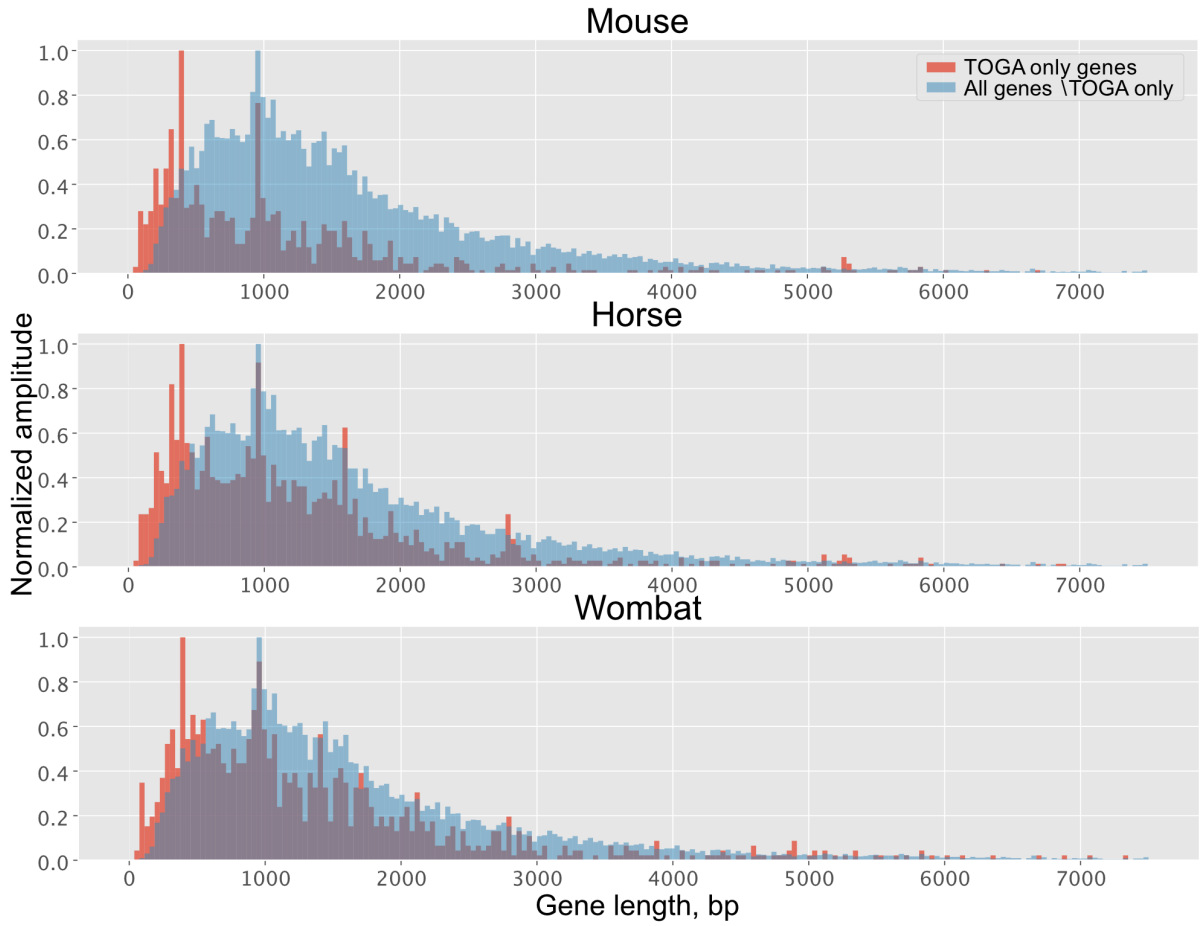

Figure 5.8 Predicted gene lengths distribution

The histograms compare gene length distributions between (i) all TOGA projections excluding TOGA-specific ones (blue) and (ii) TOGA-specific predictions (red) for three species: mouse, horse, and wombat. According to these plots, TOGA-specific orthologs are substantially shorter than the average.

In general, the overall histogram's appearance does not clarify whether the gene length difference is significant between the analyzed sets. To check whether this is the case, I applied a pairwise Student's T-test to this distribution (table 5).

Species	Mean - all	Med. - all	Mean - TOGA	Med. - TOGA	T value	P value
Mouse	1787	1341	1141	849	10.5718	4.7439e-26
Horse	1801	1347	1269	972	11.2555	2.6847e-29
Wombat	1775	1326	1456	1038	5.7780	7.6721e-09

Table 5 Gene length in TOGA-specific predictions

Apparently, the results confirm that the TOGA-specific genes are significantly shorter than the primary set of TOGA orthologs predictions. These results are consistent with the worse performance of the gene tree-based methods on shorter genes. In such cases using the whole genomic context compensates for the lack of data provided by short coding sequences.

Appendix A. Software and Data

Ensembl BIOMART annotation versions: 99, 101

BUSCO dataset version: Vertebrata odb9

Python: 3.6.5

XGBoost: 1.2.1

Networkx: 2.1

Nextflow: 20.04

PRANK: .170427

MACSE: 2.04

IQTree: 1.6.12

Appendix B. Annotated genome assemblies

species name	assembly ID	assembly accession	Species name	Assembly ID	assembly accession
Acinonyx jubatus	HLaciJub2	GCF_003709585.1	**Monodon monoceros**	HLmonMon2	Private
Acomys cahirinus	HLacoCah1	GCA_004027535.1	**Mormoops blainvillei**	HLmorBla1	GCA_004026545.1
Acomys russatus	HLacoRus1	GCA_903995435.1	**Moschus berezovskii**	HLmosBer1	GCA_006459085.1
Aeorestes cinereus	HLaeoCin1	GCA_011751065.1	**Moschus chrysogaster**	HLmosChr1	GCA_006461725.1
Aepyceros melampus	HLaepMel1	GCA_006408695.1	**Moschus moschiferus**	HLmosMos1	GCA_004024705.2
Ailuropoda melanoleuca	HLailMel2	GCF_002007445.1	**Mungos mungo**	HLmunMug1	GCA_004023785.1
Ailurus fulgens	HLailFul2	None	**Muntiacus crinifrons**	HLmunCri1	GCA_006408485.1
Alces alces	HLalcAlc1	GCA_007570765.1	**Muntiacus muntjak**	HLmunMun1	GCA_008782695.1
Allactaga bullata	HLallBul1	GCA_004027895.1	**Muntiacus reevesi**	HLmunRee1	GCA_008787405.1
Allenopithecus nigroviridis	HLallNig1	None	**Murina aurata feae**	HLmurAurFea1	GCA_004026665.1
Alouatta palliata	HLaloPal1	GCA_004027835.1	**Mus caroli**	HLmusCar1	GCF_900094665.1
Ammotragus lervia	HLammLer1	GCA_002201775.1	**Mus musculus**	mm10	GCF_000001635.20
Anoura caudifer	HLanoCau1	GCA_004027475.1	**Mus musculus**	mm39	GCF_000001635.27
Antechinus flavipes	HLantFla1	GCA_016432865.1	**Mus pahari**	HLmusPah1	GCF_900095145.1
Antidorcas marsupialis	HLantMar1	GCA_006408585.1	**Mus spicilegus**	HLmusSpi1	GCA_003336285.1
Antilocapra americana	HLantAme1	GCA_007570785.1	**Mus spretus**	HLmusSpr1	GCA_001624865.1
Antrozous pallidus	HLantPal1	GCA_007922775.1	**Muscardinus avellanarius**	HLmusAve1	GCA_004027005.1
Aotus nancymaae	aotNan1	GCF_000952055.2	**Mustela erminea**	HLmusErm1	GCF_009829155.1
Aplodontia rufa	HLaplRuf1	GCA_004027875.1	**Mustela putorius**	HLmusPut1	GCA_902460205.1

Apodemus sylvaticus	HLapoSyl1	GCA_001305905.1	**Mustela putorius furo**	HLmusFur2	GCA_011764305.1
Arctocephalus gazella	HLarcGaz2	GCA_900642305.1	**Myocastor coypus**	HLmyoCoy1	GCA_004027025.1
Artibeus jamaicensis	HLartJam1	GCA_004027435.1	**Myodes glareolus**	HLmyoGla2	GCA_902806735.1
Artibeus jamaicensis	HLartJam2	GCA_014825515.1	**Myotis albescens**	HLmyoAlb1	Private
Arvicanthis niloticus	HLarvNil1	GCA_011762505.1	**Myotis alcathoe**	HLmyoAlc1	Private
Arvicola amphibius	HLarvAmp1	GCA_903992535.1	**Myotis blythii**	HLmyoBly1	Private
Ateles geoffroyi	HLateGeo1	GCA_004024785.1	**Myotis bocagii**	HLmyoBoc1	Private
Axis porcinus	HLaxiPor1	GCA_003798545.1	**Myotis brandtii**	myoBra1	GCF_000412655.1
Balaena mysticetus	HLbalMys1	None	**Myotis californicus**	HLmyoCal1	Private
Balaenoptera acutorostrata	balAcu1	GCF_000493695.1	**Myotis capaccinii**	HLmyoCap1	Private
Balaenoptera bonaerensis	HLbalBon1	GCA_000978805.1	**Myotis dasycneme**	HLmyoDas1	Private
Balaenoptera edeni	HLbalEde1	None	**Myotis daubentonii**	HLmyoDau1	Private
Balaenoptera musculus	HLbalMus1	GCA_009873245.1	**Myotis davidii**	myoDav1	GCF_000327345.1
Balaenoptera physalus	HLbalPhy1	GCA_008795845.1	**Myotis elegans**	HLmyoEle1	Private
Bassariscus sumichrasti	HLbasSum1	None	**Myotis lucifugus**	HLmyoLuc1	None
Beatragus hunteri	HLbeaHun1	GCA_004027495.1	**Myotis lucifugus**	myoLuc2	GCF_000147115.1
Bison bison bison	bisBis1	GCF_000754665.1	**Myotis myotis**	HLmyoMyo6	None
Bos frontalis	HLbosFro1	GCA_007844835.1	**Myotis mystacinus**	HLmyoMys1	Private
Bos gaurus	HLbosGau1	GCA_014182915.1	**Myotis ricketti**	HLmyoRic1	Private
Bos grunniens	HLbosGru1	GCA_005887515.2	**Myotis scotti**	HLmyoSco1	Private
Bos indicus	HLbosInd2	GCA_002933975.1	**Myotis septentrionalis**	HLmyoSep1	None
Bos mutus	HLbosMut2	GCA_007646595.3	**Myotis siligorensis**	HLmyoSil1	Private
Bos taurus	bosTau9	GCF_002263795.1	**Myotis siligorensis alticraniatus**	HLmyoSilAlt1	Private
Bubalus bubalis	HLbubBub2	GCF_003121395.1	**Myotis thysanodes**	HLmyoThy1	Private
Callithrix jacchus	HLcalJac4	GCA_011100555.1	**Myotis volans**	HLmyoVol1	Private
Callithrix pygmaea	HLcalPym1	None	**Myotis welwitschii**	HLmyoWel1	Private
Callorhinus ursinus	HLcalUrs1	GCF_003265705.1	**Myrmecophaga tridactyla**	HLmyrTri1	GCA_004026745.1
Camelus bactrianus	HLcamBac1	GCF_000767855.1	**Nanger granti**	HLnanGra1	GCA_006408635.1
Camelus dromedarius	HLcamDro2	GCF_000803125.2	**Nannospalax galili**	nanGal1	GCF_000622305.1
Camelus ferus	HLcamFer3	GCF_009834535.1	**Nasalis larvatus**	nasLar1	GCA_000772465.1
Canis lupus dingo	HLcanLupDin1	GCF_003254725.1	**Nasua narica**	HLnasNar1	None
Canis lupus familiaris	canFam4	GCA_011100685.1	**Natalus tumidirostris**	HLnatTum1	Private
Canis lupus familiaris	canFam5	GCA_005444595.1	**Neofelis nebulosa**	HLneoNeb1	None
Capra aegagrus	HLcapAeg1	GCA_000765075.1	**Neomonachus schauinslandi**	neoSch1	GCF_002201575.1

Capra hircus	HLcapHir2	GCF_001704415.1	**Neophocaena asiaeorientalis**	HLneoAsi1	GCF_003031525.1
Capra ibex	HLcapIbe1	GCA_006410555.1	**Neotoma lepida**	HLneoLep1	GCA_001675575.1
Capra sibirica	HLcapSib1	GCA_003182615.2	**Neotragus moschatus**	HLneoMos1	GCA_006410615.1
Capreolus pygargus	HLcapPyg1	GCA_012922965.1	**Neotragus pygmaeus**	HLneoPyg1	GCA_006410875.1
Carlito syrichta	tarSyr2	GCF_000164805.1	**Neovison vison**	HLneoVis1	GCA_900108605.1
Carollia perspicillata	HLcarPer2	Private	**Noctilio leporinus**	HLnocLep1	GCA_004026585.1
Carollia perspicillata	HLcarPer3	GCA_004027735.1	**Noctilio leporinus**	HLnocLep2	Private
Castor canadensis	HLcasCan3	None	**Nomascus leucogenys**	HLnomLeu4	GCF_006542625.1
Catagonus wagneri	HLcatWag1	GCA_004024745.2	**Notamacropus eugenii**	HLnotEug3	None
Cavia aperea	cavApe1	GCA_000688575.1	**Notamacropus eugenii**	macEug2	GCA_000004035.1
Cavia porcellus	cavPor3	GCF_000151735.1	**Nyctereutes procyonoides**	HLnycPro3	Private
Cavia tschudii	HLcavTsc1	GCA_004027695.1	**Nycticebus coucang**	HLnycCou1	GCA_004027815.1
Cebus albifrons	HLcebAlb1	GCA_004027755.1	**Nycticeius humeralis**	HLnycHum2	GCA_007922795.1
Cebus capucinus imitator	cebCap1	GCF_001604975.1	**Ochotona princeps**	ochPri3	GCF_000292845.1
Cephalophus harveyi	HLcepHar1	GCA_006410635.1	**Octodon degus**	octDeg1	GCF_000260255.1
Ceratotherium simum cottoni	HLcerSimCot1	GCA_004027795.1	**Odobenus rosmarus**	HLodoRos1	None
Ceratotherium simum simum	cerSim1	GCF_000283155.1	**Odobenus rosmarus divergens**	odoRosDiv1	GCF_000321225.1
Cercocebus atys	cerAty1	GCF_000955945.1	**Odocoileus hemionus hemionus**	HLodoHem1	GCA_004115125.1
Cercopithecus mona	HLcerMon1	GCA_014849445.1	**Odocoileus virginianus**	HLodoVir2	None
Cercopithecus neglectus	HLcerNeg1	GCA_004027615.1	**Odocoileus virginianus**	HLodoVir3	GCA_014726795.1
Cervus elaphus hippelaphus	HLcerEla1	GCA_002197005.1	**Odocoileus virginianus texanus**	HLodoVir1	GCF_002102435.1
Cervus hanglu yarkandensis	HLcerHanYar1	GCA_010411085.1	**Okapia johnstoni**	HLokaJoh2	None
Cheirogaleus medius	HLcheMed1	GCA_008086735.1	**Ondatra zibethicus**	HLondZib1	GCA_004026605.1
Chinchilla lanigera	chiLan1	GCF_000276665.1	**Onychomys torridus**	HLonyTor1	GCA_903995425.1
Chlorocebus sabaeus	chlSab2	GCF_000409795.2	**Orcinus orca**	orcOrc1	GCF_000331955.2
Choloepus didactylus	HLchoDid1	GCA_004027855.1	**Oreamnos americanus**	HLoreAme1	GCA_009758055.1
Choloepus didactylus	HLchoDid2	GCF_015220235.1	**Oreotragus oreotragus**	HLoreOre1	GCA_006410675.1
Choloepus hoffmanni	HLchoHof3	None	**Ornithorhynchus anatinus**	HLornAna3	GCA_004115215.1
Chrysochloris asiatica	chrAsi1	GCF_000296735.1	**Orycteropus afer afer**	oryAfe1	GCF_000298275.1
Coendou prehensilis	HLcoePre1	None	**Oryctolagus cuniculus**	HLoryCun3	GCA_009806435.1
Colobus angolensis palliatus	colAng1	GCF_000951035.1	**Oryctolagus cuniculus**	oryCun2	GCF_000003625.2
Condylura cristata	conCri1	GCF_000260355.1	**Oryctolagus cuniculus cuniculus**	HLoryCunCun4	GCA_013371645.1
Connochaetes taurinus	HLconTau2	None	**Oryx dammah**	HLoryDam1	None
Craseonycteris thonglongyai	HLcraTho1	GCA_004027555.1	**Oryx gazella**	HLoryGaz1	GCA_003945745.1

Cricetomys gambianus	HLcriGam1	GCA_004027575.1	**Osphranter rufus**	HLospRuf1	None
Cricetulus griseus	HLcriGri3	GCF_003668045.1	**Otolemur garnettii**	otoGar3	GCF_000181295.1
Crocuta crocuta	HLcroCro1	GCA_008692635.1	**Ovis ammon**	HLoviAmm1	GCA_003121645.1
Cryptoprocta ferox	HLcryFer2	None	**Ovis aries**	HLoviAri5	GCA_011170295.1
Ctenodactylus gundi	HLcteGun1	GCA_004027205.1	**Ovis canadensis**	HLoviCan2	GCA_004026945.1
Ctenomys sociabilis	HLcteSoc1	GCA_004027165.1	**Ovis canadensis canadensis**	HLoviCan1	GCA_001039535.1
Cynomys gunnisoni	HLcynGun1	GCA_011316645.1	**Ovis nivicola lydekkeri**	HLoviNivLyd1	GCA_903231385.1
Cynopterus brachyotis	HLcynBra1	GCA_009793145.1	**Ovis orientalis**	HLoviOri1	GCA_014523465.1
Damaliscus lunatus	HLdamLun1	GCA_006408505.1	**Pan paniscus**	panPan3	GCF_013052645.1
Dasyprocta punctata	HLdasPun1	GCA_004363535.1	**Pan troglodytes**	panTro6	GCF_002880755.1
Dasypus novemcinctus	dasNov3	GCF_000208655.1	**Panthera leo**	HLpanLeo1	GCA_008795835.1
Daubentonia madagascariensis	HLdauMad1	GCA_004027145.1	**Panthera onca**	HLpanOnc1	GCA_004023805.1
Delphinapterus leucas	HLdelLeu2	GCF_002288925.2	**Panthera onca**	HLpanOnc2	None
Desmodus rotundus	HLdesRot2	Private	**Panthera pardus**	HLpanPar1	GCF_001857705.1
Dicerorhinus sumatrensis	HLdicSum1	GCA_002844835.1	**Panthera tigris altaica**	panTig1	GCF_000464555.1
Diceros bicornis	HLdicBic1	GCA_004027315.2	**Pantholops hodgsonii**	panHod1	GCF_000400835.1
Didelphis virginiana	HLdidVir1	None	**Papio anubis**	HLpapAnu5	GCF_008728515.1
Dinomys branickii	HLdinBra1	GCA_004027595.1	**Paradoxurus hermaphroditus**	HLparHer1	GCA_004024585.1
Dipodomys ordii	dipOrd2	GCF_000151885.1	**Pedetes capensis**	HLpedCap1	GCA_007922755.1
Dipodomys stephensi	HLdipSte1	GCA_004024685.1	**Peponocephala electra**	HLpepEle1	None
Dolichotis patagonum	HLdolPat1	GCA_004027295.1	**Perognathus longimembris**	HLperLonPac1	GCA_004363475.1
Dugong dugon	HLdugDug1	GCA_015147995.1	**Peromyscus californicus insignis**	HLperCal2	GCA_007827085.2
Echinops telfairi	echTel2	GCF_000313985.1	**Peromyscus crinitus**	HLperCri1	None
Eidolon dupreanum	HLeidDup1	None	**Peromyscus eremicus**	HLperEre1	GCA_902702925.1
Eidolon helvum	HLeidHel2	None	**Peromyscus leucopus**	HLperLeu1	GCF_004664715.1
Elaphurus davidianus	HLelaDav1	GCA_002443075.1	**Peromyscus maniculatus bairdii**	HLperManBai2	GCA_003704035.1
Elephantulus edwardii	eleEdw1	GCF_000299155.1	**Peromyscus nasutus**	HLperNas1	None
Elephas maximus	HLeleMax1	None	**Peromyscus polionotus**	HLperPol1	GCA_003704135.2
Ellobius lutescens	HLellLut1	GCA_001685075.1	**Petromus typicus**	HLpetTyp1	GCA_004026965.1
Ellobius talpinus	HLellTal1	GCA_001685095.1	**Phalanger gymnotis**	HLphaGym1	None
Enhydra lutris kenyoni	enhLutKen1	GCF_002288905.1	**Phascolarctos cinereus**	HLphaCin1	GCF_002099425.1
Enhydra lutris nereis	enhLutNer1	GCA_006410715.1	**Phataginus tricuspis**	HLphaTri2	None
Eonycteris spelaea	HLeonSpe1	GCA_003508835.1	**Philantomba maxwellii**	HLphiMax1	GCA_006410695.1
Eptesicus fuscus	eptFus1	GCF_000308155.1	**Phoca vitulina**	HLphoVit1	GCF_004348235.1

Equus asinus	HLequAsi1	GCF_001305755.1	**Phocoena phocoena**	HLphoPho1	GCA_004363495.1
Equus asinus asinus	HLequAsiAsi2	GCA_003033725.1	**Phocoena phocoena**	HLphoPho2	None
Equus burchellii boehmi	HLequQuaBoe1	None	**Phocoena sinus**	HLphoSin1	GCF_008692025.1
Equus caballus	equCab3	GCF_002863925.1	**Phyllostomus discolor**	HLphyDis3	None
Equus przewalskii	equPrz1	GCF_000696695.1	**Physeter catodon**	HLphyCat2	GCF_002837175.2
Erethizon dorsatum	HLereDor1	GCA_006547115.1	**Physeter catodon**	phyCat1	GCF_000472045.1
Erignathus barbatus	HLeriBar1	None	**Piliocolobus tephrosceles**	HLpilTep2	GCF_002776525.3
Erinaceus europaeus	eriEur2	GCF_000296755.1	**Pipistrellus kuhlii**	HLpipKuh2	None
Erythrocebus patas	HLeryPat1	GCA_004027335.1	**Pipistrellus pipistrellus**	HLpipPip1	GCA_004026625.1
Eschrichtius robustus	HLescRob1	GCA_004363415.1	**Pipistrellus pipistrellus**	HLpipPip2	GCA_903992545.1
Eubalaena glacialis	HLeubGla1	None	**Pithecia pithecia**	HLpitPit1	GCA_004026645.1
Eubalaena japonica	HLeubJap1	GCA_004363455.1	**Platanista minor**	HLplaMin1	GCA_004363435.1
Eudorcas thomsonii	HLeudTho1	GCA_006408755.1	**Plecturocebus donacophilus**	HLpleDon1	GCA_004027715.1
Eulemur flavifrons	HLeulFla1	None	**Pongo abelii**	ponAbe3	GCF_002880775.1
Eulemur flavifrons	eulFla1	GCA_001262665.1	**Pontoporia blainvillei**	HLponBla1	GCA_011754075.1
Eulemur fulvus	HLeulFul1	GCA_004027275.1	**Potos flavus**	HLpotFla1	None
Eulemur macaco	eulMac1	GCA_001262655.1	**Prionailurus bengalensis euptilurus**	HLpriBen1	GCA_005406085.1
Eulemur mongoz	HLeulMon1	None	**Procapra przewalskii**	HLproPrz1	GCA_006410515.1
Eumetopias jubatus	HLeumJub1	GCF_004028035.1	**Procavia capensis**	HLproCap3	GCA_004026925.2
Felis catus	felCat9	GCF_000181335.3	**Procyon lotor**	HLproLot1	None
Felis nigripes	HLfelNig1	GCA_004023925.1	**Prolemur simus**	HLproSim1	GCA_003258685.1
Fukomys damarensis	HLfukDam2	GCF_012274545.1	**Propithecus coquereli**	proCoq1	GCF_000956105.1
Galeopterus variegatus	HLgalVar2	GCA_004027255.2	**Przewalskium albirostris**	HLprzAlb1	GCA_006408465.1
Giraffa camelopardalis	HLgirCam1	GCA_006408565.1	**Psammomys obesus**	HLpsaObe1	GCA_002215935.2
Giraffa camelopardalis	HLgirCam2	None	**Pseudocheirus occidentalis**	HLpseOcc1	None
Giraffa tippelskirchi	HLgirTip1	GCA_001651235.1	**Pseudochirops corinnae**	HLpseCor1	None
Glis glis	HLgliGli1	GCA_004027185.1	**Pseudochirops cupreus**	HLpseCup1	None
Globicephala melas	HLgloMel1	GCF_006547405.1	**Pteronotus parnellii**	ptePar1	GCA_000465405.1
Glossophaga soricina	HLgloSor2	Private	**Pteronura brasiliensis**	HLpteBra1	GCA_004024605.1
Gorilla gorilla gorilla	gorGor6	GCF_008122165.1	**Pteronura brasiliensis**	HLpteBra2	None
Gracilinanus agilis	HLgraAgi1	GCA_016433145.1	**Pteropus alecto**	pteAle1	GCF_000325575.1
Grammomys surdaster	HLgraSur1	GCF_004785775.1	**Pteropus giganteus**	HLpteGig1	GCA_902729225.1
Graphiurus kelleni	HLgraKel3	Private	**Pteropus pselaphon**	HLptePse1	GCA_014363405.1
Graphiurus murinus	HLgraMur1	GCA_004027655.1	**Pteropus rufus**	HLpteRuf1	None

Gulo gulo	HLgulGul1	GCA_900006375.2	**Pteropus vampyrus**	HLpteVam2	GCF_000151845.1
Gymnobelideus leadbeateri	HLgymLea1	GCA_011680675.1	**Puma concolor**	HLpumCon1	GCF_003327715.1
Halichoerus grypus	HLhalGry1	GCA_012393455.1	**Puma yagouaroundi**	HLpumYag1	GCA_014898765.1
Helogale parvula	HLhelPar1	GCA_004023845.1	**Pygathrix nemaeus**	HLpygNem1	GCA_004024825.1
Hemitragus hylocrius	HLhemHyl1	GCA_004026825.1	**Rangifer tarandus**	HLranTar1	GCA_004026565.1
Heterocephalus glaber	hetGla2	GCF_000247695.1	**Rangifer tarandus granti**	HLranTarGra2	GCA_014898785.1
Heterohyrax brucei	HLhetBru1	GCA_004026845.1	**Raphicerus campestris**	HLrapCam1	GCA_006410735.1
Hippopotamus amphibius	HLhipAmp1	GCA_002995585.1	**Rattus norvegicus**	HLratNor7	GCA_015227675.1
Hippopotamus amphibius	HLhipAmp3	GCA_004027065.2	**Rattus norvegicus**	rn6	GCF_000001895.5
Hipposideros armiger	HLhipArm1	GCF_001890085.1	**Rattus rattus**	HLratRat7	GCF_011064425.1
Hipposideros galeritus	HLhipGal1	GCA_004027415.1	**Redunca redunca**	HLredRed1	GCA_006410935.1
Hippotragus equinus	HLhipEqu1	GCA_016433095.1	**Rhinoceros unicornis**	HLrhiUni1	None
Hippotragus niger niger	HLhipNig1	GCA_006942125.1	**Rhinolophus ferrumequinum**	HLrhiFer5	None
Homo sapiens	hg38	GCF_000001405.38	**Rhinolophus sinicus**	HLrhiSin1	GCF_001888835.1
Hyaena hyaena	HLhyaHya1	GCA_003009895.1	**Rhinopithecus bieti**	rhiBie1	GCF_001698545.1
Hydrochoerus hydrochaeris	HLhydHyd1	GCA_004027455.1	**Rhinopithecus roxellana**	HLrhiRox2	GCF_007565055.1
Hydrodamalis gigas	HLhydGig1	GCA_013391785.1	**Rhizomys pruinosus**	HLrhiPru1	GCA_009823505.1
Hydropotes inermis	HLhydIne1	GCA_006459105.1	**Rhombomys opimus**	HLrhoOpi1	GCA_010120015.1
Hylobates moloch	HLhylMol2	GCF_009828535.2	**Rousettus aegyptiacus**	HLrouAeg4	None
Hystrix cristata	HLhysCri1	GCA_004026905.1	**Rousettus leschenaultii**	HLrouLes1	GCA_015472975.1
Ictidomys tridecemlineatus	speTri2	GCF_000236235.1	**Rousettus madagascariensis**	HLrouMad1	None
Indri indri	HLindInd1	GCA_004363605.1	**Saccopteryx bilineata**	HLsacBil1	Private
Inia geoffrensis	HLlniGeo1	GCA_004363515.1	**Saguinus imperator**	HLsagImp1	GCA_004024885.1
Jaculus jaculus	jacJac1	GCF_000280705.1	**Saiga tatarica**	HLsaiTat1	GCA_004024985.1
Kobus ellipsiprymnus	HLkobEll1	GCA_006410655.1	**Saimiri boliviensis**	HLsaiBol1	None
Kobus leche leche	HLkobLecLec1	GCA_014926565.1	**Saimiri boliviensis boliviensis**	saiBol1	GCF_000235385.1
Kogia breviceps	HLkogBre1	GCA_004363705.1	**Sapajus apella**	HLsapApe1	GCF_009761245.1
Lagenorhynchus obliquidens	HLlagObl1	GCF_003676395.1	**Sarcophilus harrisii**	HLsarHar2	GCF_902635505.1
Lama glama	HLlamGla1	None	**Scalopus aquaticus**	HLscaAqu1	GCA_004024925.1
Lama glama chaku	HLlamGlaCha1	GCA_013239585.1	**Scarturus elater**	HLallEla1	Private
Lama guanicoe cacsilensis	HLlamGuaCac1	GCA_013239625.1	**Sciurus carolinensis**	HLsciCar1	GCA_902686445.1
Lasiurus borealis	HLlasBor1	GCA_004026805.1	**Sciurus vulgaris**	HLsciVul1	GCA_902686455.1
Lemur catta	HLlemCat1	GCA_004024665.1	**Semnopithecus entellus**	HLsemEnt1	GCA_004025065.1
Leptonychotes weddellii	lepWed1	GCF_000349705.1	**Sigmodon hispidus**	HLsigHis1	GCA_004025045.1

Leptonycteris yerbabuenae	HLlepYer1	None	**Solenodon paradoxus**	HLsolPar1	GCA_004363575.1
Lepus americanus	HLlepAme1	GCA_004026855.1	**Sorex araneus**	sorAra2	GCF_000181275.1
Lepus timidus	HLlepTim1	GCA_009760805.1	**Sousa chinensis**	HLsouChi1	GCA_007760645.1
Lipotes vexillifer	lipVex1	GCF_000442215.1	**Spermophilus dauricus**	HLspeDau1	GCA_002406435.1
Litocranius walleri	HLlitWal1	GCA_006410535.1	**Spilogale gracilis**	HLspiGra1	GCA_004023965.1
Lontra canadensis	HLlonCan1	GCF_010015895.1	**Sturnira hondurensis**	HLstuHon1	GCA_014824575.1
Loxodonta africana	HLloxAfr4	None	**Submyotodon latirostris**	HLsubLat1	Private
Lutra lutra	HLlutLut1	GCA_902655055.1	**Suricata suricatta**	HLsurSur1	GCF_006229205.1
Lycaon pictus	HLlycPic2	GCA_004216515.1	**Suricata suricatta**	HLsurSur2	GCA_004023905.1
Lycaon pictus	HLlycPic3	None	**Sus scrofa**	susScr11	GCF_000003025.6
Lynx canadensis	HLlynCan1	GCF_007474595.1	**Sylvicapra grimmia**	HLsylGri1	GCA_006408735.1
Lynx pardinus	HLlynPar1	GCA_900661375.1	**Sylvilagus bachmani**	HLsylBac1	None
Macaca fascicularis	HLmacFas6	GCA_012559485.1	**Syncerus caffer**	HLsynCaf1	GCA_902500845.1
Macaca fuscata	HLmacFus1	None	**Tachyglossus aculeatus**	HLtacAcu1	GCA_015852505.1
Macaca mulatta	rheMac10	GCF_003339765.1	**Tadarida brasiliensis**	HLtadBra1	GCA_004025005.1
Macaca nemestrina	macNem1	GCF_000956065.1	**Tadarida brasiliensis**	HLtadBra2	Private
Macroglossus sobrinus	HLmacSob1	GCA_004027375.1	**Talpa occidentalis**	HLtalOcc1	GCA_014898055.1
Macropus fuliginosus	HLmacFul1	None	**Tamandua tetradactyla**	HLtamTet1	GCA_004025105.1
Macropus giganteus	HLmacGig1	None	**Tapirus indicus**	HLtapInd1	GCA_004024905.1
Macrotus californicus	HLmacCal1	GCA_007922815.1	**Tapirus indicus**	HLtapInd2	None
Madoqua kirkii	HLmadKir1	GCA_006408675.1	**Tapirus terrestris**	HLtapTer1	GCA_004025025.1
Mandrillus leucophaeus	manLeu1	GCF_000951045.1	**Taxidea taxus jeffersonii**	HLtaxTax1	GCA_003697995.1
Mandrillus sphinx	HLmanSph1	GCA_004802615.1	**Theropithecus gelada**	HLtheGel1	GCF_003255815.1
Manis javanica	HLmanJav1	GCF_001685135.1	**Thryonomys swinderianus**	HLthrSwi1	GCA_004025085.1
Manis javanica	HLmanJav2	GCA_014570535.1	**Thylacinus cynocephalus**	HLthyCyn1	GCA_007646695.1
Manis pentadactyla	HLmanPen2	GCA_014570555.1	**Tolypeutes matacus**	HLtolMat1	GCA_004025125.1
Manis pentadactyla	manPen1	GCA_000738955.1	**Tonatia saurophila**	HLtonSau1	GCA_004024845.1
Manis tricuspis	HLmanTri1	GCA_004765945.1	**Trachypithecus francoisi**	HLtraFra1	GCF_009764315.1
Marmota flaviventris	HLmarFla1	GCA_003676075.2	**Tragelaphus imberbis**	HLtraImb1	GCA_006410775.1
Marmota himalayana	HLmarHim1	GCA_005280165.1	**Tragelaphus scriptus**	HLtraScr1	GCA_006410495.1
Marmota marmota marmota	HLmarMar1	GCF_001458135.1	**Tragelaphus strepsiceros**	HLtraStr1	GCA_006410795.1
Marmota monax	HLmarMon1	GCA_901343595.1	**Tragulus javanicus**	HLtraJav1	GCA_004024965.2
Marmota monax	HLmarMon2	GCA_014533835.1	**Tragulus kanchil**	HLtraKan1	GCA_006408655.1
Marmota vancouverensis	HLmarVan1	GCA_005458795.1	**Trichechus manatus latirostris**	triMan1	GCF_000243295.1

Martes zibellina	HLmarZib1	GCA_012583365.1	**Trichosurus vulpecula**	HLtriVul1	GCA_011100635.1
Mastomys coucha	HLmasCou1	GCF_008632895.1	**Tupaia belangeri**	tupBel1	GCA_000181375.1
Megaderma lyra	HLmegLyr2	GCA_004026885.1	**Tupaia chinensis**	tupChi1	GCF_000334495.1
Megaptera novaeangliae	HLmegNov1	GCA_004329385.1	**Tursiops aduncus**	HLturAdu1	GCA_003227395.1
Mellivora capensis	HLmelCap1	GCA_004024625.1	**Tursiops aduncus**	HLturAdu2	None
Meriones unguiculatus	HLmerUng1	GCF_002204375.1	**Tursiops truncatus**	HLturTru3	GCF_001922835.1
Mesocricetus auratus	mesAur1	GCF_000349665.1	**Tursiops truncatus**	HLturTru4	GCF_011762595.1
Mesoplodon bidens	HLmesBid1	GCA_004027085.1	**Tursiops truncatus**	turTru2	GCF_000151865.1
Microcebus murinus	micMur3	GCF_000165445.2	**Urocitellus parryii**	HLuroPar1	GCF_003426925.1
Microcebus sp. 3 GT-2019	HLmicSpe31	GCA_008750915.1	**Uropsilus gracilis**	HLuroGra1	GCA_004024945.1
Microcebus tavaratra	HLmicTav1	GCA_008750935.1	**Urotrichus talpoides**	HLuroTal1	Private
Microgale talazaci	HLmicTal1	GCA_004026705.1	**Ursus americanus**	HLursAme1	GCA_003344425.1
Micronycteris hirsuta	HLmicHir1	GCA_004026765.1	**Ursus americanus**	HLursAme2	None
Microtus agrestis	HLmicAgr2	GCA_902806775.1	**Ursus arctos horribilis**	HLursArc1	GCF_003584765.1
Microtus arvalis	HLmicArv1	GCA_007455615.1	**Ursus maritimus**	ursMar1	GCF_000687225.1
Microtus fortis	HLmicFor1	GCA_014885135.1	**Ursus thibetanus thibetanus**	HLursThi1	GCA_009660055.1
Microtus ochrogaster	micOch1	GCF_000317375.1	**Vicugna pacos**	vicPac2	GCF_000164845.1
Microtus oeconomus	HLmicOec1	GCA_007455595.1	**Vicugna pacos huacaya**	HLvicPacHua3	GCA_000767525.1
Miniopterus natalensis	HLminNat1	GCF_001595765.1	**Vicugna vicugna mensalis**	HLvicVicMen1	GCA_013265495.1
Miniopterus schreibersii	HLminSch1	GCA_004026525.1	**Vombatus ursinus**	HLvomUrs1	GCF_900497805.2
Mirounga angustirostris	HLmirAng2	None	**Vulpes lagopus**	HLvulLag1	GCA_004023825.1
Mirounga leonina	HLmirLeo1	GCF_011800145.1	**Vulpes vulpes**	HLvulVul1	GCF_003160815.1
Mirza coquereli	HLmirCoq1	GCA_004024645.1	**Vulpes zerda**	HLvulZer1	Private
Mirza zaza	HLmirZaz1	GCA_008750895.1	**Xerus inauris**	HLxerIna1	GCA_004024805.1
Mogera wogura	HLmogWog1	Private	**Zalophus californianus**	HLzalCal1	GCA_009762305.1
Molossus molossus	HLmolMol2	None	**Zapus hudsonius**	HLzapHud1	GCA_004024765.1
Monodelphis domestica	monDom5	GCF_000002295.2	**Ziphius cavirostris**	HLzipCav1	GCA_004364475.1
Monodon monoceros	HLmonMon1	GCF_005190385.1			

References

Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, Fernandez Banet J, Billis K, García Girón C, Hourlier T, Howe K, Kähäri A, Kokocinski F, Martin FJ, Murphy DN, Nag R, Ruffier M, Schuster M, Tang YA, Vogel J-H, White S, Zadissa A, Flicek P, Searle SMJ. 2016. The Ensembl gene annotation system. Database, 2016:baw093 DOI: 10.1093/database/baw093.

Alioto TS. 2007. U12DB: a database of orthologous U12-type spliceosomal introns. Nucleic Acids Research, 35(Database):D110–D115 DOI: 10.1093/nar/gkl796.

Altenhoff AM, Glover NM, Dessimoz C. 2019. Inferring Orthology and Paralogy. In: Anisimova M (ed) Evolutionary Genomics: Statistical and Computational Methods. Springer New York, New York, NY, pp. 149–175 DOI: 10.1007/978-1-4939-9074-0_5.

Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C. 2012. Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs. In: Eisen JA (ed) PLoS Comput Biol, 8(5):e1002514 DOI: 10.1371/journal.pcbi.1002514.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol, 215(3):403–410 DOI: 10.1016/S0022-2836(05)80360-2.

Aminzadeh M, Kurfess TR. 2019. Online quality inspection using Bayesian classification in powder-bed additive manufacturing from high-resolution visual camera images. J Intell Manuf, 30(6):2505–2523 DOI: 10.1007/s10845-018-1412-0.

Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, Fang Q, Xie D, Feng S, Stiller J, Genereux D, Johnson J, Marinescu VD, Alföldi J, Harris RS, Lindblad-Toh K, Haussler D, Karlsson E, Jarvis ED, Zhang G, Paten B. 2020. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature, 587(7833):246–251 DOI: 10.1038/s41586-020-2871-y.

Baralle FE, Giudice J. 2017. Alternative splicing as a regulator of development and tissue identity. Nat Rev Mol Cell Biol, 18(7):437–451 DOI: 10.1038/nrm.2017.27.

Behe MJ. 2010. Experimental evolution, loss-of-function mutations, and 'the first rule of adaptive evolution'. Q Rev Biol, 85(4):419–445 DOI: 10.1086/656902.

Blanchard JL, Lynch M. 2000. Organellar genes. Trends in Genetics, 16(7):315–320 DOI: 10.1016/S0168-9525(00)02053-9.

Brandes N, Linial N, Linial M. 2020. PWAS: proteome-wide association study—linking genes and phenotypes by functional variation in proteins. Genome Biol, 21(1):173 DOI: 10.1186/s13059-020-02089-x.

Braun EL, Kimball RT. 2002. Examining Basal avian divergences with mitochondrial sequences: model complexity, taxon sampling, and sequence length. Syst Biol, 51(4):614–625 DOI: 10.1080/10635150290102294.

Brawand D, Wahli W, Kaessmann H. 2008. Loss of egg yolk genes in mammals and the origin of lactation and placentation. PLoS Biol, 6(3):e63 DOI: 10.1371/journal.pbio.0060063.

Brent MR. 2005. Genome annotation past, present, and future: How to define an ORF at each locus. Genome Research, 15(12):1777–1786 DOI: 10.1101/gr.3866105.

Bryzghalov O, Szcześniak MW, Makałowska I. 2019. SyntDB: defining orthologues of human long noncoding RNAs across primates. Nucleic Acids Research:gkz941 DOI: 10.1093/nar/gkz941.

Burge C. 1997. Identification of genes in human genomic DNA. Stanford University, Stanford, CA 94305, Ph.D. Thesis URL: https://www.eecis.udel.edu/~shatkay/Course/papers/burgeThesis.pdf.

Burge CB, Padgett RA, Sharp PA. 1998. Evolutionary Fates and Origins of U12-Type Introns. Molecular Cell, 2(6):773–785 DOI: 10.1016/S1097-2765(00)80292-0.

Capowski EE, Esnault S, Bhattacharya S, Malter JS. 2001. Y box-binding factor promotes

eosinophil survival by stabilizing granulocyte-macrophage colony-stimulating factor mRNA. J Immunol, 167(10):5970–5976 DOI: 10.4049/jimmunol.167.10.5970.
Carlton VEH, Harris BZ, Puffenberger EG, Batta AK, Knisely AS, Robinson DL, Strauss KA, Shneider BL, Lim WA, Salen G, Morton DH, Bull LN. 2003. Complex inheritance of familial hypercholanemia with associated mutations in TJP2 and BAAT. Nat Genet, 34(1):91–96 DOI: 10.1038/ng1147.
Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, Brar GA, Weissman JS, Regev A, Thierry-Mieg N, Cusick ME, Vidal M. 2012. Proto-genes and de novo gene birth. Nature, 487(7407):370–374 DOI: 10.1038/nature11184.
Cassandri M, Smirnov A, Novelli F, Pitolli C, Agostini M, Malewicz M, Melino G, Raschellà G. 2017. Zinc-finger proteins in health and disease. Cell Death Discov, 3(1):17071 DOI: 10.1038/cddiscovery.2017.71.
Charlesworth B, Campos JL, Jackson BC. 2018. Faster-X evolution: Theory and evidence from *Drosophila*. Mol Ecol, 27(19):3753–3771 DOI: 10.1111/mec.14534.
Chattopadhyay R, Das S, Maiti AK, Boldogh I, Xie J, Hazra TK, Kohno K, Mitra S, Bhakat KK. 2008. Regulatory role of human AP-endonuclease (APE1/Ref-1) in YB-1-mediated activation of the multidrug resistance gene MDR1. Mol Cell Biol, 28(23):7066–7080 DOI: 10.1128/MCB.00244-08.
Chen CY, Gherzi R, Andersen JS, Gaietta G, Jürchott K, Royer HD, Mann M, Karin M. 2000. Nucleolin and YB-1 are required for JNK-mediated interleukin-2 mRNA stabilization during T-cell activation. Genes Dev, 14(10):1236–1248.
Chen N-C, Solomon B, Mun T, Iyer S, Langmead B. 2021. Reference flow: reducing reference bias using multiple population genomes. Genome Biology, 22(1):8 DOI: 10.1186/s13059-020-02229-3.
Chen T, Guestrin C. 2016. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, San Francisco California USA, pp. 785–794 DOI: 10.1145/2939672.2939785.
Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. 2012. Predicting the Functional Effect of Amino Acid Substitutions and Indels. In: de Brevern AG (ed) PLoS ONE, 7(10):e46688 DOI: 10.1371/journal.pone.0046688.
Clark DP, Pazdernik NJ, McGehee MR. 2019. Molecular Evolution. In: Molecular Biology. Elsevier, pp. 925–969 DOI: 10.1016/B978-0-12-813288-3.00029-X.
Collins TM, Fedrigo O, Naylor GJP. 2005. Choosing the best genes for the job: the case for stationary genes in genome-scale phylogenetics. Syst Biol, 54(3):493–500 DOI: 10.1080/10635150590947339.
Cortes C, Vapnik V. 1995. Support-vector networks. Mach Learn, 20(3):273–297 DOI: 10.1007/BF00994018.
Curwen V. 2004. The Ensembl Automatic Gene Annotation System. Genome Research, 14(5):942–950 DOI: 10.1101/gr.1858004.
Dada EG, Bassi JS, Chiroma H, Abdulhamid SM, Adetunmbi AO, Ajibuwa OE. 2019. Machine learning for email spam filtering: review, approaches and open research problems. Heliyon, 5(6):e01802 DOI: 10.1016/j.heliyon.2019.e01802.
Di Franco A, Poujol R, Baurain D, Philippe H. 2019. Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. BMC Evol Biol, 19(1):21 DOI: 10.1186/s12862-019-1350-2.
Doronina L, Churakov G, Kuritzin A, Shi J, Baertsch R, Clawson H, Schmitz J. 2017. Speciation network in Laurasiatheria: retrophylogenomic signals. Genome Res, 27(6):997–1003 DOI: 10.1101/gr.210948.116.
Ejigu GF, Jung J. 2020. Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing. Biology, 9(9):295 DOI: 10.3390/biology9090295.
Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome

comparisons dramatically improves orthogroup inference accuracy. Genome Biol, 16(1):157 DOI: 10.1186/s13059-015-0721-2.
Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639):115–118 DOI: 10.1038/nature21056.
Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst Biol, 61(5):717–726 DOI: 10.1093/sysbio/sys004.
Felsenstein J. 1978. Cases in which Parsimony or Compatibility Methods will be Positively Misleading. Systematic Biology, 27(4):401–410 DOI: 10.1093/sysbio/27.4.401.
Fitch WM. 1970. Distinguishing Homologous from Analogous Proteins. Systematic Zoology, 19(2):99 DOI: 10.2307/2412448.
Fitch WM. 2000. Homology. Trends in Genetics, 16(5):227–231 DOI: 10.1016/S0168-9525(00)02005-9.
Foley NM, Springer MS, Teeling EC. 2016. Mammal madness: is the mammal tree of life not yet resolved? Phil Trans R Soc B, 371(1699):20150140 DOI: 10.1098/rstb.2015.0140.
Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland JE, Mudge JM, Sisu C, Wright JC, Armstrong J, Barnes I, Berry A, Bignell A, Boix C, Carbonell Sala S, Cunningham F, Di Domenico T, Donaldson S, Fiddes IT, García Girón C, Gonzalez JM, Grego T, Hardy M, Hourlier T, Howe KL, Hunt T, Izuogu OG, Johnson R, Martin FJ, Martínez L, Mohanan S, Muir P, Navarro FCP, Parker A, Pei B, Pozo F, Riera FC, Ruffier M, Schmitt BM, Stapleton E, Suner M-M, Sycheva I, Uszczynska-Ratajczak B, Wolf MY, Xu J, Yang YT, Yates A, Zerbino D, Zhang Y, Choudhary JS, Gerstein M, Guigó R, Hubbard TJP, Kellis M, Paten B, Tress ML, Flicek P. 2021. GENCODE 2021. Nucleic Acids Research, 49(D1):D916–D923 DOI: 10.1093/nar/gkaa1087.
Gabaldón T. 2008. Large-scale assignment of orthology: back to phylogenetics? Genome Biol, 9(10):235 DOI: 10.1186/gb-2008-9-10-235.
Gabaldón T, Koonin EV. 2013. Functional and evolutionary implications of gene orthology. Nat Rev Genet, 14(5):360–366 DOI: 10.1038/nrg3456.
Gaudreault I, Guay D, Lebel M. 2004. YB-1 promotes strand separation in vitro of duplex DNA containing either mispaired bases or cisplatin modifications, exhibits endonucleolytic activities and binds several DNA repair proteins. Nucleic Acids Res, 32(1):316–327 DOI: 10.1093/nar/gkh170.
Gawryluk RMR, Eme L, Roger AJ. 2015. Gene fusion, fission, lateral transfer, and loss: Not-so-rare events in the evolution of eukaryotic ATP citrate lyase. Molecular Phylogenetics and Evolution, 91:12–16 DOI: 10.1016/j.ympev.2015.05.010.
Gremme G, Brendel V, Sparks ME, Kurtz S. 2005. Engineering a software tool for gene structure prediction in higher organisms. Information and Software Technology, 47(15):965–978 DOI: 10.1016/j.infsof.2005.09.005.
Gyles C, Boerlin P. 2014. Horizontally transferred genetic elements and their role in pathogenesis of bacterial disease. Vet Pathol, 51(2):328–340 DOI: 10.1177/0300985813511131.
Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol, 9(1):R7 DOI: 10.1186/gb-2008-9-1-r7.
Hahn Y, Jeong S, Lee B. 2007. Inactivation of MOXD2 and S100A15A by Exon Deletion during Human Evolution. Molecular Biology and Evolution, 24(10):2203–2212 DOI: 10.1093/molbev/msm146.
Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E, Taipale J. 2006. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. Cell, 124(1):47–59 DOI: 10.1016/j.cell.2005.10.042.
Harris RS. 2007. Improved pairwise alignment of genomic DNA. The Pennsylvania State University, Ph.D. Thesis URL:

http://www.bx.psu.edu/~rsharris/rsharris_phd_thesis_2007.pdf.
Hecker N, Hiller M. 2020. A genome alignment of 120 mammals highlights ultraconserved element variability and placenta-associated enhancers. GigaScience, 9(1):giz159 DOI: 10.1093/gigascience/giz159.
van der Heijden RT, Snel B, van Noort V, Huynen MA. 2007. Orthology prediction at scalable resolution by phylogenetic tree analysis. BMC Bioinformatics, 8(1):83 DOI: 10.1186/1471-2105-8-83.
Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, Vilella AJ, Searle SMJ, Amode R, Brent S, Spooner W, Kulesha E, Yates A, Flicek P. 2016. Ensembl comparative genomics resources. Database, 2016:bav096 DOI: 10.1093/database/bav096.
Hiller M, Agarwal S, Notwell JH, Parikh R, Guturu H, Wenger AM, Bejerano G. 2013. Computational methods to detect conserved non-genic elements in phylogenetically isolated genomes: application to zebrafish. Nucleic Acids Research, 41(15):e151–e151 DOI: 10.1093/nar/gkt557.
Hiller M, Schaar BT, Indjeian VB, Kingsley DM, Hagey LR, Bejerano G. 2012. A “Forward Genomics” Approach Links Genotype to Phenotype using Independent Phenotypic Losses among Related Species. Cell Reports, 2(4):817–823 DOI: 10.1016/j.celrep.2012.08.032.
Hu W, Liu X, Huang Y, Wang Y, Zhang M, Zhao H. 2020. Structured Data Encoder for Neural Networks Based on Gradient Boosting Decision Tree. In: Qiu M (ed) Algorithms and Architectures for Parallel Processing. Springer International Publishing, Cham (Lecture Notes in Computer Science), pp. 603–618 DOI: 10.1007/978-3-030-60239-0_41.
Huang Y, Chen S-Y, Deng F. 2016. Well-characterized sequence features of eukaryote genomes and implications for ab initio gene prediction. Comput Struct Biotechnol J, 14:298–303 DOI: 10.1016/j.csbj.2016.07.002.
Huelsmann M, Hecker N, Springer MS, Gatesy J, Sharma V, Hiller M. 2019. Genes lost during the transition from land to water in cetaceans highlight genomic changes associated with aquatic adaptations. Sci Adv, 5(9):eaaw6671 DOI: 10.1126/sciadv.aaw6671.
Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldón T. 2007. The human phylome. Genome Biol, 8(6):R109 DOI: 10.1186/gb-2007-8-6-r109.
Hughes DP, Crispe IN. 1995. A naturally occurring soluble isoform of murine Fas generated by alternative splicing. J Exp Med, 182(5):1395–1401 DOI: 10.1084/jem.182.5.1395.
Hughes GM, Boston ESM, Finarelli JA, Murphy WJ, Higgins DG, Teeling EC. 2018. The Birth and Death of Olfactory Receptor Gene Families in Mammalian Niche Adaptation. In: Satta Y (ed) Molecular Biology and Evolution, 35(6):1390–1406 DOI: 10.1093/molbev/msy028.
International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature, 409(6822):860–921 DOI: 10.1038/35057062.
Jalal ASB, Tran NT, Stevenson CE, Chan EW, Lo R, Tan X, Noy A, Lawson DM, Le TBK. 2020. Diversification of DNA-Binding Specificity by Permissive and Specificity-Switching Mutations in the ParB/Noc Protein Family. Cell Reports, 32(3):107928 DOI: 10.1016/j.celrep.2020.107928.
Jebb D, Hiller M. 2018. Recurrent loss of HMGCS2 shows that ketogenesis is not essential for the evolution of large mammalian brains. eLife, 7:e38906 DOI: 10.7554/eLife.38906.
Jebb D, Huang Z, Pippel M, Hughes GM, Lavrichenko K, Devanna P, Winkler S, Jermiin LS, Skirmuntt EC, Katzourakis A, Burkitt-Gray L, Ray DA, Sullivan KAM, Roscito JG, Kirilenko BM, Dávalos LM, Corthals AP, Power ML, Jones G, Ransome RD, Dechmann DKN, Locatelli AG, Puechmaille SJ, Fedrigo O, Jarvis ED, Hiller M, Vernes SC, Myers EW, Teeling EC. 2020. Six reference-quality genomes reveal evolution of bat adaptations. Nature, 583(7817):578–584 DOI: 10.1038/s41586-020-2486-3.

Kabza M, Ciomborowska J, Makałowska I. 2014. RetrogeneDB--a database of animal retrogenes. Mol Biol Evol, 31(7):1646–1648 DOI: 10.1093/molbev/msu139.
Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. Genome Research, 20(10):1313–1326 DOI: 10.1101/gr.101386.109.
Kay E, Vogel TM, Bertolla F, Nalin R, Simonet P. 2002. In situ transfer of antibiotic resistance genes from transgenic (transplastomic) tobacco plants to bacteria. Appl Environ Microbiol, 68(7):3345–3351 DOI: 10.1128/aem.68.7.3345-3351.2002.
Keeling PJ, Palmer JD. 2008. Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet, 9(8):605–618 DOI: 10.1038/nrg2386.
Kent WJ. 2002. BLAT---The BLAST-Like Alignment Tool. Genome Research, 12(4):656–664 DOI: 10.1101/gr.229202.
Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. 2003. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. Proceedings of the National Academy of Sciences, 100(20):11484–11489 DOI: 10.1073/pnas.1932072100.
Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler a. D. 2002. The Human Genome Browser at UCSC. Genome Research, 12(6):996–1006 DOI: 10.1101/gr.229102.
Khan I, Maldonado E, Vasconcelos V, O'Brien SJ, Johnson WE, Antunes A. 2014. Mammalian keratin associated proteins (KRTAPs) subgenomes: disentangling hair diversity and adaptation to terrestrial and aquatic environments. BMC Genomics, 15(1):779 DOI: 10.1186/1471-2164-15-779.
Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. 2011. Adaptive seeds tame genomic sequence comparison. Genome Research, 21(3):487–493 DOI: 10.1101/gr.113985.110.
Kirilenko BM, Hagey LR, Barnes S, Falany CN, Hiller M. 2019. Evolutionary Analysis of Bile Acid-Conjugating Enzymes Reveals a Complex Duplication and Reciprocal Loss History. In: Das S (ed) Genome Biology and Evolution, 11(11):3256–3268 DOI: 10.1093/gbe/evz238.
Klug A. 2010. The Discovery of Zinc Fingers and Their Applications in Gene Regulation and Genome Manipulation. Annu Rev Biochem, 79(1):213–231 DOI: 10.1146/annurev-biochem-010909-095056.
Kohl S, Baumann B, Rosenberg T, Kellner U, Lorenz B, Vadalà M, Jacobson SG, Wissinger B. 2002. Mutations in the cone photoreceptor G-protein alpha-subunit gene GNAT2 in patients with achromatopsia. Am J Hum Genet, 71(2):422–425 DOI: 10.1086/341835.
Koonin EV. 2005. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet, 39:309–338 DOI: 10.1146/annurev.genet.39.073003.114725.
Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics, 5(1):59 DOI: 10.1186/1471-2105-5-59.
Korf I, Flicek P, Duan D, Brent MR. 2001. Integrating genomic homology into gene structure prediction. Bioinformatics, 17 Suppl 1:S140-148 DOI: 10.1093/bioinformatics/17.suppl_1.s140.
Krizhevsky A, Sutskever I, Hinton GE. 2017. ImageNet classification with deep convolutional neural networks. Commun ACM, 60(6):84–90 DOI: 10.1145/3065386.
Kuo RI, Cheng Y, Zhang R, Brown JWS, Smith J, Archibald AL, Burt DW. 2020. Illuminating the dark side of the human transcriptome with long read transcript sequencing. BMC Genomics, 21(1):751 DOI: 10.1186/s12864-020-07123-7.
Kurian L, Palanimurugan R, Gödderz D, Dohmen RJ. 2011. Polyamine sensing by nascent ornithine decarboxylase antizyme stimulates decoding of its mRNA. Nature, 477(7365):490–494 DOI: 10.1038/nature10393.
Lane RP, Cutforth T, Young J, Athanasiou M, Friedman C, Rowen L, Evans G, Axel R, Hood L, Trask BJ. 2001. Genomic analysis of orthologous mouse and human olfactory receptor loci. Proceedings of the National Academy of Sciences, 98(13):7390–7395 DOI: 10.1073/pnas.131215398.
Lange S, Xiang F, Yakovenko A, Vihola A, Hackman P, Rostkova E, Kristensen J,

Brandmeier B, Franzen G, Hedberg B, Gunnarsson LG, Hughes SM, Marchand S, Sejersen T, Richard I, Edström L, Ehler E, Udd B, Gautel M. 2005. The kinase domain of titin controls muscle gene expression and protein turnover. Science, 308(5728):1599–1603 DOI: 10.1126/science.1110463.
Lee J-H, Lewis KM, Moural TW, Kirilenko B, Borgonovo B, Prange G, Koessl M, Huggenberger S, Kang C, Hiller M. 2018. Molecular parallelism in fast-twitch muscle proteins in echolocating mammals. Sci Adv, 4(9):eaat9660 DOI: 10.1126/sciadv.aat9660.
Leonard G, Richards TA. 2012. Genome-scale comparative analysis of gene fusions, gene fissions, and the fungal tree of life. Proceedings of the National Academy of Sciences, 109(52):21402–21407 DOI: 10.1073/pnas.1210909110.
Levine A, Durbin R. 2001. A computational scan for U12-dependent introns in the human genome sequence. Nucleic Acids Research, 29(19):4006–4013 DOI: 10.1093/nar/29.19.4006.
Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, Durbin R, Edwards SV, Forest F, Gilbert MTP, Goldstein MM, Grigoriev IV, Hackett KJ, Haussler D, Jarvis ED, Johnson WE, Patrinos A, Richards S, Castilla-Rubio JC, van Sluys M-A, Soltis PS, Xu X, Yang H, Zhang G. 2018. Earth BioGenome Project: Sequencing life for the future of life. Proc Natl Acad Sci USA, 115(17):4325–4333 DOI: 10.1073/pnas.1720115115.
Li L. 2003. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Research, 13(9):2178–2189 DOI: 10.1101/gr.1224503.
Li Y, Liu Z, Shi P, Zhang J. 2010. The hearing gene Prestin unites echolocating bats and whales. Current Biology, 20(2):R55–R56 DOI: 10.1016/j.cub.2009.11.042.
Liu Z, Qi F-Y, Zhou X, Ren H-Q, Shi P. 2014. Parallel Sites Implicate Functional Convergence of the Hearing Gene Prestin among Echolocating Mammals. Molecular Biology and Evolution, 31(9):2415–2424 DOI: 10.1093/molbev/msu194.
Löytynoja A. 2014. Phylogeny-aware alignment with PRANK. Methods Mol Biol, 1079:155–170 DOI: 10.1007/978-1-62703-646-7_10.
Ma B, Tromp J, Li M. 2002. PatternHunter: faster and more sensitive homology search. Bioinformatics, 18(3):440–445 DOI: 10.1093/bioinformatics/18.3.440.
MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, Albers CA, Zhang ZD, Conrad DF, Lunter G, Zheng H, Ayub Q, DePristo MA, Banks E, Hu M, Handsaker RE, Rosenfeld JA, Fromer M, Jin M, Mu XJ, Khurana E, Ye K, Kay M, Saunders GI, Suner M-M, Hunt T, Barnes IHA, Amid C, Carvalho-Silva DR, Bignell AH, Snow C, Yngvadottir B, Bumpstead S, Cooper DN, Xue Y, Romero IG, 1000 Genomes Project Consortium, Wang J, Li Y, Gibbs RA, McCarroll SA, Dermitzakis ET, Pritchard JK, Barrett JC, Harrow J, Hurles ME, Gerstein MB, Tyler-Smith C. 2012. A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. Science, 335(6070):823–828 DOI: 10.1126/science.1215040.
Malik HS, Henikoff S. 2003. Phylogenomics of the nucleosome. Nat Struct Biol, 10(11):882–891 DOI: 10.1038/nsb996.
Mason L, Baxter J, Bartlett P, Frean M. Boosting Algorithms as Gradient Descent. In: Proceedings of the 12th International Conference on Neural Information Processing Systems. MIT Press, Denver, CO (NIPS'99), pp. 512–518.
Matsui A, Go Y, Niimura Y. 2010. Degeneration of Olfactory Receptor Gene Repertories in Primates: No Direct Link to Full Trichromatic Vision. Molecular Biology and Evolution, 27(5):1192–1200 DOI: 10.1093/molbev/msq003.
Meredith RW, Zhang G, Gilbert MTP, Jarvis ED, Springer MS. 2014. Evidence for a single loss of mineralized teeth in the common avian ancestor. Science, 346(6215):1254390 DOI: 10.1126/science.1254390.
Minoche AE, Dohm JC, Schneider J, Holtgräwe D, Viehöver P, Montfort M, Rosleff Sörensen T, Weisshaar B, Himmelbauer H. 2015. Exploiting single-molecule transcript sequencing for eukaryotic gene prediction. Genome Biol, 16(1):184 DOI:

10.1186/s13059-015-0729-7.
Mower JP, Stefanović S, Hao W, Gummow JS, Jain K, Ahmed D, Palmer JD. 2010. Horizontal acquisition of multiple mitochondrial genes from a parasitic plant followed by gene conversion with host mitochondrial genes. BMC Biol, 8(1):150 DOI: 10.1186/1741-7007-8-150.
Nachtweide S, Stanke M. 2019. Multi-Genome Annotation with AUGUSTUS. In: Kollmar M (ed) Gene Prediction. Springer New York, New York, NY (Methods in Molecular Biology), pp. 139–160 DOI: 10.1007/978-1-4939-9173-0_8.
Nei M, Xu P, Glazko G. 2001. Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. Proceedings of the National Academy of Sciences, 98(5):2497–2502 DOI: 10.1073/pnas.051611498.
Neme R, Tautz D. 2014. Evolution: Dynamics of De Novo Gene Emergence. Current Biology, 24(6):R238–R240 DOI: 10.1016/j.cub.2014.02.016.
Nevers Y, Defosset A, Lecompte O. 2020. Orthology: Promises and Challenges. In: Pontarotti P (ed) Evolutionary Biology—A Transdisciplinary Approach. Springer International Publishing, Cham, pp. 203–228 DOI: 10.1007/978-3-030-57246-4_9.
Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution, 32(1):268–274 DOI: 10.1093/molbev/msu300.
Niimura Y, Matsui A, Touhara K. 2014. Extreme expansion of the olfactory receptor gene repertoire in African elephants and evolutionary dynamics of orthologous gene groups in 13 placental mammals. Genome Res, 24(9):1485–1496 DOI: 10.1101/gr.169532.113.
Opitz D, Maclin R. 1999. Popular Ensemble Methods: An Empirical Study. jair, 11:169–198 DOI: 10.1613/jair.614.
Osipova E, Hecker N, Hiller M. 2019. RepeatFiller newly identifies megabases of aligning repetitive sequences and improves annotations of conserved non-exonic elements. GigaScience, 8(11):giz132 DOI: 10.1093/gigascience/giz132.
Parra G, Agarwal P, Abril JF, Wiehe T, Fickett JW, Guigó R. 2003. Comparative gene prediction in human and mouse. Genome Res, 13(1):108–117 DOI: 10.1101/gr.871403.
Pasek S, Risler J-L, Brezellec P. 2006. Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. Bioinformatics, 22(12):1418–1423 DOI: 10.1093/bioinformatics/btl135.
Patel AA, Steitz JA. 2003. Splicing double: insights from the second spliceosome. Nat Rev Mol Cell Biol, 4(12):960–970 DOI: 10.1038/nrm1259.
Paulding CA, Ruvolo M, Haber DA. 2003. The Tre2(USP6) oncogene is a hominoid-specific gene. Proc Natl Acad Sci USA, 100(5):2507 DOI: 10.1073/pnas.0437015100.
Pearson WR. 2013. An introduction to sequence similarity ('homology') searching. Curr Protoc Bioinformatics, Chapter 3:Unit3.1 DOI: 10.1002/0471250953.bi0301s42.
Peterson ME, Chen F, Saven JG, Roos DS, Babbitt PC, Sali A. 2009. Evolutionary constraints on structural similarity in orthologs and paralogs. Protein Sci, 18(6):1306–1315 DOI: 10.1002/pro.143.
Pizarro D, Divakar PK, Grewe F, Crespo A, Dal Grande F, Lumbsch HT. 2020. Genome-Wide Analysis of Biosynthetic Gene Cluster Reveals Correlated Gene Loss with Absence of Usnic Acid in Lichen-Forming Fungi. In: Jason S (ed) Genome Biology and Evolution, 12(10):1858–1868 DOI: 10.1093/gbe/evaa189.
Polikar R. 2006. Ensemble based systems in decision making. IEEE Circuits Syst Mag, 6(3):21–45 DOI: 10.1109/MCAS.2006.1688199.
Ranwez V, Douzery EJP, Cambon C, Chantret N, Delsuc F. 2018. MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. In: Wilke C (ed) Molecular Biology and Evolution, 35(10):2582–2584 DOI: 10.1093/molbev/msy159.
Ranwez V, Harispe S, Delsuc F, Douzery EJP. 2011. MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons. In: Murphy WJ (ed) PLoS

ONE, 6(9):e22594 DOI: 10.1371/journal.pone.0022594.
Reimand J, Kull M, Peterson H, Hansen J, Vilo J. 2007. g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Research, 35(suppl_2):W193–W200 DOI: 10.1093/nar/gkm226.
Rodriguez JM, Maietta P, Ezkurdia I, Pietrelli A, Wesselink J-J, Lopez G, Valencia A, Tress ML. 2013. APPRIS: annotation of principal and alternative splice isoforms. Nucleic Acids Research, 41(D1):D110–D117 DOI: 10.1093/nar/gks1058.
Rosati M, Marino M, Franzé A, Tramontano A, Girmaldi G. 1991. Members of the zinc finger protein gene family sharing a conserved N-terminal module. Nucl Acids Res, 19(20):5661–5667 DOI: 10.1093/nar/19.20.5661.
Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, Rowan TN, Low WY, Zimin A, Couldrey C, Hall R, Li W, Rhie A, Ghurye J, McKay SD, Thibaud-Nissen F, Hoffman J, Murdoch BM, Snelling WM, McDaneld TG, Hammond JA, Schwartz JC, Nandolo W, Hagen DE, Dreischer C, Schultheiss SJ, Schroeder SG, Phillippy AM, Cole JB, Van Tassell CP, Liu G, Smith TPL, Medrano JF. 2020. De novo assembly of the cattle reference genome with single-molecule sequencing. GigaScience, 9(3):giaa021 DOI: 10.1093/gigascience/giaa021.
Roy B, Haupt LM, Griffiths LR. 2013. Review: Alternative Splicing (AS) of Genes As An Approach for Generating Protein Complexity. Curr Genomics, 14(3):182–194 DOI: 10.2174/1389202911314030004.
Salzberg SL. 2019. Next-generation genome annotation: we still struggle to get it right. Genome Biol, 20(1):92 DOI: 10.1186/s13059-019-1715-2.
Schwartz S. 2003. Human-Mouse Alignments with BLASTZ. Genome Research, 13(1):103–107 DOI: 10.1101/gr.809403.
Sharma V, Elghafari A, Hiller M. 2016. Coding exon-structure aware realigner (CESAR) utilizes genome alignments for accurate comparative gene annotation. Nucleic Acids Res, 44(11):e103–e103 DOI: 10.1093/nar/gkw210.
Sharma V, Hecker N, Roscito JG, Foerster L, Langer BE, Hiller M. 2018. A genomics approach reveals insights into the importance of gene losses for mammalian adaptations. Nat Commun, 9(1):1215 DOI: 10.1038/s41467-018-03667-1.
Sharma V, Hecker N, Walther F, Stuckas H, Hiller M. 2020. Convergent Losses of TLR5 Suggest Altered Extracellular Flagellin Detection in Four Mammalian Lineages. In: Yeager M (ed) Molecular Biology and Evolution, 37(7):1847–1854 DOI: 10.1093/molbev/msaa058.
Sharma V, Hiller M. 2019. Coding Exon-Structure Aware Realigner (CESAR): Utilizing Genome Alignments for Comparative Gene Annotation. In: Kollmar M (ed) Gene Prediction. Springer New York, New York, NY (Methods in Molecular Biology), pp. 179–191 DOI: 10.1007/978-1-4939-9173-0_10.
Sharma V, Schwede P, Hiller M. 2017. CESAR 2.0 substantially improves speed and accuracy of comparative gene annotation. In: Kelso J (ed) Bioinformatics, 33(24):3985–3987 DOI: 10.1093/bioinformatics/btx527.
Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL. 2015. Less Is More: An Adaptive Branch-Site Random Effects Model for Efficient Detection of Episodic Diversifying Selection. Molecular Biology and Evolution, 32(5):1342–1353 DOI: 10.1093/molbev/msv022.
Smooker PM, Whisstock JC, Irving JA, Siyaguna S, Spithill TW, Pike RN. 2000. For the record: A single amino acid substitution affects substrate specificity in cysteine proteinases from Fasciola hepatica. Protein Sci, 9(12):2567–2572 DOI: 10.1110/ps.9.12.2567.
Soltis DE, Soltis PS. 2003. The Role of Phylogenetics in Comparative Genetics. Plant Physiol, 132(4):1790–1800 DOI: 10.1104/pp.103.022509.
Sonnhammer ELL, Gabaldon T, Sousa da Silva AW, Martin M, Robinson-Rechavi M, Boeckmann B, Thomas PD, Dessimoz C, the Quest for Orthologs consortium. 2014. Big data and other challenges in the quest for orthologs. Bioinformatics, 30(21):2993–2998 DOI: 10.1093/bioinformatics/btu492.

Springer MS, Emerling CA, Gatesy J, Randall J, Collin MA, Hecker N, Hiller M, Delsuc F. 2019. Odontogenic ameloblast-associated (ODAM) is inactivated in toothless/enamelless placental mammals and toothed whales. BMC Evolutionary Biology, 19(1):31 DOI: 10.1186/s12862-019-1359-6.
Springer MS, Gatesy J. 2019. An ABBA-BABA Test for Introgression Using Retroposon Insertion Data. Evolutionary Biology DOI: 10.1101/709477.
Stigler SM. 1986. The history of statistics: the measurement of uncertainty before 1900. Belknap Press of Harvard University Press, Cambridge, Mass.
Suarez HG, Langer BE, Ladde P, Hiller M. 2017. chainCleaner improves genome alignment specificity and sensitivity. Bioinformatics:btx024 DOI: 10.1093/bioinformatics/btx024.
Sutskever I, Vinyals O, Le QV. 2014. Sequence to Sequence Learning with Neural Networks. arXiv:14093215 [cs] [accessed: 07/04/2021] URL: http://arxiv.org/abs/1409.3215.
Taher L, Rinner O, Garg S, Sczyrba A, Brudno M, Batzoglou S, Morgenstern B. 2003. AGenDA: homology-based gene prediction. Bioinformatics, 19(12):1575–1577 DOI: 10.1093/bioinformatics/btg181.
Tatusov RL. 1997. A Genomic Perspective on Protein Families. Science, 278(5338):631–637 DOI: 10.1126/science.278.5338.631.
Taylor JS, Raes J. 2004. Duplication and Divergence: The Evolution of New Genes and Old Ideas. Annu Rev Genet, 38(1):615–643 DOI: 10.1146/annurev.genet.38.072902.092831.
The Bovine Genome Sequencing and Analysis Consortium. 2009. The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution. Science, 324(5926):522–528 DOI: 10.1126/science.1169588.
Thompson RF, Langford GM. 2002. Myosin superfamily evolutionary history. Anat Rec, 268(3):276–289 DOI: 10.1002/ar.10160.
Tin Kam Ho. 1995. Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition. 3rd International Conference on Document Analysis and Recognition. IEEE Comput. Soc. Press, Montreal, Que., Canada, pp. 278–282 DOI: 10.1109/ICDAR.1995.598994.
Van Oss SB, Carvunis A-R. 2019. De novo gene birth. PLoS Genet, 15(5):e1008160 DOI: 10.1371/journal.pgen.1008160.
Vandamme A-M. 2009. Basic concepts of molecular evolution. The phylogenetic handbook, 2:3–32.
Vicoso B, Charlesworth B. 2006. Evolution on the X chromosome: unusual patterns and processes. Nat Rev Genet, 7(8):645–653 DOI: 10.1038/nrg1914.
Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2008. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Research, 19(2):327–335 DOI: 10.1101/gr.073585.107.
Wang B, Tseng E, Regulski M, Clark TA, Hon T, Jiao Y, Lu Z, Olson A, Stein JC, Ware D. 2016. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. Nat Commun, 7(1):11708 DOI: 10.1038/ncomms11708.
Wang Y, Coleman-Derr D, Chen G, Gu YQ. 2015. OrthoVenn: a web server for genome wide comparison and annotation of orthologous clusters across multiple species. Nucleic Acids Res, 43(W1):W78–W84 DOI: 10.1093/nar/gkv487.
Watson M, Warr A. 2019. Errors in long-read assemblies can critically affect protein prediction. Nat Biotechnol, 37(2):124–126 DOI: 10.1038/s41587-018-0004-z.
Wu D-D, Irwin DM, Zhang Y-P. 2008. Molecular evolution of the keratin associated protein gene family in mammals, role in the evolution of mammalian hair. BMC Evol Biol, 8(1):241 DOI: 10.1186/1471-2148-8-241.
Xia J, Guo Z, Yang Z, Han H, Wang S, Xu H, Yang X, Yang F, Wu Q, Xie W, Zhou X, Dermauw W, Turlings TCJ, Zhang Y. 2021. Whitefly hijacks a plant detoxification gene that neutralizes plant toxins. Cell, 184(7):1693-1705.e17 DOI: 10.1016/j.cell.2021.02.014.
Yang X, Coulombe-Huntington J, Kang S, Sheynkman GM, Hao T, Richardson A, Sun S, Yang F, Shen YA, Murray RR, Spirohn K, Begg BE, Duran-Frigola M, MacWilliams A,

Pevzner SJ, Zhong Q, Trigg SA, Tam S, Ghamsari L, Sahni N, Yi S, Rodriguez MD, Balcha D, Tan G, Costanzo M, Andrews B, Boone C, Zhou XJ, Salehi-Ashtiani K, Charloteaux B, Chen AA, Calderwood MA, Aloy P, Roth FP, Hill DE, Iakoucheva LM, Xia Y, Vidal M. 2016. Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. Cell, 164(4):805–817 DOI: 10.1016/j.cell.2016.01.029.

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution, 24(8):1586–1591 DOI: 10.1093/molbev/msm088.

Yang Z, Rannala B. 2012. Molecular phylogenetics: principles and practice. Nature Reviews Genetics, 13(5):303–314 DOI: 10.1038/nrg3186.

Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J, Billis K, Boddu S, Marugán JC, Cummins C, Davidson C, Dodiya K, Fatima R, Gall A, Giron CG, Gil L, Grego T, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, Kay M, Lavidas I, Le T, Lemos D, Martinez JG, Maurel T, McDowall M, McMahon A, Mohanan S, Moore B, Nuhn M, Oheh DN, Parker A, Parton A, Patricio M, Sakthivel MP, Abdul Salam AI, Schmitt BM, Schuilenburg H, Sheppard D, Sycheva M, Szuba M, Taylor K, Thormann A, Threadgold G, Vullo A, Walts B, Winterbottom A, Zadissa A, Chakiachvili M, Flint B, Frankish A, Hunt SE, IIsley G, Kostadima M, Langridge N, Loveland JE, Martin FJ, Morales J, Mudge JM, Muffato M, Perry E, Ruffier M, Trevanion SJ, Cunningham F, Howe KL, Zerbino DR, Flicek P. 2019. Ensembl 2020. Nucleic Acids Research:gkz966 DOI: 10.1093/nar/gkz966.

Zerbino DR, Frankish A, Flicek P. 2020. Progress, Challenges, and Surprises in Annotating the Human Genome. Annu Rev Genomics Hum Genet, 21:55–79 DOI: 10.1146/annurev-genom-121119-083418.

Zhang J. 2003. Evolution by gene duplication: an update. Trends in Ecology & Evolution, 18(6):292–298 DOI: 10.1016/S0169-5347(03)00033-8.

Zhang ZD, Frankish A, Hunt T, Harrow J, Gerstein M. 2010. Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. Genome Biol, 11(3):R26 DOI: 10.1186/gb-2010-11-3-r26.

Zhou Y, Shearwin-Whyatt L, Li J, Song Z, Hayakawa T, Stevens D, Fenelon JC, Peel E, Cheng Y, Pajpach F, Bradley N, Suzuki H, Nikaido M, Damas J, Daish T, Perry T, Zhu Z, Geng Y, Rhie A, Sims Y, Wood J, Haase B, Mountcastle J, Fedrigo O, Li Q, Yang H, Wang J, Johnston SD, Phillippy AM, Howe K, Jarvis ED, Ryder OA, Kaessmann H, Donnelly P, Korlach J, Lewin HA, Graves J, Belov K, Renfree MB, Grutzner F, Zhou Q, Zhang G. 2021. Platypus and echidna genomes reveal mammalian biology and evolution. Nature DOI: 10.1038/s41586-020-03039-0.

Zmasek CM, Eddy SR. 2001. A simple algorithm to infer gene duplication and speciation events on a gene tree. Bioinformatics, 17(9):821–828 DOI: 10.1093/bioinformatics/17.9.821.

Technische Universität Dresden
Medizinische Fakultät Carl Gustav Carus
Promotionsordnung vom 24. Juli 2011

Erklärungen zur Eröffnung des Promotionsverfahrens

1. Hiermit versichere ich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.
2. Bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskripts habe ich Unterstützungsleistungen von folgenden Personen erhalten:
……..
3. Weitere Personen waren an der geistigen Herstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich nicht die Hilfe eines kommerziellen Promotionsberaters in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.
4. Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.
5. Die Inhalte dieser Dissertation wurden in folgender Form veröffentlicht:
……..
6. Ich bestätige, dass es keine zurückliegenden erfolglosen Promotionsverfahren gab.
7. Ich bestätige, dass ich die Promotionsordnung der Medizinischen Fakultät der Technischen Universität Dresden anerkenne.
8. Ich habe die Zitierrichtlinien für Dissertationen an der Medizinischen Fakultät der Technischen Universität Dresden zur Kenntnis genommen und befolgt.

Ort, Datum

Unterschrift des Doktoranden

(Diese Erklärungen sind an das Ende der Arbeit einzubinden) Formblatt 1.2.1, Seite 1-1, erstellt 18.10.2013