THE UNIVERSITY *of* EDINBURGH

Evolutionary Insights Enabled by Assembly and Annotation of Dog's Y Chromosome

Wengang Zhang

THE UNIVERSITY *of* EDINBURGH

Doctor of Philosophy

College of Medicine and Veterinary Medicine, The University of Edinburgh

2023

Declaration

I declare that the work contained within this thesis is my own with the exception of specific experiments which are clearly indicated in the text. This thesis has been composed by myself and has not been submitted for any other degree or qualification.

Wengang Zhang

Lay Summary

Most mammals use the XX/XY system for determining sex, where males have two distinct types of sex chromosomes (X and Y) while females have two of the same sort (X). The X and Y chromosomes differ from one another in numerous ways including size, organisation, and structure. It is thought that mammalian sex chromosomes descended from a pair of autosomes in the common therian ancestor around 180 million years ago. Mammalian X chromosomes contain many genes; they are largely conserved among mammals in terms of their gene content and organisation. In contrast, Y chromosomes are typically shorter than X chromosomes and they are characterised by their large amounts of repetitive DNA and multiple-copy of gene families. The Y chromosome is known to have biological functions in spermatogenesis, sex determination, and sexual dimorphism, and its defects can cause several diseases. Therefore, defining mammalian Y chromosome biology will facilitate our understanding of male fertility and development, provide a genetic record for male demography, and help elucidate the mechanisms of certain diseases.

The dog is a member of the *Carnivora*. This order represents the fourth main branch of the mammalian tree. Currently, our understanding of the dog's Y chromosomes is limited due to the lack of sequencing data and an imperfect assembly. The objective of my project was to improve our knowledge of the dog Y chromosome, identify the novelties within its sequence, and advance the theory of mammalian Y chromosome evolution. To do so, a male Labrador retriever dog's genome was sequenced. The resulting data was assembled into a 6.6 megabase-long male-specific Y (MSY) region of highly contiguous DNA sequence using a hybrid methodology I developed.

As observed in other species, the dog MSY was characterised by its repetitive nature, with a concentration of mobile DNA elements. Using multiple strategies, I identified 23 coding genes on MSY. These genes were further deciphered by looking at their copy numbers, expression patterns, divergence rate, and polymorphisms. My analysis found in extant dogs at least two copies of the SRY gene, which were found in

the centre of palindromes. MSY genes were categorised by their gene activities: widespread, low-expression, and unique to the testis. Through construction of a Y-chromosome-based phylogeny, I was able to infer historical instances of gene evolution, indicating a dynamic evolution of the dog's Y chromosome.

Of particular note is the pseudoautosomal boundary (PAB), the point at which dog X and Y chromosomes lose their ability to exchange DNA through recombination. The extant dog PAB was inferred to be of *Canidae* origin, and X-linked and Y-linked PAB included different genes. Moreover, I observed the enrichment of mobile DNA elements called SINEs around the PAB. This work suggests that SINEs may have played a role in suppressing recombination near the PAB, leading to a divergence between X and Y chromosomes.

As in other mammals, the dog X chromosome maintained its gene content whereas the Y chromosome purged most of its genes throughout the development of sex chromosomes. One exception to this generalisation is the Y chromosome-based *PRSSLY* gene that I first detected on the dog Y chromosome. Extending my studies to other species, I found that the *PRSSLY* was broadly preserved on the Y chromosomes and absent on the X chromosome of other mammals. This makes *PRSSLY* the first example of a gene that originated from the paired ancestral autosomes, but was only retained on the Y chromosome following sex chromosome evolution. Based on characterisation of its gene activity in the testis, my data supports the hypothesis that the *PRSSLY* protein supports sperm production.

My studies generated the most complete assembly of the dog Y chromosome to date. In turn, the assembly enabled me to define aspects of gene evolution, structural features, and boundary regions, highlighting its dynamic nature. These discoveries enhanced our comprehension of the Y chromosome's evolution and its role in dogs. My assembly of the dog Y chromosome has provided the dog genetic research community with a resource to improve data analysis.

Abstract

Mammalian sex chromosomes, the oldest XX\XY system, are thought to have evolved from a pair of autosomes around 180 million years ago. During evolution, X and Y chromosomes differentiated along their lengths progressively. On the Y chromosome, this manifested as inhibition of recombination, genetic decay, and inversions. The mammalian Y chromosome, which is widely known for its diverse and complex repetitive sequences, differs from other chromosomes in terms of its size, genomic structure, gene content, and evolutionary trajectory. It is well known that the Y chromosome is crucial for testis development and gametogenesis. Morbidities are also related to Y chromosome dysfunction: deletions located on the Y chromosome can cause defective spermatogenesis and male sterility, while tumour susceptibility is also linked to Y chromosome genes. The Y chromosome is also a unique tracer of ancestry; its paternal genetic information enables the investigation of male demography and the application of forensic medicine. Humans, mice, rhesus macaques, chimpanzees, and cattle are just a few of the model species for which whole or almost complete Y chromosome sequences have been obtained. In this effort, a dog Y chromosome was assembled to high contiguity and used to shed light on genome structure, the course of Y chromosome evolution, and gene features.

Dog Y chromosome sequences are based on reads from a Labrador retriever dog produced with various sequencing platforms. Long reads of PacBio sequencing were assembled using Falcon and Flye, followed by scaffolding steps with Bionano and Dovetail Hi-C technologies. Two generated superscaffolds were then oriented and connected resulting in dog male-specific Y chromosome sequences of 6.78 Mb in length. Constituents of the assembled Y chromosome include a single-copy region, multiple-copy region, X-transposed regions, and autosomal homologous sequences. Other unique features of the chromosome include the detection of massive repetitive sequences, such as the enrichment of LINE transposable elements at the distal end of MSY and 0.86 Mb of LINE1_CF that occur as tandem repeats.

MSY genes were annotated and characterised to determine their copy number, transcriptional expression, phylogeny, divergence rate, and polymorphisms. It is inferred

that dog MSY genes arose from three evolutionary strata, and five genes -- *TSPY*, *CUL4BY*, *BCORY*, *SRY*, and *UBE1Y* – occur as multiple copies. Based on their expression, MSY genes were grouped into three categories: ubiquitous, low-expression, and testis-specific. These categories displayed significantly different evolutionary rates, potentially as a consequence, or in response, to their presumed different functional roles. Phylogenetic analysis identified evidence of conversion events in seven MSY genes, revealing a dynamic evolution of dog Y chromosomes.

The pseudoautosomal boundary (PAB) of dog sex chromosomes was also defined. The dog PAB descended from the common ancestor of the *Canidae*. The PAB contains *CLDN34* and *TETY2* in the X-linked and Y-linked PAB, respectively. *CLDN34* and *TETY2* appear to share the same promoter in the PAB and they are co-expressed in the testes. The PAB's SINE content accumulated near the PAB, suggesting that these small mobile elements may have catalysed or reinforced inhibition of recombination in this region.

Finally, the dog's Y chromosome was found to have a novel gene called *PRSSLY*. This unusual gene appears to be the first gene to arise from the paired ancestral gene that was lost from mammalian X chromosomes, but maintained on Y chromosomes. Single-cell transcriptomics and *in situ* hybridisation analyses revealed that *PRSSLY* expression occurs within the semineferours tubules of the testes, suggesting that its encoded protein facilitates spermatogenesis.

Taken together, this project elucidated lineage-specific characteristics of the dog Y chromosome and underlined its dynamic nature through gene activity, structural features, and boundary sequences. These discoveries add to the understanding of mammalian Y chromosome evolution and provide the scientific community with a valuable resource to improve whole genome sequence analysis.

Table of Contents

Table of Contents

List of Figures

Chapter 1

Chapter 3

Chapter 4

List of Tables

Abbreviations

µL	microlitre
µM	micromolar
bp	base pair
CAGE	Cap Analysis Gene Expression
cDNA	complementary DNA
CDS	coding sequences
DNA	deoxyribose nucleic acid
dpb	days post birth
E	embryo
FISH	Fluorescence In-Situ Hybridization
FSX	female-specific X
g	gram
gDNA	genomic DNA
Hi-C	high-throughput chromatin interaction
INDEL	small insertion/deletions
Ka	non-synonymous substitutio
Kb	kilo-base pair
KO	knockout
Ks	synonymous substitutio
L	litre
LC	liquid chromatography
LC-MS	liquid chromatography-mass spectrometry
M	molar
MAF	minor allele frequency
Mb	megabase

mins	minutes
mL	millilitre
mM	milli-molar
MRCA	most recent common ancestor
mRNA	messenger RNA
MSY	male-specific region of the Y chromosome
MYA	million years ago
NAHR	non-allelic homologous recombination
NCBI	National Center for Biotechnology Information
ng	nanogram
NGS	next-generation sequencing
NOR	nucleolus organizer region
OP	optical mapping
ORF	open reading frame
PAB	pseudoautosomal boundary
PAR	pseudoautosomal region
PCA	principal component analysis
PCR	polymerase chain reaction
qPCR	quantitative polymerase chain reaction
rDNA	ribosomal gene
RIN	RNA integrity number
RNA	ribonucleic acid
rpm	revolutions per minute
rRNA	ribosomal RNA
RT-PCR	reverse transcription PCR
SINE	short interspersed nuclear element
SMRT	Single Molecule Real Time

SNV	single-nucleotide variant
SRA	Sequence Read Archive
TEs	transposable elements
TPM	transcripts per million
TSD	temperature-dependent sex determination
TSS	transcription start site
t-SNE	t-distributed stochastic neighbor embedding
UTR	untranslated region
WGS	whole-genome sequencing
WBP	week before present
YBP	years before present

Foreword

The experiments within this thesis contributed to the production of two manuscripts. Firstly, the dog Y chromosome assembly, annotation and pseudoautosomal boundary analyses are in preparation for a scientific publication. For the purposes of this thesis, the results in this manuscript are distributed across Chapters 3, 4 and 5 to permit a significant extension and inclusion of additional experiments and discussions that could not be included in the manuscript. The content of these chapters has been explicitly written for this thesis and are distinct from the published manuscript.

We are currently developing a second manuscript that describes the *PRSSLY* gene including its evolution and putative functions. However, due to other researchers pre-empting our experiment and most of our results being similar to theirs, we have decided to conduct a more extensive study. Therefore, we are only presenting a portion of our experimental outcomes at this stage, with the remaining results to be disseminated in an academic publication.

All scripts have been deposited on GitHub at https://github.com/WengangXbio/script_bio. All figures in the manuscript were generated exclusively by Wengang Zhang and have been modified for the inclusion in the thesis. The experiments and manuscript were primarily executed and prepared by Wengang with analytical and written contributions from colleagues, who are duly acknowledged in this thesis.

Acknowledgments

I would like to take this opportunity to express my gratitude and appreciation to everyone who has supported me during my academic journey. Your help and encouragement have been invaluable in shaping my career and personal growth.

First and foremost, I want to express my deepest gratitude to my supervisor, Dr. Jeffrey Schoenebeck. Your meticulous guidance, encouragement, and dialectical thinking have been instrumental in shaping my research. I appreciate your patience, understanding, and financial support throughout my academic work, and your unwavering assistance in finding more research resources to help me complete my studies. Thank you for caring for my physical and mental well-being during the COVID-19 pandemic. I can still vividly recall the scenes of academic discussions we had in Teviot and other pubs. You are truly one of the most important people in my academic career.

I would also like to extend my heartfelt thanks to my secondary supervisor, Dr. Lel Eory. Your guidance and enlightenment in the field of bioinformatics have been invaluable. Your diverse perspectives have enriched my experiments, making them more comprehensive. Thank you for always being there to help me find solutions and for being an amazing supervisor and friend.

I would like to thank all the members in Schoenebeck's lab, especially Dr. Jenni Irving Mc Grath, Dr. Derya Odemir, and Dr. Melany Jackson, for their companionship and assistance. I am also grateful to all the staff members at Roslin, including Dr. James Prendergast, Dr. Emily Clark, Prof. Alan Archibald, Dr. Gregor Gorjanc, Dr. Jacqueline Smith, Prof. Ross Houston, Dr. Mazdak Salavati, Dr. Richard Kuo, Dr. Megan Davey, and Dr. Dominic Kurian, for their selfless help in my academic pursuits.

My heartfelt thanks also go out to my collaborators, Prof. Gregor Larson, Prof. Laurent Frantz, Dr. Shelagh Boyle, Jiaqi Yang, Prof. Jeff Kidd, Dr. Melissa Wilson, Dr. Steven Fiddaman, and Prof. Hannes Lohi for their help in my experiments.

I would like to thank all of my Chinese friends at Roslin Institute, especially Dr. Debiao Zhao, Dr. Zhiguang Wu, and Dr. Tuanjun Hu, for their care and concern over the past four years. I feel at home in Edinburgh thanks to you. And to my great friends in Edinburgh, Dr. Alan Zhao, Xi Tao, Di Shi, Dr. Lun Yao, Dr. Xiang Xu, Zejun Yan, Dr.

Guangwen Yin, and Dr. Chuifan Zhou, thank you for your friendship and all the wonderful memories.

Lastly, I would like to express my heartfelt gratitude to my family. To my ~~girlfriend~~ **wife**, Dr. Jingjing Wang, thank you for your waiting. Your spiritual encouragement, understanding, and acceptance of me have been instrumental in shaping who I am today. And to my parents, I am indebted to you for your unconditional love and care. I hope to make you proud and repay your sacrifices by becoming a successful and compassionate individual.

PS: After my viva, I would like to add my appreciation to my two examiners Professor Qi Zhou and Professor Dan Macqueen. Qi gave me many valuable comments on my work and Dan reviewed my thesis very carefully, pointing out many typos and inappropriate expressions.

CHAPTER 1: Introduction

1.1 The Mammalian Y Chromosome

Mammalian sex chromosomes are well known as the oldest sex system, which evolved from paired autosomes prior to the split of the marsupial and placental mammals at around 180 million years ago (MYA) (1). Unlike autosomes, whose genes are distributed in a shuffled way such that their expression pattern is heterogeneous for various tissues, the sex chromosomes present an enrichment of sex-related genes whose functions include sexual identity, development, and reproduction. Heteromorphic sex chromosomes are the consequence of collective genetic decay on Y chromosomes with different mutation rates, sexual selection, effective population sizes, and recombination rates. Mammalian X chromosomes are enriched with genes and conserved in gene contents, whereas mammalian Y chromosomes of different species display enormous differentiation in terms of size, structure, and gene contents.

Our knowledge regarding the biological roles of the Y chromosome comes from a few well-studied model organisms, particularly humans (2,3) and mice (4). The development of molecular genetics and genomics contributed to the discovery of the biological and medical significance of the Y chromosome. Y chromosomes are implicated in the biological roles of sex determination, spermatogenesis, and generalised regulation of transcription and translation (5). Nevertheless, despite these roles, our biological understanding of the Y chromosome still lags behind that of autosomes and the X chromosome. Beyond its normal biological functions, several pieces of evidence reported (6–8) suggest that the Y chromosome contributes to sexual dimorphism of disease presentation and health. The Y chromosome also facilitates male demography (9) and sex-biased admixture (10) for population genetic studies.

The process of domestication involves the transformation of wild animals into forms that are better suited for human use and companionship. The dog, a domesticated descendant of the wolf, belongs to the family *Canidae*. The genus *Canis* encompasses a range of species including not only dogs, but also wolves, coyotes, and jackals (11). This

diverse group falls under the subfamily *Caninae* within the order *Carnivora*, which comprises various mammals (e.g. cats, bears, and seals).

Genetic and archaeological evidence indicates that dogs existed over 30,000 years ago, with gene flow observed between dogs and wolves (11,12). Despite this knowledge, the specific geographic and temporal origins of dogs continue to elude researchers, and the contentious topic of how many domestication events existed in prehistory remains unresolved.

Many perspectives make canines interesting to study. First, its genome encodes a staggering variety of phenotypes that have been modified through domestication, adaptation to human-dominated surroundings, and artificial selection. Because of this special bond, canine genomic research has the chance to not only elucidate what makes a dog a dog, but also inspire comparative genomic techniques in their human companions. Second, the identification of variants underlying genetic disorders is an important use of trait mapping in dogs. Since numerous canine diseases have related human analogs, causal variants in dogs can inform the genetic underpinnings of human conditions. Dogs are particularly helpful in this perspective behind such problems because some diseases that are polygenetic and/or rare in humans are monogenic and/or common in specific dog breeds. Third, the coevolution of dogs and humans facilitates our understanding of their genomic relationships corresponding to certain phenotypes. Dogs have lived alongside people for a longer period of time than any other domestic animal, sharing our dietary and pathogenic milieu as our species transitioned from a hunter-gatherer lifestyle to agriculture. Hence, some human adaptations, immune and digestive systems might have evolved in parallel in dogs.

The generation of dog Y chromosome sequence allows the study of population history and male demography, as well as the comprehensive understandings of genes gain insights about the functions of the Y chromosome that are dog-specific. Broadly, the elucidation of the dog Y chromosome can enable broader comparative analysis within the mammalian Y chromosomes, expanding our knowledge on the general evolutionary process of eutherian Y chromosomes, and studying dog Y chromosomes provides

alternative approaches to understanding mechanisms and finding solutions for infertility in humans.

1.1.1 Origins and Evolution of Mammalian Y Chromosomes

Sexual reproduction is a ubiquitous attribute of vertebrate life cycles (13), and involves the process of parents' gametes being fused to create a zygote. The gamete is a reproductive cell carrying only one set of dissimilar chromosomes (haploid) and the zygote is a fertilized egg cell that is composed of cells with two sets of chromosomes (diploid) (14,15). Gametes being morphologically similar or different in size are called isogamy and anisogamy, respectively (16,17). Anisogamy is universal among eukaryotes: following meiosis females produce a large gamete (ovum). In contrast, males produce a small gamete (sperm). Molecularly, the eukaryotic sex-determination system mainly occurs as either male-heterogametic (XX/XY system) or female-heterogametic (ZW/ZZ system). The XX/XY system predominates among mammals while the ZW/ZZ system is common among birds and snakes (18).

A mammal is a vertebrate animal of the class Mammalia, which is characterised by the presence of milk-producing mammary glands for feeding offspring. Taxonomically, monotremes are one of the three groups of living mammals. This ancient group of mammals includes the echidna and platypus and is notable for laying eggs rather than birthing live young as occurs in other mammal groups. The other two mammalian groups include placental (eutherians) and marsupial (metatherians) mammals. These latter two are often subclassified as therians. Most mammals use the male-heterogametic XX/XY sex system. Exceptional species include the Amami spiny rat and the Tokunoshima spiny rat (19). Males of both species lack a Y chromosome, therefore their sex system is regarded as XX/X0. Cytologically, placental and marsupial sex chromosomes are highly dimorphic such that the X chromosome is commonly over double the size of the Y chromosome (**Figure 1.1 A-E**). In contrast, the monotreme's sex chromosomes are small in size and present multiple chromosomes in the haploid genome (**Figure 1.1 F**) (20,21). For example,

the platypus has 5 distinct X and Y chromosomes and the echidna has 5 X chromosomes and 4 Y chromosomes which are not homologous with those of the platypus (22).

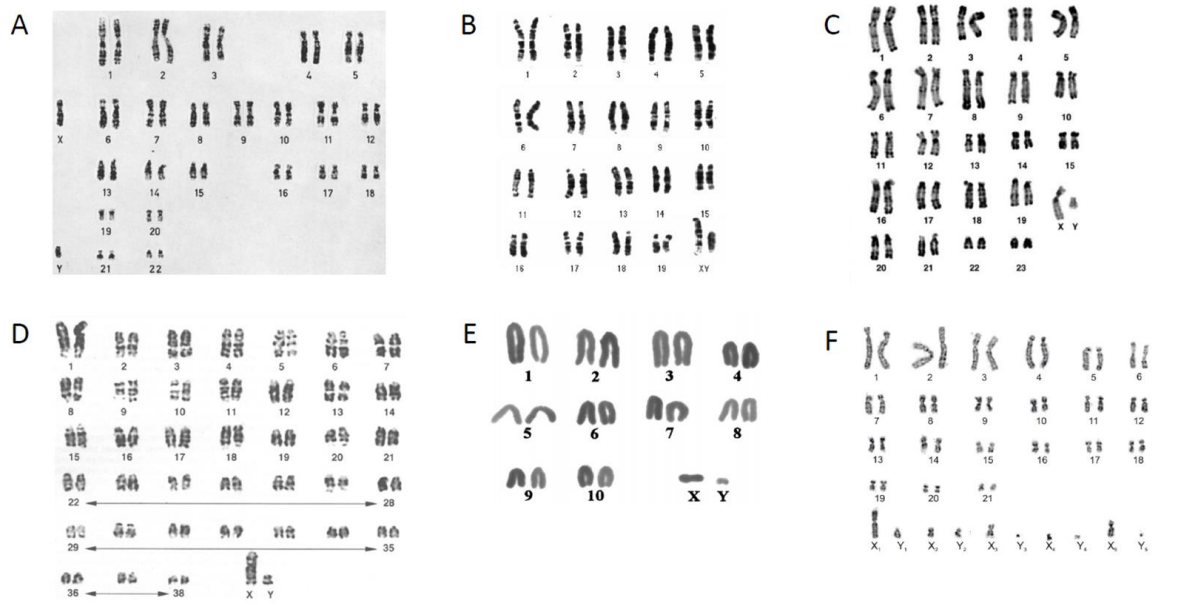


Figure 1.1. Banded karyotypes in male mammals. Human (A), mouse (B), rabbit (C), dog (D), possum (E), and platypus (F) are shown. In morphology, the X chromosome of placental mammals is more than twice the size of the Y chromosome. The eutherian X chromosome represents approximately 5% of the haploid genome. In marsupials, the X chromosome is smaller, representing 2/3 of that of eutherian mammals. The platypus, which belongs to the monotremes, has five X chromosomes and five Y chromosomes in males. Images are modified from (22–27).

In terms of the origin of the mammalian XX/XY system, mammals have obviously different sex chromosomes from birds and reptiles. So it has been assumed that the mammalian XX/XY system originated *de novo* following the split of the mammalian and avian lineages around 300 million years ago (MYA) (**Figure 1.2A**). However, this assumption is confused by the findings of the platypus with its multiple sex chromosomes (28). The X chromosome of the platypus was reported as sharing homology with one end of the sex chromosomes of the eutherian and marsupial mammals (29). This hypothesis has been recently challenged due to the finding that the X chromosome of therian mammals lacks any homologous regions with it. Instead, it appears that the homologous regions are scattered and resemble a chain-like structure similar to the chicken sex chromosome (Z). This discovery raises questions about the previous assumptions and requires further investigation to understand the evolutionary implications fully (22,30)

(**Figure 1.2B**). Although the origins of the platypus's sex chromosomes are unclear, it is widely accepted that the sex chromosomes of placental mammals and marsupials have been derived after the therian-monotreme split. The therian is a common ancestor of placental and marsupial species. The therian X chromosome shares homology with platypus chromosome 6, indicating the therian sex chromosomes were derived from the ancestral paired autosomes (30). Also, the analysis of gene repertoires on the Y chromosomes showed the oldest Y-specific genes originated in the common therian ancestor at around 180 MYA, which marked the beginnings of divergence between the X and Y chromosomes (1,31).

The ancestral paired autosomes are referred to as "proto-sex chromosomes", which existed at the common ancestor of the therian 180 MYA ago. The proto-X and proto-Y became the modern X and Y chromosomes under a series of evolutionary events. What forces drove their specialisation? The initial step was generally believed to be the acquisition of a male-determining gene by one of the proto-sex chromosomes. In the therian, the male-determining gene was *SRY*, which enabled the proto-Y chromosome to take a role in sex determination (5,32,33). Following the emergence of *SRY* and selection for sexually antagonistic mutations in the vicinity of *SRY*, recombination between the proto-X and proto-Y chromosomes was suppressed via inversions to resolve sexual conflict (32).

The inhibition of recombination enabled the proto-X and proto-Y chromosomes to evolve in distinct trajectories. On one hand, the structure, organisation, and gene content of the X chromosome are highly conserved across evolutionarily diverse mammals (34–37). This is explained by Ohno's law, which states that since the X chromosome follows the dosage compensation, which is different from that of autosomes, any rearrangement between the X chromosome and autosomes is regarded as detrimental and will be eliminated (38,39). Also, the recombination landscape is constrained within the mammalian X chromosome evolution. A huge recombination coldspot that covered about the central one-third of the X chromosome and extends tens of megabases distally from the centromere was shared by cats, dogs, pigs, and humans. And the huge coldspot was

flanked by conserved hotspots, where the recombination rate elevated extremely in studied mammals (40,41). Hence the reduced recombination could explain the reduction in diversity in this region and is favourable for recurrent selective sweeps. But the mechanism underlying the driving force of low recombination rate is unclear, and the relationship between the collinear recombination landscape and the dosage compensation has not been addressed (34).

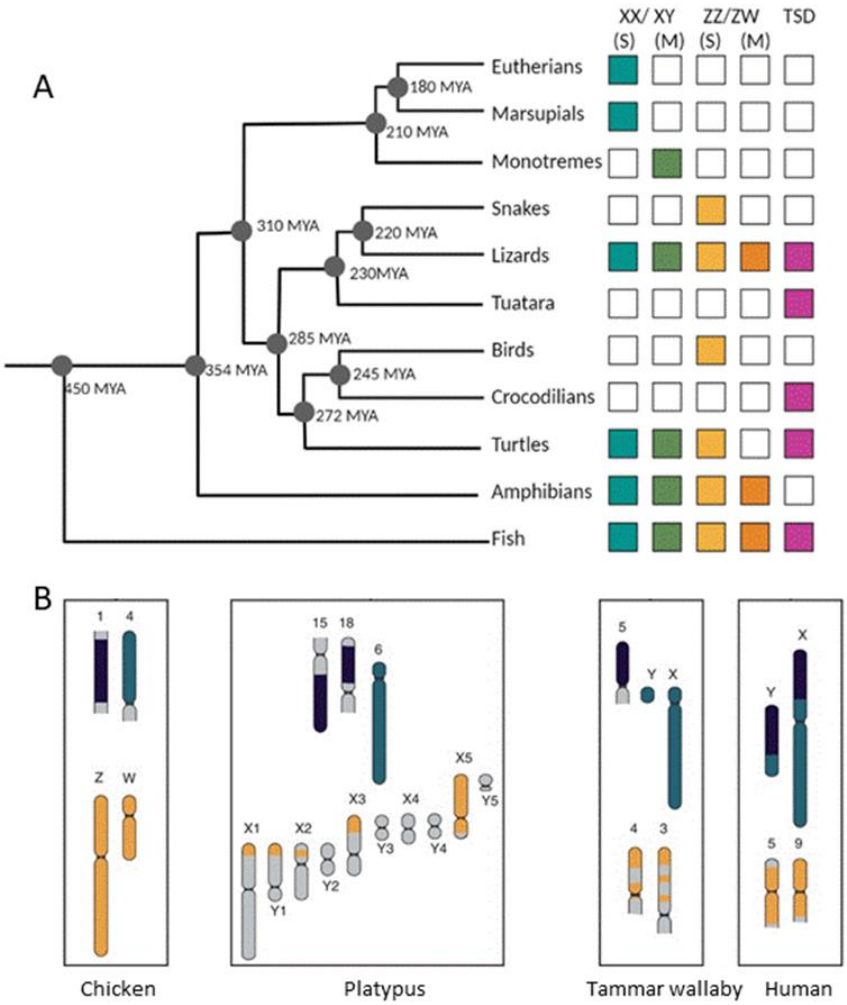


Figure 1.2. Sex determination systems in vertebrates. Evolution versus sex chromosome systems (A). Therians (eutherians and marsupials) are under the control of the XX/ XY system with singleton X and Y chromosomes. Monotremes have multiple X and Y chromosomes. Reptiles and fish present a great variety of sex-determining systems. TSD stands for temperature-dependent sex determination, and S and M present single and multiple sex chromosomes, respectively. Orthology of sex chromosomes and autosomes among chicken, platypus, tammar wallaby, and human (B). The sex chromosomes of the platypus share homology with the chicken's Z and W chromosomes, but not with the sex chromosomes of therians. Therians' sex chromosomes originate from paired ancestral

chromosomes that are homologous to chromosome 4 of chicken and chromosome 6 of platypus. Figures are modified from previous studies (18,42).

On the other hand, the therian Y chromosome evolved rapidly and species-specifically (4,43,44). However, some timings of evolutionary events are unclear to date, for example, whether inversions catalyse or were a consequence of recombination suppression between sex chromosomes, and whether the accumulation of sexually antagonistic loci was the cause or the result of halting recombination (45,46). It is certain that the efficacy of natural selection was reduced due to the absence of recombination for some regions of the proto-Y chromosome. Detrimental mutations accumulated at recombination-free regions of the proto-Y chromosome and led to the current Y chromosome's degeneration. The non-functional and recombination-free regions of the proto-Y chromosome can be inverted, amplified, and deleted leading to the current Y chromosome in reduced physical size (**Figure 1.3**).

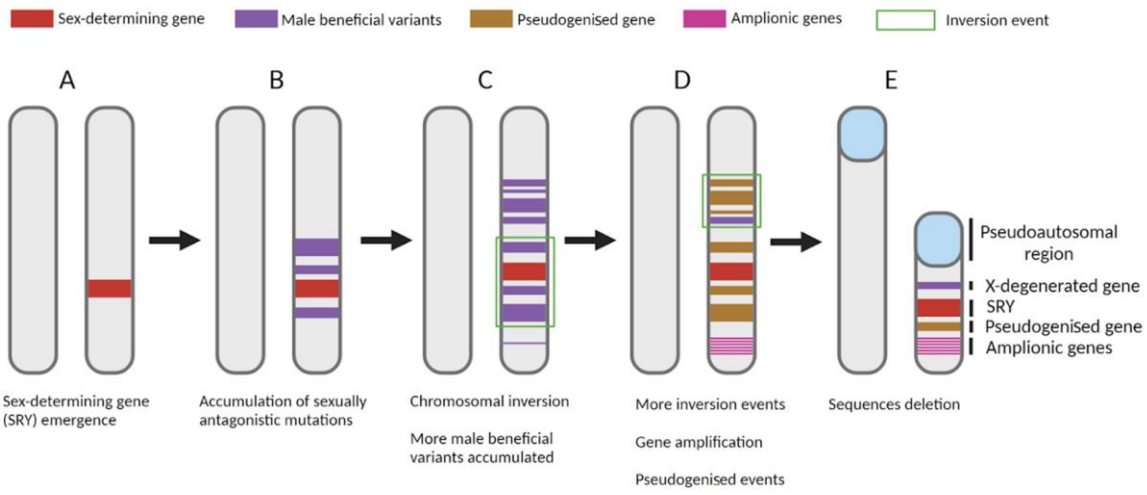


Figure 1.3. An evolutionary model of mammalian Y chromosome evolution. (A) The mammalian sex chromosomes originated from autosomes following tranposition of a male-determining gene which is generally believed to be *SRY*. (B) Accumulation of sexually antagonistic mutations that occurred in the vicinity of *SRY*. (C) Chromosomal inversion on the Y chromosomes eliminated the recombination between the proto-sex chromosomes. (D) Accumulation of deleterious mutations led to the pseudogenisation of coding genes on the proto-Y chromosomes, and sequences were repeated accompanying the expansion of repetitive genes on the proto-Y chromosomes. (E) Segments of DNA without function were lost resulting in the reduced physical size of the Y chromosomes. Figures are modified from previous studies (32).

Why do mammalian Y chromosomes normally degenerate quickly? Recombination-free of heterochromatic Y chromosomes is the cause of the degeneration of Y chromosomes. Cessation of recombination accumulates deleterious mutations and incorporates fewer favourable mutations on the Y chromosomes. As a result, Y chromosomes frequently display a lower level of adaptation than autosomes and X chromosomes. Also, sexually antagonistic mutations - which are beneficial to males but detrimental to females- are supposed to be selected in the population. This leads to a restricted activation of genes on Y chromosomes, specifically associated with male functions such as spermatogenesis and male development (32).

An extended question: what are the underlying mechanisms to explain the accumulation of deleterious mutations and the incorporation of a few beneficial mutations on the Y chromosome? Various models have been proposed, including Muller's ratchet, genetic hitchhiking, Hill-Robertson interference, and background selection (**Figure 1.4**). The theory of Muller's ratchet refers to the irreversible accumulation of mutational decay on the Y chromosomes due to their uniparental inheritance, high mutation rates, and lack of effective recombination (47,48). Genetic hitchhiking is a complement model for Muller's ratchet that states when beneficial mutations are selected, linked deleterious mutations on the Y chromosomes are dragged along to fixation (49). The Hill-Robertson interference describes a phenomenon where a strong linkage between two adaptive mutations will decline the efficiency of selection in a finite population (50). Thus, the hemizygous nature of Y chromosomes, in the long term, has less chance to accumulate advantageous mutations than the X chromosomes. Finally, background selection is a process where mild fitness mutations are at risk of being eliminated when their linked deleterious mutations are strongly selected (51,52). The background selection can lead to a reduction in the level of adaptation of Y chromosomes in a large population (53).

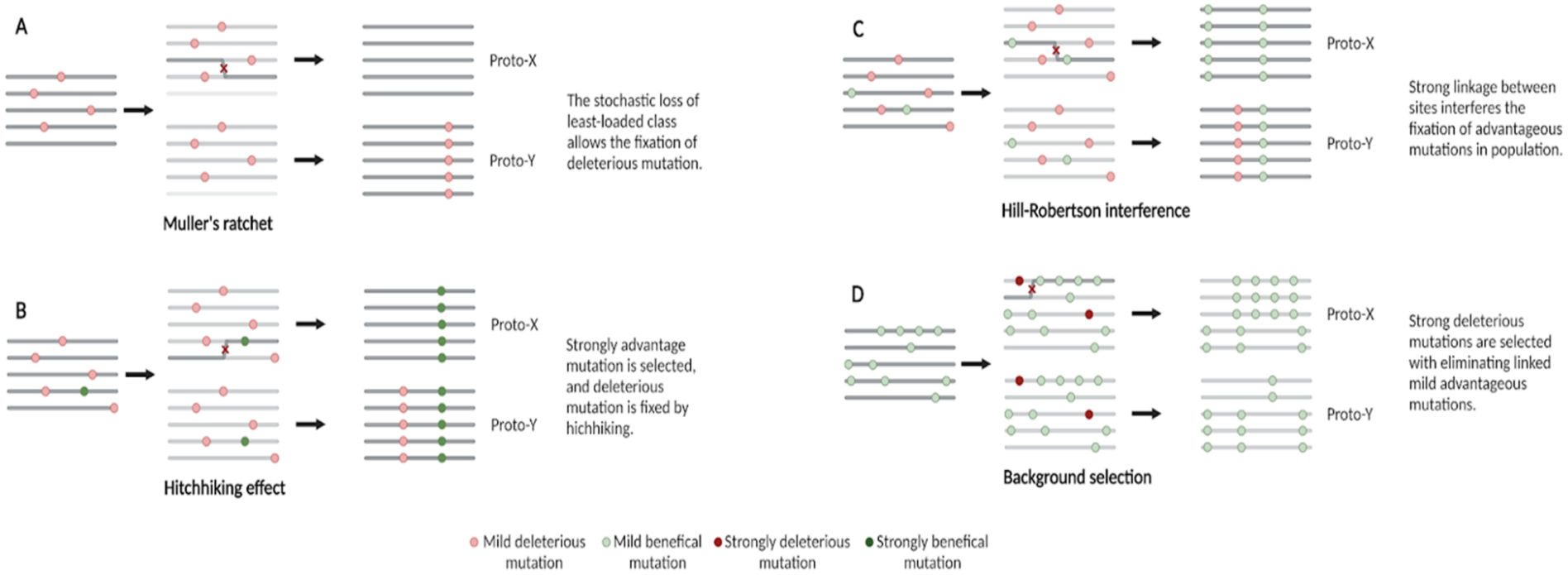


Figure 1.4. Models for Y chromosome degeneration.
(A) Muller's ratchet theory refers to the fixation of deleterious mutation on the Y chromosomes followed by the stochastic loss of the least-loaded class. The least-loaded class of chromosomes is the one with the fewest deleterious mutations. Once lost, the least-loaded chromosome cannot be restored in the absence of recombination. (B) Hitchhiking effect enables deleterious mutation(s) to be fixed within the population when it is linked with a strongly advantageous mutation. The X chromosomes avoid permanent inclusion of deleterious variants by their ability to recombine. (C) Hill-Robertson interference inhibits the selection of advantageous mutation on the Y chromosome. The recombination-free Y chromosomes have no chance to fix multiple fitness variants from different lineages. (D) Background selection is the process that eliminates accumulated mild advantageous mutations by the selection against a strongly deleterious mutation on the Y chromosomes. The X chromosomes can store the advantageous variants due to their ability to recombine during meiotic recombination in females. For all the four figures, each grey line represents a haplotype in the population.

1.1.2 Features of Eutherian Y Chromosomes

1.1.2.1 Structure of Y Chromosomes

Therians' Y chromosomes are invariably composed of at least five distinct regions: one or more pseudoautosomal region (PAR), X-degenerate sequence, X-transposed sequence, ampliconic regions, and heterochromatin (**Figure 1.5**). The total length varies widely across species, with 11 Mb in rhesus macaque, 22 Mb in humans, and 90 Mb in mice (4,54). PARs are the only regions that maintain homology with the X chromosome and synapse during meiosis. In humans, two PAR are located at either tip of the Y chromosome (44). In other mammals including dogs and cats, the PAR comprises a single segment (55). Excluding the PAR, the remaining four regions (i.e. X-degenerate, X-transposed, ampliconic, and heterochromatic sequences) are located within the so-called male-specific region of the Y chromosome (MSY) where the crossover with the X chromosome does not occur during meiosis. The X-degenerate region originates from the ancestral autosome and contains genes that are differentiated from X-linked homologous genes. Acknowledging their common evolutionary origins, the genes within the X-degenerate region and their X chromosome counterparts are commonly called "gametologs", e.g. *HSFX* and *HSFY* are X-linked and Y-linked gametologs, respectively. As its name suggests, X-transposed sequences are believed to derive from a recent transposition event and therefore display high sequence conservation and homology to X chromosome sequences. X-transposed sequences can happen as multiple blocks on the Y chromosomes and usually contain genes that display high similarity with X-linked gametolog sequences (43,44,55).

Heterochromatin on the Y chromosome is a form of DNA that is densely packed with satellite repeats sequence and genetically inert. Different species display significantly variable heterochromatin content, from sheep (small) to elephants (large) (56). In humans, heterochromatin is found on the distal part of the long arm and covers 29% to 54% of the Y chromosome in length, with a median of 44% (57–59). In mice, heterochromatin is almost absent from the Y chromosomes.

In the Y chromosomes, the ampliconic areas are widely distributed and are intrachromosomal, self-similar (even identical) sequences often contain multiple-copy gene families within. The ampliconic regions are highly variable in gene contents and copy numbers even between close species, and the amplified gene families usually are predominantly expressed in testes. These genes are thought to be associated with male gametogenesis (4,60–62) and also potentially engaged in an arms race with X-linked gametologs for transmission of respective X or Y chromosomes to the sperm cells during male spermatogenesis (63–65).

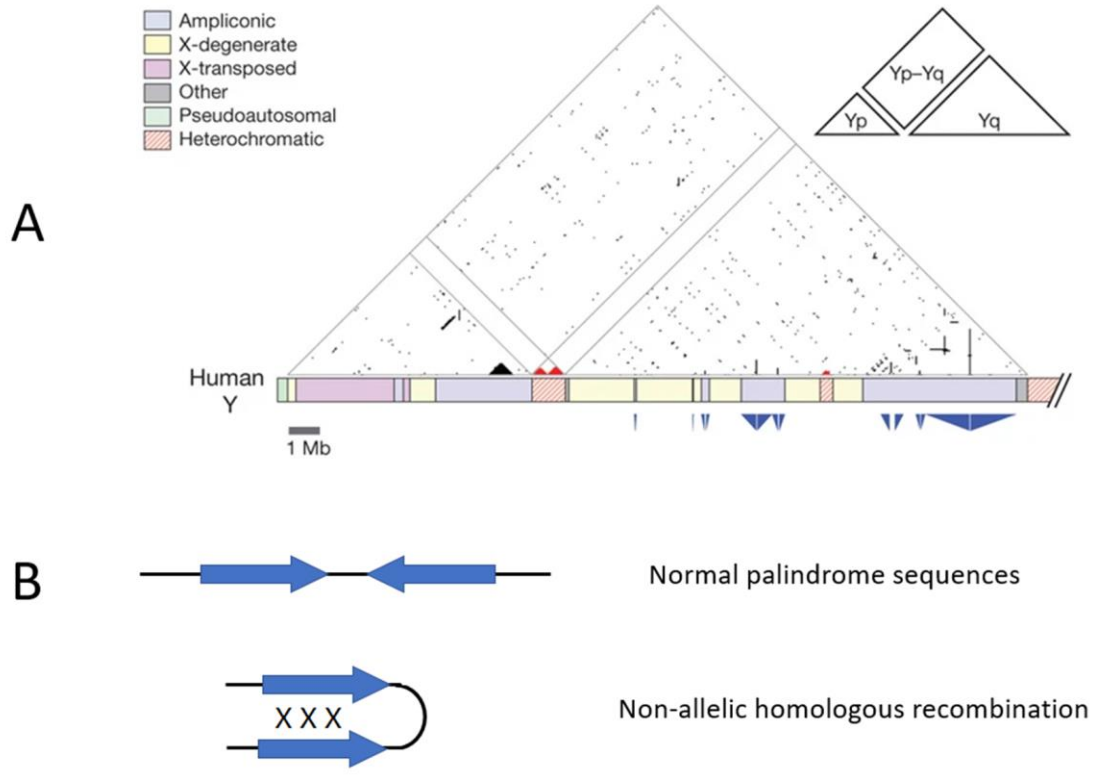


Figure 1.5. Structure and repeat features of the human Y chromosome. (A) The triangular dot plot of DNA sequences illustrates repeat structures in the MSY. Tandem repeats appear as horizontal lines, inverted repeats as vertical lines, and palindromes as vertical lines that nearly intersect the baseline. There are 8 palindromes identified in the long arm of the Y chromosome. Schematic representations of chromosomes are shown below plots; colour-coding indicates sequence class. (B) Non-allelic homologous recombination (NAHR) makes two arms crossover and prevents gene degeneration from accumulation of deleterious mutations. The NAHR is supposed to be a mechanism of gene conversion on the Y chromosomes. Figure A is modified from (66).

1.1.2.2 Repetitive DNA of the Y Chromosome

The abundance of repetitive DNA is a notable feature of the Y chromosome, which is found not only in the ampliconic region but also in the X-degenerate region and heterochromatin. To better understand the evolution of the Y chromosome, the Y chromosome sequences of several species were assembled accurately, and their repetitive sequences are well characterised. In general, repetitive DNA can be distributed in the pattern of tandem repeats (head to tail), inverted repeats (palindrome sequences), and interspersed repeats such as transposons and retrotransposons. Repeats are composed of a great number of sequences that present striking homology, as much as 99.99%, with other MSY sequences. To date, only four eutherian species have fully assembled Y chromosomes in euchromatin: humans (44), rhesus macaques (54), chimpanzees (66), and mice (4). Of these species, humans and chimpanzees have similarly sized Y chromosomes and their repetitive regions are 10.2 Mb and 14.7 Mb in length, occupying 44.7% and 57.0% of MSY sequences respectively (**Figure 1.5A**). The ampliconic region of humans contains a total of 60 copies of genes corresponding to nine gene families. Despite ~7 million years of separation from humans, the chimpanzee Y has only 25 copies of genes belonging to six gene families. Mice and rhesus monkeys represent the extremes of ampliconic and repeat content: In mice, most MSY sequences are made of amplicons (97.9%) with a total length of 87.7 Mb, and nine gene families are located within the ampliconic regions with 633 multicopy genes. In contrast, repetitive sequences of rhesus macaques collectively span 0.5 Mb with only 4.5% of MSY sequences. Although incomplete assemblies, the Y chromosomes of dogs (55), cats (55,62), bulls (64), and horses (43) have also revealed enrichment of repetitive sequences. Notably, ampliconic sequences roughly span 35 Mb of the bull MSY's long arm, with four basic repeat units. Bull's Y chromosome amplicons are similar to mice in that they contain three basic repeat units, though there is no homology between the repetitive sequences in bulls and mice. Taken together, massive MSY sequence amplification is a broad characteristic of eutherian MSYs, however distinct chromosome lengths and non-homologous amplicons

between species indicate their respective Y chromosomes experienced amplifications independently.

The enrichment of the palindrome sequences in amplicons is another common feature of the therian Y chromosome. In humans, 54% of ampliconic regions are palindrome sequences (5.5 Mb) from 8 palindromes with the length of arms ranging from 9 Kb to 1.45 Mb. Six of eight palindromes carry coding genes and self-recombining between two paired arms prevents genes from accumulating deleterious mutations by enabling arm-to-arm gene conversion (**Figure 1.5B**) (44). Rhesus macaques have three palindromes, two of which are orthologous to humans' (54). Chimpanzees have 19 palindromic sequences with a length of 7.5 Mb collectively and 12 palindromes are chimpanzee-specific (66). The draft assembly of the Y chromosome of the gorilla displays 13 gorilla-specific palindromes and 8 orthologous palindromes with humans (67). In mice, the massive ampliconic sequences of the Y chromosome are distributed as either head-to-tail tandem sequences (515 Kb) or head-to-head units as palindrome repeats (400 Kb) (4). Among *Artiodactyla*, the order of even-toed ungulate species, bulls are reported to have two palindromes on the Y chromosomes, one of which carries pseudogenes (68). In pigs, two palindromes of 120 Kb are formed around *SRY* and *CUL4BY* genes (69). For species belonging to the *Carnivora* order, no palindromes are reported despite the partial assembly of Y chromosomes (55,62,70). This deficit can be attributed to incomplete assembly or alternatively, few palindrome sequences occur among *Carnivora* species. In theory, the palindromes can exchange genetic components by self-crossover of two arms, which is called non-allelic homologous recombination (NAHR) (**Figure 1.5B**). This process has two potential features. On one hand, NAHR can reduce the sequence divergence between two palindromic arms, providing a potential way of eliminating detrimental mutations followed by escaping Muller's ratchet (71,72). As seen, in humans, most of the ampliconic testis-specific genes are located within the palindrome sequences and are still functional. On the other hand, NAHR can exchange across different palindromes resulting in structural rearrangement and copy number variants (73). Overall, palindromes evolve as species-specific for the eutherian Y chromosomes with

different lengths, copy numbers, and gene contents, and the structure of palindromes rescues coding genes from deleterious mutations implying their importance in functional maintenance related to spermatogenesis and male development.

The interspersed repeats refer to transposable elements that inserted into the host genome. In primates, the MSYs were occupied with 48.8%, 47.0%, and 43.7% of interspersed element sequences for humans, gorillas, and chimpanzees respectively (67). Interspersed repeats occur as short interspersed elements (SINEs), long interspersed elements (LINEs), and long-terminal repeat retrotransposons (LTRs). LINE1 is the predominant element in all three primate species, occupying 35.8% to 42.5% of total repeat sequences. The density of interspersed repeats occurs heterogeneously in different Y chromosome regions. In both humans and gorillas, the X-degenerate regions have a significantly higher interspersed content than the ampliconic regions (44,67). In mice, a similar density of repeat sequences to primate MSYs is observed, whereas the ampliconic region is just slightly lower than the X-degenerate region (4). When compared with autosomes, the MSYs have more interspersed repeats in all named species.

1.1.2.3 Gene Content on the Y Chromosome

In mammals, the recombination-free regions of the X and Y chromosomes allow for their evolution in different trajectories. By comparing orthologs of a basal tetrapod (*Xenopus tropicalis*), mammalian X chromosomes retain more than 95% of their ancestral genes and display conservation among species (74–76). Conversely, mammals' largely degraded Y chromosomes bear just a few dozen genes – a specialised set with an extraordinary evolutionary lifespan. Here, the Y chromosome genes are reviewed in different aspects.

'Evolutionary strata' are used to classify Y chromosome genes based on the time points when they stopped recombining with their counterparts on the X chromosome. The number of strata differs between mammalian species, as do the criteria researchers used to define them. Strata were defined by several methods including phylogenetic methods, synonymous versus nonsynonymous substitution rates of coding genes, and inversion analyses (77). Comprehensive surveys of mammalian Y chromosome gene

repertoires were performed by Bellott (1) and Cortez (31) independently, and both showed the same results that the oldest stratum (S1) genes stem from the last common therian ancestor at around 180 MYA. S1 genes in eutherians include *SRY*, *RPS4Y*, *HSFY*, *RBMY*, and *CUL4BY*, and some of them are dispensable in some species (**Figure 1.6**). The definition of the strata that follow S1 is controversial in eutherians, especially the existence of a stratum 3 (S3) and its boundary with stratum 2 (S2) (44,74,78). Similarly, a boundary between S3 and stratum 4 (S4) is also a source of debate (44,79–81). However, the paucity of gametologs within S3 and S4 due to the genetic decay limits the ability of these methods to define evolutionary strata stably.

Here, S2 and S3 genes are discussed together due to their similar differentiation time and rate of genetic divergence (43,44,78). In total, Bellott's (1) and Cortez's (31) studies of eutherians identified 14 S2/S3 genes: *AMELY*, *CYorf15*, *DDX3Y*, *EIF1AY*, *EIF2S3Y*, *KDM5D*, *OFD1Y*, *TMSB4Y*, *TSPY*, *UBE1Y*, *USP9Y*, *UTY*, *ZFY*, and *ZRSR2Y*. During the evolution of over 110 MYA, S2/S3 genes were maintained independently in different species, and closer species have more similar gene contents (**Figure 1.6**).

Some species' Y chromosomes evolved with an even younger stratum (S4 and S5), which was estimated to occur no more than 50 MYA (43,54). For primates, S4 and S5 genes emerged to be lineage-specific. Four S4/S5 genes (*TBL1Y*, *NLGN4Y*, *PRKY*, and *MXRA5Y*) are conserved in old-world monkeys (humans, chimpanzees, and rhesus macaques) in either coding or pseudogenised states, whereas the new-world monkey (marmoset) developed three novel S4/S5 genes (*AKAP17AY*, *P2RY8Y*, and *ZBED1Y*) and lost *TBL1Y* and *PRKY* (54). In horses, five S4 genes, *SHROOM2Y*, *TBL1Y*, *ANOS1Y*, *STSY*, and *NLGN4Y* were detected following assembly of the Y chromosome, three of which (*SHROOM2Y*, *ANOS1Y*, *STSY*) were not detected in primates (43). To date, S4 or more strata have not been described for mice, rats, bulls, cats, and dogs.

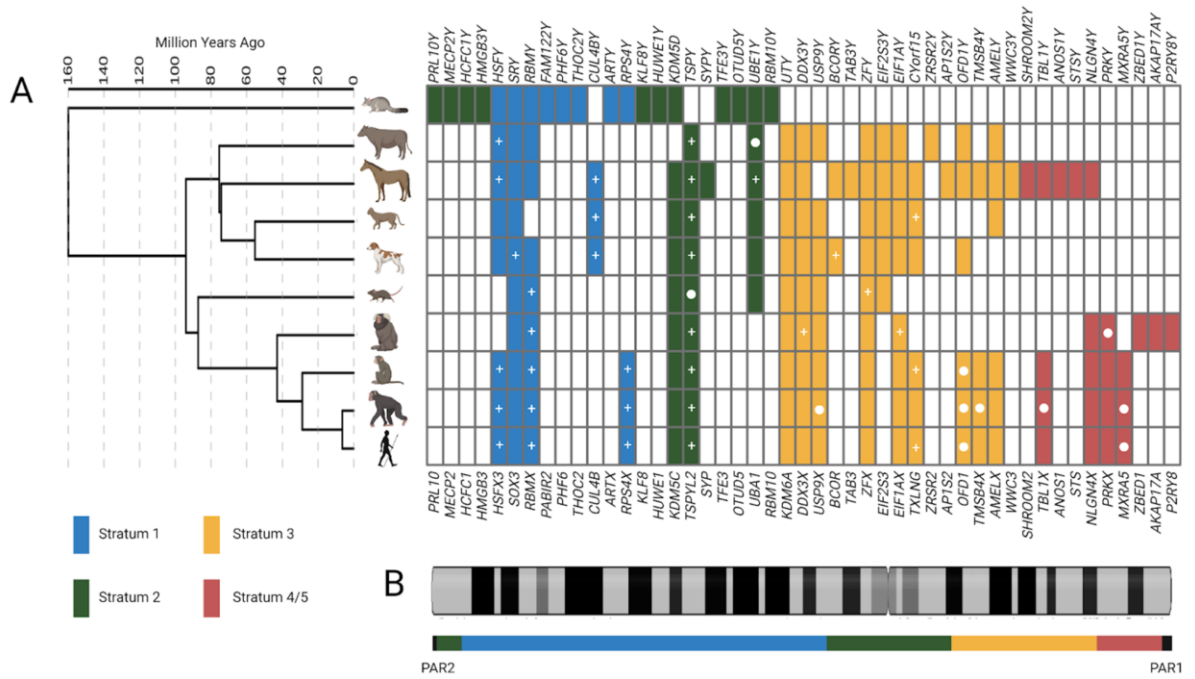


Figure 1.6. Gene repertoires of the mammalian Y chromosomes and the structure of human Y chromosomes. (A) Y chromosome gene repertoires. The mammalian species with a complete or nearly complete Y chromosome assembly are selected. The phylogenic tree is built using TreeTime (http://www.timetree.org/). The gene contents for primates (humans, chimpanzees, rhesus macaques, and marmoset), mice, dogs, cats, horses, and opossums are collected based on previous publications (1,31,43,62) and are ordered from left to right according to the coordinates of human X homologs. "+" represents multi-copy and "◦" represents pseudogenes. Y-linked gene and X-linked gene names are shown on the top and bottom, respectively. The proto-Y chromosome of marsupials and eutherians evolved in the same stratum 1 (S1) event triggered by the sex-determining gene *SRY*, which is conserved across therians. Marsupials stratum 2 (S2) and eutherians S2 occurred independently, leading to distinct sets of genes maintained on the Y chromosomes. *KDM5D* and *UBE1Y* span both therians and marsupials indicating the evolutionary convergence of the mammalian Y chromosomes. Marsupials experienced only two evolutionary strata compared with eutherian species whose Y chromosomes carry a further stratum 3 (S3). Within eutherians, Y chromosomes evolved at a high rate in horses, marmosets, rhesus macaques, chimpanzees, and humans with the emergence of stratum 4 (S4) or even stratum 5 (S5). (B) Schematic representation of the evolutionary stratum projecting on the human X chromosome. Pseudoautosomal regions (*PAR1* and *PAR2*) are located at the terminal ends of the Y chromosome. Each stratum progressively shortened the recombination regions between proto-sex chromosomes. The first two strata (S1 and S2) encompass the long arm and proximal short arm, respectively. S3 and S4 extend on the distal short arm until the PAR1.

Marsupials experienced the same S1 events with eutherians, while their retained genes (n=8) are twice the number of eutherian S1 genes. A marsupial S2 event generated

ten X-degenerate genes after the marsupial-eutherian split, containing the convergent differentiation of two genes (*KDM5D* and *UBE1Y*) (82).

In terms of origin, not all mammal Y chromosome genes derived from ancestral proto-sex chromosomes. MSY genes can be acquired from autosomes via transposition or retrotransposition events. Thus, they are usually single-copy and lack introns compared to their autosomal paralogs due to their retrotransposition through mRNA sequences. In humans, there are seven acquired genes (*CDY*, *DAZ*, *BPY2*, *PRY*, *VCY*, *TGIF2LY*, and *PCDH11Y*), some of which are shared with other primates as well (44). For example, the azoospermia factor c (AZFc) region is composed of palindrome sequences and enriched with three acquired genes (*CDY*, *DAZ*, and *BPY2*) (83). The AZFc amplicons occur as copy number variants and structural rearrangements on the Y chromosomes of primate species, indicating several independent transposition events occurred during primates' Y chromosome evolution (84). In mice, there are two acquired genes reported, *TEYorf1* and *PRSSLY* (4), however, these conclusions are controversial. Recent studies debated whether *TEYorf1*, which is called *TETY2* in cats, is the Y-linked gametolog of *CLDN34,* a gene located on the X chromosome (46,49). For *PRSSLY,* it was proved to arise from paired ancestral sex chromosomes. The loss of the X-linked gametolog made it look like an acquired gene from autosomes (85). In the cat MSY, the *CDC71LY* and *TETY1* derived from autosomal homologs from chromosomes A2 and A3, respectively. And in horses, there are 10 acquired genes reported on the MSY, which are homologous with genes located on different autosomes (43). In addition, MSY genes can arise from recurrent X-Y transposition events, for example, the *TGIF1LY* and *PCDH11Y* in humans (44), *SRSHY* and *ASSFY* in horses (43), and *OFD1* for dogs (55). These genes, called X-transposed genes, are the youngest on the MSY and display high similarity with X-linked gametologs in sequences.

The MSY genes are often classified into two categories of gene expression: broad and testis-specific. MSY genes are conserved in expression patterns for humans, mice, dogs, and horses, except for the *UBE1Y* which is broadly expressed in dogs, but testis-specific in humans and horses (**Table 1.1**). The ubiquitously expressed genes belong to S3

in most cases and appear to remain functional across various species. This phenomenon of independent functional preservation was termed "convergent survival" by Daniel et al. (1). The absence of Y-linked gametologs could lead to a broad difference across the body between XX and XY individuals when genes are ubiquitously expressed. Thus, both X-linked gametologs and Y-linked gametologs encode proteins that are at least partially redundant in function. It is believed that ubiquitous-expressed genes are selected to maintain the ancestral dosage of X-Y gene pairs, which may take functions in the regulation of transcription, translation, and protein stability (1,86). Testis-specific genes are mostly from S1 and S2 and occur as multiple copies on the MSY; these are related to male fertility and sex-determination (**Table 1.1**). Male-fitness genes were accumulated on the MSY with gene expansion events, and intrachromosomal recombination of the genes within palindromes preserves testis-specific genes from genetic decay (62). Additionally, *AMELY* exhibits expression exclusively in teeth and demonstrates remarkable longevity, being retained on the MSY in various species, including humans, horses, bulls, pigs, and cats. The phylogenetic analysis indicates *AMELY* arose independently across species, and the *AMELY* had respective gene conversion events between X and Y in different lineages of mammals. This convergent survival is essential in some species, and future explorations of *AMELY*'s roles are needed.

Table 1.1 The expression profile of S1, S2, and S3 genes in humans, mice, dogs, and horses.

Gene name	Human[1]	Mouse[2]	Dog[3]	Horse[4]	Stratum	Consensus expression	Functions
HSFY	Testis	Lost	Testis	Testis	1	Testis	Spermatogenesis
SRY	Testis	Testis	Testis	Testis	1	Testis	Sex-determining
RBMY	Testis	Testis	Testis	Testis	1	Testis	RNA splicing in germ cells
CUL4BY	Lost	Lost	Testis	Testis	1	Testis	NA
RPS4Y	Testis	Lost	Lost	Lost	1	Testis	NA
KDM5D	Broad	Broad	Broad	Broad	2	Broad	Demethylating the

							active transcriptional marks H3K4me3 and H3K4me2
TSPY	Testis	Lost	Testis	Testis	2	Testis	Sperm differentiation and proliferation
SYPY	Testis	Lost	Lost	N. E.	2	Testis	NA
UBE1Y	Lost	Testis	Broad	Testis	2	Testis	Spermatogenesis
UTY	Broad	Broad	Broad	Broad	3	Broad	Mediating protein-protein interaction
DDX3Y	Broad	Broad	Broad	Broad	3	Broad	Initiating translation in the cytoplasm and cell cycle progression at the G1-S phase
USP9Y	Broad	N. E.[6]	Broad	Lost	3	Broad	Cleaving the ubiquitin moiety from ubiquitin-fused precursors and ubiquitinylated
BCORY	Lost	Lost	Testis	Testis	3	Testis	NA
TAB3Y	Lost	Lost	Lost	Broad	3	Broad	NA
ZFY	Broad	N. E.	Broad	Broad	3	Broad	Activating transcriptional events
EIF2S3Y	Lost	Broad	Lost	Broad	3	Broad	Spermatogonial differentiation
EIF1AY	Broad	Lost	Broad	Broad	3	Broad	Stabilizing the binding of the initiator Met-tRNA to 40S ribosomal subunits
CYorf15	Broad	Lost	Broad	Broad	3	Broad	Syntaxin binding activity
AP1S2Y	Lost	Lost	Lost	Testis	3	Testis	NA
TMSB4Y	Broad	Lost	Lost	Broad	3	Broad	Cytoskeleton organization

AMELY[5]	Teeth	Lost	Lost	Teeth	3	Teeth	Biomineralization during tooth enamel development
WWC3Y	Lost	Lost	Lost	Broad	3	Broad	NA

[1,2,3,4]The expression profile of MSY genes for humans, mice, dogs, and horses are based on previous studies (4,31,43,44,55,87).
[5]The AMELY gene is only expressed in the teeth enamel.
[6]No expression data to support the category.

1.1.2.4 Dosage compensation of the sex chromosomes

In mammals, females have two copies of the X chromosome, whereas males carry one X and one Y chromosome. The X chromosome is not instructive in terms of sexual identity. Dosage compensation ensures equalisation of gene products of X chromosomes between males and females in somatic tissues. Mechanistically, X-linked gametologs are initially expressed at twice the ancestral expression level (88), followed by randomly silencing one of the two X chromosomes, named X inactivation (89,90). This process, which was hypothesised by Ohno, guarantees gene expression of X chromosomes euploid with respect to autosomal production in both males and females.

Several studies in humans suggested not all X chromosome genes strictly follow the X inactivation. First, the PAR genes, which have two copies in both males and females, resemble autosomal genes in terms of their expression. Second, for X chromosome genes that have active Y-linked gametologs with ancestrally derived functions, there is no need for X inactivation (91). Third, a proportion of X-linked genes, not belonging to either group above, escape X chromosome inactivation. In humans, this proportion is variable, ranging from 15% to 66% (89,92,93). Overall, the mammal's sex chromosomes have evolved to achieve dosage compensation, enabling balanced expression for the X-linked genes regardless of sex-specific haploid (male) or diploid (female) representation. At the same time, the genes which escape from X chromosome inactivation may contribute to sexual dimorphism (93) (next section).

1.1.3 Function of Mammalian Y Chromosomes

1.1.3.1 Sex Determination

As early as 1959, researchers found that human sex determination is controlled by a gene on the Y chromosome (94,95). By 1990, studies in humans (96) and mice (97,98) had independently concluded that *SRY* is the sex-determining gene. In mammals, *SRY* was thought to be a single-exon gene containing a high-mobility group (HMG)-box DNA-binding domain (99). Subsequent work proved the existence of a second exon in mice that produces an alternative transcript (100). *SRY* is expressed exclusively in the supporting cells of the genital ridges, between 10.5 and 12.5 days post-coitum, a tissue which will give rise to the Sertoli cells of the testis (97,101,102). After that, *SRY*, coupled with the steroidogenic factor 1 gene (*SF1*), binds to an enhancer of *SOX9* called TESCO, as a consequence of *SOX9* upregulation in males, whereas the expression of *SOX9* is inhibited in females (103). Subsequently, *SOX9* autoregulates its expression through binding to TESCO (104,105), and the expression of these genes triggers the activation of downstream targets such as *AMH* (106), *VNN* (107), *PGDS*, and *CBLN4* (108). In summary, *SRY* is necessary for male determination in mammals and initiates the cascade of activities of male gonadogenesis during foetal development.

1.1.3.2 Spermatogenesis

As discussed above, some MSY genes are only expressed in the mammalian testis, leading to the assumption they function to support testis development and gametogenesis. These genes evolved new expression patterns as a consequence of their ancestral functions being replaced by their X-linked genes. Additionally, these genes survived across distant taxa convergently, indicating their presence on the mammalian Y chromosomes is essential to male-specific functions. Among these genes, the functions of *HSFY*, *RBMY*, *TSPY*, and *UBE1Y* are extensively studied.

HSFY belongs to the heat shock transcription factor (*HSF*) gene family containing a conserved domain that enables binding to specific regions of DNA called Heat Shock Elements (HSEs). *HSFY* was predominantly expressed in round spermatids in mice (109)

and occasionally in type A spermatogonia and Sertoli cells (110,111). In humans, *HSFY* has two copies, *HSFY1* and *HSFY2*, located within the palindromes of AZFb. Maturation arrest patients, whose spermatogenesis is interrupted before the final stage without damage of Sertoli or Leydig cells, expressed low levels of *HSFY* protein in spermatogenic cells (111) and deletion of the *HSFY* gene was suggested to cause idiopathic male infertility (112). In cattle, copy number of *HSFY* shows a positive correlation with sire fertility traits (113,114).

RBMY, an RNA-binding motif (RBM) gene, has multiple copies in primate species and mice, and a single copy in other mammals (**Figure 1.6**). The RBMY protein contains an RNA-binding motif that can participate in RNA splicing events in germ cells of testes (115–117). Human evidence shows that the deficiency of *RBMY* causes failure in male meiosis (118). In mice, Rbmy protein is a testis-specific splicing factor that is expressed during spermatogenesis (108). Underscoring its role, *Rbmy* knockout mice have major spermatozoan structural abnormalities (120). Also, *RBMY* is implicated in oncogenesis, where it has been reported to be aberrantly activated and highly expressed in male liver cancer hepatocellular carcinoma (HCC) tissues, somatic cancer cells, lung adenocarcinoma, and kidney renal papillary cell carcinoma (121–123). HepG2 cells with *RBMY* loss-of-function displayed a reduction of transformation and anti-apoptotic efficiency, implying *RBMY*'s normal functions include processing androgen receptor activity and protection against male hepatocellular carcinoma (124).

TSPY is one of the most prevalent multicopy genes on mammalian Y chromosomes and its copy number varies among and within mammal species (59,62,125,126). TSPY interacts with the SET/NAP1 domain forming the TSPY/SET/NAP1 superfamily. Previous studies demonstrated that TSPY/SET/NAP1 have a variety of functions including DNA replication, transcriptional modulation, chromatin remodelling, and cell cycle regulation (127–131). TSPY exerts specialised functions regarding promoting meiotic division, whereas its X-linked gametolog, TSPX, is suggested to have generalised functions serving as a housekeeping gene in cell proliferation (132).

Additionally, TSPY is involved in the oncogenic process, as its expression is observed in HCC and other tumour cells (133,134).

The *UBE1Y* gene (previously called *SBY*) also presents on the Y chromosomes of dogs, horses, and cats as multiple-copy genes. It encodes a ubiquitin-activating enzyme E1 that is 90% identical to its X-linked gametolog at the coding level, *UBE1X*. The testis-specific expression and conservation between marsupial and eutherian suggest *UBE1Y* as a candidate for male reproduction (135,136).

1.1.3.3 Ubiquitously Expressed Genes

The expression of ubiquitously expressed genes shows a tight correlation with that of X-linked genes in a multitude of tissues (87). Their convergent maintenance on the mammalian Y chromosomes implies that their functions are similar to X-linked gametologs obeying dosage sensitivity. In general, the functions of ubiquitously expressed genes are involved in fundamental biological processes regarding broad regulation of transcription, translation, and protein stability (1,31) (**Table 1.1**).

There are *in vivo* experiments to support the functional redundancy between X-linked and Y-linked genes. For example, functions of mouse *Utx* and *Uty* overlap in embryonic development (137), *USP9Y* is dispensable for spermatogenesis (138), *DDX3Y* is able to complement the deletion of *DDX3X* in neural development (139), and overexpression of *EIF2S3X* can compensate for *EIF2S3Y* loss-of-function, resulting in fully functional male sexual development and fertility (140).

While the aforementioned studies support the view that ubiquitously expressed Y-linked gametologs are functionally redundant with their X-linked counterparts, several studies obtained contradicting results. For example, *DDX3X* (141) and *DDX3Y* (142) are both essential for embryogenesis and male fertility respectively, and *EIF2S3X* cannot replace *EIF2S3Y* in mouse spermatogenesis (143). Thus, it is suggested that ubiquitously expressed genes, on one hand, are partially equivalent to X-linked genes in functions, mainly regarding fundamental biological processes. On the other hand, they likely evolved novel functions in specific tissues that cannot be replaced by X-linked genes.

1.1.3.4 Sexual Dimorphism

Sexual dimorphism is the systematic distinction between individuals of various genders within the same species in terms of non-sex traits. In the strictest sense, "dimorphism" suggests binary presentation of morphologies. Here, I use the term more broadly to include a variety of non-physical traits, physiology, and incidence of some hereditary diseases that differ between sexes. In mammals, sexual dimorphism can manifest as body masses in humans, appearance in mandrills and manes in African lions. Genetically, the sex chromosomes innately contribute to sexual differentiation. So-called "X chromosome escapee genes" are reported to be involved in sexual disparities in humans, including blood-flow (ischemia), cognition, behaviour, and pathophysiology (144–148). As discussed above, some Y-linked genes developed new functions contributing to gender-related differences. Sexual disparities in the brain (149), cardiac (150), and kidney development (151) related to Y-linked genes have been explored. Third, the biased expression between X-linked and Y-linked genes may yield sexual dimorphism related to some diseases. For example, *EIF1AY* is expressed more than *EIF1AX* in human hearts, potentially in association with diseases of the heart, many of which manifest more frequently or severely in males (87).

1.1.3.5 Diseases beyond reproduction

Both gain- and loss-of function mutations of Y-chromomsome genes are known to cause, or to be associated with various health conditions including Parkinson's disease (152), Alzheimer disease (153), a variety of cancers (154–157), and heart disease (8,158). Proteome analysis found that DDX3Y protein may play an important role in neuronal differentiation (159), while the Y-linked *KDM5D-4* is involved in fatty liver and cellular inflammation related to atherosclerosis and cardiovascular diseases (160). Also, the type of Y chromosome haplogroup is reported to be associated with prostate cancer (6). Overall, mounting data reveals that the Y chromosome not only confers male features but also has a significant impact on disease and cellular phenotypes.

1.2 Genome Assembly and Gene Annotation

1.2.1 *De novo* Assembly of Sequences

Reference genome assembly underpins the study of functional, evolutionary, and genetic processes for vertebrate species. *De novo* whole-genome assembly is a hierarchical process where the initial step is to create continuous but short stretches of sequence, called contigs. Scaffolds are structures formed by linking contigs together in their respective positions and orientations. This process involves utilizing additional information to accurately connect the contigs, resulting in a sequential arrangement. Scaffolds provide a framework that captures the relative order and orientation of contigs, contributing to a more comprehensive representation of the genome. The genetic linkage map, whose genetic markers are physically arranged in order on the chromosomes, is employed as a "backbone" to sort scaffolds into the correct order and orientation. The achieved chromosome-level assembly, together with the following annotation, is regarded as a reference genome for genetics and genomics studies.

The earliest approach used for genome sequencing was the shotgun strategy together with Sanger sequencing (161,162). The human (163) and mouse (164) genomes, released in 2001 and 2002, respectively, were the first examples of assembled vertebrate species by merging Sanger sequencing and genetic map technologies. A further revolution in sequencing technology was marked by the emergence of next-generation sequencing (NGS) in 2005 which produced massive throughputs of read quantity in a more cost-effective way (165). Just five years later, the Illumina HiSeq 2000 could generate DNA reads more than 10000 times faster than Sanger sequencing at less than 1/10000 of the latter's cost. However, one of the major limitations of NGS for the use of reference genome production was its short read lengths whose contig assembly was incapable of spanning complicated regions of a genome. Now, so-called third-generation sequencing technologies enabled longer reads to be sequenced, and this innovation allows more genomes to be finished completely.

1.2.1.1 Genome Sequencing Methods

As far back as 2005, the 454 system by 454 Life Sciences pioneered the market for NGS. Illumina started offering the most popular NGS technology in 2007, and they continue to lead the NGS platform market today. Other significant platforms based on various technologies include the Ion Torrent by Life Technologies in 2011 and the "sequencing-by-ligation" platform developed by SOLiD system in 2007. Illumina technology was the most widely used sequencer to assemble the whole genomes of vertebrate species before the emergence of third-generation sequencing technologies. Prior to Illumina sequencing, library preparation is a crucial step that involves various components: DNA fragmentation, DNA repair and end-polishing, and ligation of adaptors. Each step impacts on how well NGS works. Depending on the instrument system and the library, Illumina sequencing can be classified into single-end, mate-pair, and pair-end sequencing. Single-end sequencing has been deprecated since its low accuracy and shorter generated reads compared with the other two. Pair-end sequencing libraries are the common method used today because of the little amount of DNA needed and the simplicity of libraries made. In contrast, mate-pair libraries are relatively expensive in terms of DNA due to the low yield of big DNA circularization. However, mate-pair sequencing can offer useful information regarding larger structural changes (166). Pair-end Illumina sequencing can produce raw reads with a length from 2 X 50 bp to 2 X 300 bp with a base calling accuracy as high as 99.9%. Repeat content and segmental duplications pose a challenge for assembly with short reads, resulting in gaps and misassemblies (167). Now Illumina sequencing is generally used for polishing draft genomes assembled by third-generation sequencing (168).

Third-generation sequencing methods are designed for sequencing longer DNA molecules. The current commercialised applications are led by Pacific Biosciences (PacBio) (169) and Oxford Nanopore Technologies (Nanopore) (170). PacBio technology, called Single Molecule Real Time (SMRT), enables up to 20 Kb fragments to be sequenced using its Sequel and Sequel II systems. The SMRT method involves generating SMRTbells, which are closed, single-stranded DNA molecules. Each SMRTbell travels to the bottom

of a well, called zero-mode waveguide, where a single molecule's light emissions, emitted during DNA replication, are captured and read in a process (169,171). A "polymerase read" or Continuous Long Read (CLR) is the term used to describe the contiguous DNA replication carried out by the polymerase during sequencing. Given that this CLR read repeatedly circles the circular template, it is possible for it to contain sequences from adapters and numerous copies of inserts. The adapter sequences are removed from the CLRs during processing, leaving behind only the insert sequence, sometimes known as "subreads". Multiple copies of subreads produced from a single SMRTBell can then create a single, consensus sequence known as the "read of insert" or Circular Consensus Sequence (CCS) (171,172). The Sequel system relays on the CLR mode, which emphasizes the longest possible reads but has a high error rate. The Sequel II system was able to generate high-throughput HiFi reads using the CCS mode to provide base-level resolution with >99% single-molecule read accuracy. Most recently, PacBio released the newest platform, the Revio system, which increases read throughput 15 times over that of the Sequel II system.

The Nanopore sequencing uses a "nanopore," a protein pore that is nanoscale in size and inserted in a polymer membrane with electrical resistance and measures the signal changes as different nucleotides travel through the pore. The single-stranded DNA passes through the nanopore from the negatively charged side to the positively charged side under the control of motor protein in moving speed. Ionic current waves corresponding to the nucleotide changes are decoded using computational methods (173). Oxford Nanopore Technologies launched its first Nanopore sequencer, the MinION in 2014, followed by the commercialization of the GridION in 2017 and the PromethION in 2018. The GridION platform is a medium-scale sequencer that parallels five MinION flow cells together, and the PromethION is a high-throughput platform with 24 or 48 flow cells. Advancements both in library preparation chemistry and sequencing platform have led to significant improvements in DNA sequencing technology over the years. Before 2015, each flowcell (R6 and R7) produced less than 0.2 Gb (gigabases) of data, and the sequencing accuracy ranged from 60% to 70%. Over the course of four years, there have

been remarkable developments. R9.5 has substantially increased the output data to 2.5 Gb and the sequencing accuracy has increased by over 90% (173). The most recent R10.4.1 represents a significant milestone in sequencing technology, achieving an impressive 99.5% sequencing accuracy, very comparable with PacBio Sequel II and Illumina sequencing technologies. Besides, Nanopore technology has found application in the detection of DNA methylation patterns. By measuring changes in electrical current as DNA molecules pass through nanopores, it becomes possible to identify methylated and unmethylated regions along the DNA strand. This technology offers a direct and real-time approach to studying DNA methylation, providing insights into epigenetic modifications and their implications in various biological processes and diseases.

After sequencing genomic DNA (gDNA), scaffolding methods are needed to order and orient contigs to chromosomal levels. Optical mapping is one of the solutions for scaffolding, its name referring to a fluorescence microscope-based technique (174). In brief, extracted long DNA molecules are labelled at specific, common restriction enzyme sites and travel lengthwise into NanoChannel arrays. As the molecule passes along the channel, restriction sites are visualised by their fluorescent emissions and positioned resolved with respect to one another at a single pair resolution. Molecules and their labels are detected and imaged by optical mapping instruments, after which bioinformatic software is implemented to assemble contigs into scaffolds. The high molecular weight DNA molecules captured by modern optical mapping often span more than 250 Kb.

High-throughput Chromatin interaction (Hi-C) technology is a powerful molecular technique widely used in genome assembly to improve the scaffolding process. By identifying and capturing spatially close DNA segments, HiC helps establish long-range interactions (i.e. DNA-to-DNA contacts) between genomic regions. When integrated into the assembly process, the HiC data provides valuable information about the physical proximity of different sequences in the genome. Hi-C is now the gold standard for scaffolding to chromosome level and most widely used in vertebrate genome assemblies (175).

1.2.1.2 Genome Assembly Approaches and Tools

With steadily decreasing sequencing costs, the production of high-quality raw sequencing data is no longer the reason for genome assembly projects being impeded. Rather, efficient *de novo* assembly computational algorithms are demanded to keep pace with the rising availability of sequencing data. At present, computational strategies for third-generation sequencing data can be classified by one of two strategies: Overlap-layout-consensus (OLC) algorithm and de Bruijn graph (DBG).

The OLC algorithm is a traditional strategy of assembly whose steps include overlap, layout, and consensus (**Figure 1.7A**). During overlap, all pairwise matches between reads are calculated based on sequence similarity. Next a general layout is constructed according to overlaps and the best path is detected in various ways. Lastly, a consensus sequence is generated from multiple sequence alignments. *De novo* assembly tools such as Canu, Falcon, and hifiasm belong to this classic approach. Canu (176) is applicable for high-noise long reads featured as applications of *k*-mer weighting, sparse graph construction, and graph fragment assembly; it is often used with ONT data whose reads are error-prone. Falcon (177) is a diploid-aware genome assembler designed for PacBio long reads. Its hierarchical genome assembly process allows it to phase diploid genomes into primary and alternative haplotypes. The hifiasm (178) *de novo* assembler is designed for PacBio Sequel II HiFi reads and faithfully preserves the contiguity of all haplotypes. A promising feature of hifiasm is its ability to utilise Hi-C data directly during assembly.

The DBG method applies *k*-mer to decrease the complexity of overlapping computation and reduce memory utilisation (**Figure 1.7B**). The first step is to build the de Bruijn graph according to the adjacent connections of reads' *k*-mers. Then the best path that reaches each edge only once is selected to obtain the primary contig. Finally, the order and orientation of primary contigs are determined by remapping the raw reads on the contigs, and gaps between contigs are filled accordingly. Assemblers such as Flye (179) and wtdbg2 (180) are developed utilising the DBG algorithm. Flye generates disjointigs by picking arbitrary paths in a complicated repeat graph and constructs an

accurate single string from these disjointigs. The wtdbg2 introduces the concept of a Fuzzy-Bruijn graph based on *k*-bin to tolerate noisy reads and reconstruct the graph and contigs correspondingly, making it efficient in the calculation.

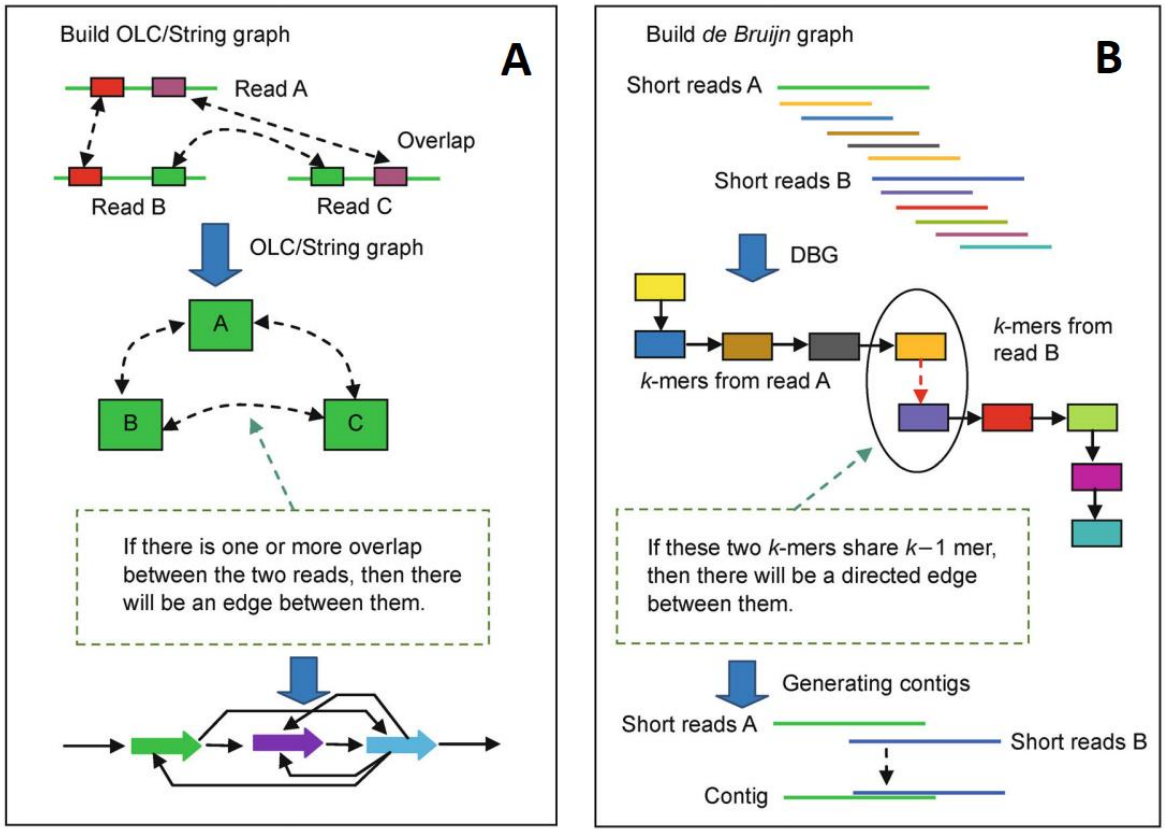


Figure 1.7. The illustration of the pipeline of de novo assembly. (A) The principle of building OLC graphs. (B) The principle of building DBG. This figure is modified from "Current challenges and solutions of de novo assembly" DOI: 10.1007/s40484-019-0166-9.

1.2.2 Gene Annotation

Coding genes are segments of DNA that provide instructions for building proteins, the essential molecules responsible for various biological functions. On the other hand, non-coding genes produce different types of RNA molecules that regulate gene expression and participate in vital cellular processes. Untranslated Regions (UTRs) are sections of genes that do not code for proteins but play crucial roles in mRNA regulation. They influence mRNA stability, localization, and translation efficiency. Promoters are specific regions at the beginning of genes that signal the initiation of transcription, the process of converting DNA into RNA. They provide binding sites for RNA polymerase and transcription factors. Enhancers, located at various distances from genes, serve as switches

that modulate gene expression. They interact with transcription factors to either enhance or repress the transcription process. These key functional features work in harmony to ensure precise control of gene expression and maintain the proper functioning of cells and organisms.

The availability of genome assemblies serves as a first step toward the categorization and understanding of functional elements within the genome. One of the key components of these elements is their definition of protein-coding genes. Gene annotation describes the physical locations of coding exons and UTRs, which are represented as gene models. Genome and transcriptome analyses, such as gene abundance estimation and regulatory element mapping, are all dependent on high-quality gene annotation (181). Also, as scientists turn to whole genome sequencing to discover genetic variation, gene models are crucial for functional annotation of genetic variants that contribute to complex traits or cause diseases. In this context, the location of variants with respect to all types of functional regions is crucial, as are the predicted effects of variants based on codon changes and regulation sequences.

The most popular and easily accessible sources of annotation are those provided by RefSeq and Ensembl using complex annotation pipelines and multiple sources of genomic datasets (such as RNA-Seq, EST, and ortholog coding sequences). RefSeq, short for Reference Sequence, is a comprehensive and curated database of reference sequences for genomes, transcripts, and proteins (182). It begins with data collection from various sources, including experimental submissions and computational predictions. The collected sequences are then aligned to a reference genome to identify coding regions, non-coding elements, and other features. Functional annotations are added through a combination of computational methods and expert curation, ensuring accurate representation of genes, transcripts, and other genomic elements. The main goal of RefSeq is to provide a standardized and well-annotated set of sequences that serve as a reliable reference for researchers and bioinformaticians.

Ensembl is a collaborative project involving several research institutions and is led by the European Bioinformatics Institute (EBI) and the Wellcome Trust Sanger Institute.

The primary goal of Ensembl is to provide a user-friendly and up-to-date platform for researchers to explore and analyze genomic information. Its annotation workflow integrates diverse gene prediction approaches, encompassing *ab initio* prediction, homology-based analysis, and RNA-seq data utilisation. These predictions are harmonised through consensus building, manual curation, functional annotation, comparative genomics, and rigorous quality control to produce a comprehensive and accurate gene annotation. As described previously, gene models, even from these respected sources, may differ substantially in their content (183). One of the efforts to solve the disagreement between RefSeq and Ensembl is called MANE (https://www.ncbi.nlm.nih.gov/refseq/MANE/). It aims to address disparities in human genome annotations by selecting a single representative transcript, known as the "MANE Select" transcript, for each protein-coding gene from RefSeq and Ensembl databases. This approach provides a standardised and accurate reference point for researchers and clinicians, ensuring more consistent interpretations of genomic data.

1.2.2.1 *ab initio* Gene Annotation

In the early 1990s, statistically-based *ab initio* gene prediction was the most widely used approach to profile gene structures (184–186). The term "*ab initio*" is derived from Latin and means "from the beginning" or "from first principles." In the context of gene prediction, it refers to the fact that this method starts from scratch, using only the intrinsic properties of the DNA sequence to make predictions. With the accumulation of proteins in the database of Pfam, Swiss-Prot, and Entrez, protein alignment-based tools including GeneWise (187) and PROCRUSTES (188) emerged. These tools utilize the idea that proteins exhibit a certain level of similarity and conservation among related organisms, allowing for the identification of gene-coding regions. These were followed by several hybrid programs, like GENOMESCAN (189) and AUGUSTUS (190), which combine *ab initio* and protein alignment methods to improve the accuracy and sensitivity of gene predictions. Other tools, including TWINSCAN (191), SPG2 (192), and DOUBLESCAN (193), ‘project’ coding sequences (CDSs) from one ‘mature’ genome to another on the basis of conservation and homology.

1.2.2.2 RNA Sequencing-Based Annotation

The commercial debut of short-read, high-throughput sequencing methods enabled genome-wide profiling of transcriptomic data (121,194). These were leveraged by Maker (195) and AceView (196) to analyse expressed sequence tags (ESTs). Subsequently, annotation software such as PASA (197), MAKER2 (198), and BRAKER (199) were developed. These use RNA-seq or ESTs/protein sequences and *ab initio* predictions to create gene annotations automatically. RNA-seq data is frequently combined with *ab initio* predictions to detect exon-intron structures and improve annotation quality (198,199).

RNA-seq data can be used independently to annotate genomes through transcript assembly. Alignment-based and *de novo* transcript assemblies are two major approaches for defining the transcriptome. For the alignment-based method, Cufflinks (200) and its successor, StringTie (201), use mapped reads to build a graph representing all possible pathways and traverse the graph to resolve individual isoforms. *De novo* methods use assemblers such as Trinity (202) and Rnnotator (203) to construct transcripts from overlapping reads. Afterward, proteins can be predicted from transcript sequences using either a 'screening machine' (204–206) to distinguish coding and non-coding transcripts or through use of specific software like TransDecoder (207) to identify open reading frames (ORFs) that align against the proteins database.

Despite these innovations, annotating the genes within eukaryotic genomes remains challenging, especially for gene structures with large numbers of exons, short exons, long introns, and non-canonical splice junctions. MAKER2, AUGUSTUS, and the annotation pipelines used by Ensembl and RefSeq all rely on the accuracy of *ab initio* predictions whose performance decreases in complicated genomes (208). RNA-seq-based transcript assembly using short reads is particularly prone to producing 'overhang' events, where two adjacent genes are mis-merged into one gene model (209–211).

1.2.2.3 Long Read RNA Sequencing-Based Annotation

The complexity of assembling and predicting genes from RNA-seq's short reads is circumvented by third-generation sequencing technologies such as Iso-seq and Oxford Nanopore RNA sequence; both technologies can routinely sequence full transcripts in a single molecule. In both PacBio and Nanopore pipelines for full-length RNA sequencing, the typical process involves converting mRNA into cDNA (complementary DNA) to facilitate sequencing. Many studies have discussed novel transcript detection with the aid of Iso-seq (212–215), and there are several pipelines (216–219) that combine full-length reads with existing tools to predict coding genes. Several studies have investigated the performance of Iso-seq and Oxford Nanopore data in genome annotation (214,217), and a few annotation tools are designed to use Iso-seq data, such as IsoAnnot Lite (https://isoannot.tappas.org/isoannot-lite/), cDNA_Cupcake (https://github.com/Magdoll/cDNA_Cupcake), and TAMA (220).

1.2.3 Genome Assembly of Mammalian Y Chromosomes

The difficulties of assembling mammalian Y chromosomes are due to their highly repetitive sequences and haploid nature. Particular strategies have been designed to target Y chromosome sequences distinct from autosome assembly. The methods can be classified into pre-assembly strategy and post-assembly strategy. The pre-assembly strategy includes two methods, Single-Haplotype Iterative Mapping and Sequencing (SHIMS) and chromosomal flow-sorting.

The SHIMS method was developed to resolve extremely repetitive sequences (83). First, nonredundant overlapping bacterial artificial chromosomes (BACs) are constructed and checked for mismatches, which are called sequence family variants (SFVs). Detection of SFVs between neighbouring BACs indicates this region has multiple copies. Polymerase chain reaction (PCR) of surrounding BACs coupled with Sanger sequencing used to determine the SFV patterns for each repeat unit and BAC sequences are sorted according to SFVs. BAC sequencing is conducted iteratively until all repeat units are connected together (221).

The principle of the flow-sorting method is to enrich and sequence isolated Y chromosomes. The technique uses intercalating dyes to label chromosomes. Due to its relatively small size, flow cytometry is able to sort and capture the Y chromosome. Sequencing library construction can be generated from the resulting pool of Y chromosomes to assemble the MSY independently (67). The enrichment of the DNA for Y chromosomes dramatically increases the sequencing depth of Y chromosomes and reduces non-Y chromosomal reads that mislead the assembly. However, the configuration of flow cytometry optics that is required for chromosome sorting is non-standard (personal observation), thus limiting the number of facilities that can support this approach.

The post-assembly strategy is to first assemble the whole genome sequences and filter out MSY scaffolds afterward. MSY sequences can be recognised in several ways. First, the short read depth of male samples on MSY sequences is theoretically half of the autosomal level, whereas female samples should lack coverage completely. Second, the Y-specific scaffolds can be determined by BLAST alignment results using MSY genes as queries. Thirdly, reads that correspond to the MSY can be filtered out by mapping them against a female assembly, as such reads should appear as unmapped. The disadvantage of this strategy is that MSY sequences can sometimes be erroneously combined with autosomal or X chromosome sequences during the assembly steps.

To date, there are complete MSY sequences for five species produced by SHIMS, including human (44), mouse (4), chimpanzee (66), rhesus (54), and bull (64). SHIMS provides good resolution of the ampliconic regions, but its application is labour-intensive. The flow-sorting method was conducted on gorillas (67), humans (222), pigs (69), mice (223), and alpacas (224), showing comparable assembly results with SHIMS. Similarly, the horse MSY was assembled using a combination of FISH and BACs sequencing (43). With the development of long-read sequencing, the post-assembly strategy can produce MSY sequences along with whole genome sequences. The MSY sequences of bonobos, orangutans (225), goats (226), sheep (227), foxes (228), and polar bears (70) were generated by this method.

1.3 Thesis Aims

With the creation of whole or almost complete Y chromosome sequences for a number of model mammals (primates and mouse), the evolutionary process, sequence characteristics, and gene content of the mammalian Y chromosome is becoming increasingly well understood. Mammalian Y chromosome evolution has been shown to be lineage-specific, with discrete gene content, a distinctive genomic structure, and novel evolutionary processes. As a model organism, our knowledge of dog Y chromosomes is currently restricted due to a paucity of sequencing data, incompleteness, and imperfect assembly. This species represents the fourth major branch -*Carnivora*- of the mammalian phylogenetic tree. Deciphering the dog Y chromosome permits more extensive comparative analysis among mammalian Y chromosomes, enhancing our understanding of the general evolutionary process of eutherian Y chromosomes, and providing insights into the dog-specific functions of the Y chromosome through an improved understanding of genes.

The major research objectives of this thesis can be summarised as follows:

1. Establish a workflow for assembling the dog's Y chromosome sequence and exploring its sequence elements and genomic structures.
2. Annotate genes on the dog's Y chromosome and investigate their expression profile and selective pressure.
3. Define and characterise the sex chromosomes' pseudoautosomal boundaries to ascertain between them, infer their origins, and explore the driving force of the inhibition of recombination between sex chromosomes.
4. Annotate a novel gene, called *PRSSLY*, accurately across mammals, and understanding its evolutionary process.

CHAPTER 2: Materials and Methods

2.1 Postmortem Sampling of Assembled Dog

The Schoenebeck group is authorised by the Royal (Dick) School of Veterinary Studies to conduct research on canine postmortems that are donated to the school. A 4 year old male Labrador retriever with severe osteoarthritis was humanely euthanized at the Royal (Dick) School for Veterinary Studies Hospital for Small Animals. The postmortem remains were scanned by computed tomography and dissected within 30 minutes of euthanasia to sample varying soft tissues (i.e. brain, muscle, lung, liver, heart etc.). Tissues were taken and immediately frozen in liquid nitrogen. For RNA applications, tissues were fixed with RNAlater solution (#AM7020, Invitrogen) guided by the manufacturer's manual. Liquid nitrogen-processed and RNAlater-processed samples were preserved at -70°C for further use. A modified procedure was also used to harvest skin fibroblasts from an ear sample (229). Low passage fibroblasts were cryopreserved in liquid nitrogen and post-thaw cells were propagated and used for sequencing library construction.

2.2 Genomic DNA Sequencing of a Labrador Retriever Dog

As **Table 2.1** shown, we conducted four technologies to sequence the whole genome of a Labrador retriever dog.

Table 2.1 Brief Description of Sequencing data generated in-house.

Sequencing platform	Purpose	Sequenced tissues	Sequenced dogs
PacBio Long Reads Sequencing	Whole-genome assembly	Soft palate	The Labrador retriever dog
Illumina Short Reads Sequencing	Sequence polishing	Soft palate	The Labrador retriever dog
Bionano optical mapping	Contigs scaffolding	Fibroblasts	The Labrador retriever dog

Dovetail® Hi-C	Contigs scaffolding	Fibroblasts	The Labrador retriever dog
PacBio Iso-Seq	Transcriptome profiling	Adrenal gland, Cerebellum, Liver, Testis, Tracheal cartilage, Lymph node	Each tissue pooled with three dog samples[1]
CAGE-Seq	Transcriptional starting site detection	Cerebellum, Liver, Tracheal cartilage, Lymph node, Heart, Bone marrow, Retina	Each tissue with three biological replicates[2]
Ultra-long Nanopore Sequencing	LINE_CF1 array detection	Fibroblasts	The Labrador retriever dog[3]

[1]See details in **Section 2.3.2**
[2]See details in **Section 2.4**
[3]See details in **Section 3.2.9**

2.2.1 PacBio Long Reads Sequencing

Snap-frozen tissue was sent to Novogene for extraction, library preparation, and sequencing. The service provider describes their procedure as follows: "Starting with approximately 0.5g of snap-frozen soft palate tissue, an SDS-based extraction protocol was used to recover high molecular weight genomic DNA (gDNA). Briefly, 0.14M NaCl-0.15M EDTA solution was added to the tissue sample to make the total volume 44 ml, then 3 ml of 25% SDS solution was added dropwise whilst stirring. The solution was then placed in a 60-degree water bath for 10 minutes with constant stirring. When the solution became viscous and slightly transparent, it was removed from the bath and cooled to room temperature. Next, 10 mL of 5M NaCl was added to the solution to reach a final concentration of 1M NaCl, stirring for 10 minutes. One volume of chloroform-isoamyl alcohol mixed solution was added, followed by shaking for 20 minutes, and centrifugation for 10 minutes at 4,000 rpm/min. The supernatant was decanted to isolate it from the precipitation. Approximately 1.5-2 times 95% ethanol was added to the supernatant slowly, resulting in DNA precipitation. Using a glass rod, DNA filaments were spooled and removed.

The following ethanol-based precipitation was repeated twice: The crude gDNA was added to 27 mL of 0.015 mol/L NaCl-0.0015M trisodium citrate solution, to which 3 mL of 1.5M NaCl, and 0.15M trisodium citrate solution was added with stirring. One volume of chloroform-isoamyl alcohol mixture was added, followed by shaking for 10 minutes. The solution was then centrifuged at 4,000 rpm/min for 10 minutes. The upper aqueous phase was decanted into a fresh tube, to which 1.5 volumes of 95% ethanol was added causing the DNA to precipitate. The solution was centrifuged and the resulting supernatant was discarded. Following the second extraction, the pellet was placed in 27 mL of 0.015M NaCl-0.0015M trisodium citrate solution, to which 2 volumes of 95% ethanol were added gradually while stirring. As described above, the filamentous DNA was transferred to a new tube and washed once with 70%, 80%, 95%, and absolute ethanol. The pellet was then dried under a vacuum and resuspended.

gDNA was fragmented using g-TUBE® (#010145, Covaris) under the Covaris® shearing instrument and gDNA from 500 bp to 10 Kb in size was selected. The prepared gDNA served as input for the SMRTbell Template Prep Kit v1.0, and the library was sequenced on a PacBio Sequel instrument using 18 SMRT cells. A total of 135.5 Gb of raw data was initially generated, with a reads N50 value of 19 Kb. Following this, a quality control process was implemented on the raw reads, which included the removal of adaptors and the exclusion of reads with low quality. Consequently, subreads, representing a genomic depth of 56.5 times, were employed for the purpose of genome assembly.

2.2.2 Illumina Short Reads Sequencing

Library construction and Illumina sequencing were performed by Novogene. gDNA for whole-genome sequencing was quality controlled using the Agilent TapeStation with DNA integrity numbers greater than 8. DNA libraries were constructed using Illumina TruSeq DNA nano kits (#20015964, Illumina) and then sequenced on an Illumina HiSeq X platform to > 90X depth with the paired-end library having an average insert size of 450 bp.

2.2.3 Optical Genome Mapping

Optical mapping was done by University of Nottingham's Deep Seq facility. They describe their protocol as follows: "High molecular weight gDNA (HMW gDNA) was extracted from ear skin fibroblasts using the Blood and Cell Culture DNA Isolation Kit (#80004, Bionano) and the Bionano Prep Cell Culture DNA Isolation protocol. Plugs were prepared, containing: 5 x 105, 1 x 106, or 2 x 106 cells/plug. The cells were processed as a single batch, according to the manufacturer's protocol and one plug was selected for DNA extraction. DNA quantitation, using the Qubit Fluorometer and the Qubit dsDNA BR kit (#Q32853, ThermoFisher), gave a mean concentration of 59.5 ng/ul (CV = 0.18).

Labelling was performed using the Bionano Prep Direct Label and Stain (DLS) protocol and 750 ng of HMW gDNA. The labelled sample was quantified by Qubit Fluorometer and the Qubit dsDNA HS Assay Kit (#Q32854, ThermoFisher). The average concentration of the labelled sample was 10.32 ng/ul (CV = 0.16). The labelling reaction was run over two flow cells on two separate Bionano Saphyr Chips (#20319, Bionano) on the Bionano Saphyr."

2.2.4 Chromosome Conformation Capture Sequencing

A high-throughput chromatin interaction (Hi-C) library was prepared from fibroblast nuclei using a Dovetail® Hi-C Kit according to the manufacturer's protocol (#21004, Dovetail Genomics). In the Hi-C process, formaldehyde is added to cells. This formaldehyde acts as a cross-linking agent, creating strong bonds between proteins and DNA that are close together in the nucleus. This "fixes" the chromatin structure in its current spatial organization, capturing the 3D arrangement of the genome. After cross-linking, the cells are lysed, and the nuclei are isolated. This step involves breaking open the cell membranes to release the nuclei while keeping the cross-linked chromatin intact. Fixed chromatin was digested with DpnII, the 5' overhangs filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed, and the DNA was purified from proteins. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was then sheared to ~350 bp

mean fragment size and a sequencing library was generated using Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of the library. The library was sequenced on an Illumina HiSeqX platform with paired-end, 150 bp long reads, yielding 145 Gbps (61x genome coverage) of sequence data. This work is finished by Jeffrey Schoenebeck (University of Edinburgh).

2.3 Long-read RNA Sequencing

2.3.1 Total RNA extraction from Frozen Tissues in RNAlater™

Prior to RNA extractions, RNaseZap™ (#AM9782M, Ambion) was used to decontaminate equipment and work surfaces. Less than100 mg of tissue was used for each extraction. Tissue was chopped into small pieces with a sterile scalpel in a sterile petri dish and tipped into a Lysing Matrix D tube (#MBR-247-110Y, Fisher Scientific). 1 mL chilled TRIzol™ (#15596026, Invitrogen) was added to the tube. FastPrep-24™ Classic bead beating grinder and lysis system (MP Biomedicals) were used to homogenise tissues. Soft tissues were homogenised twice at 2000 rpm/min for 20 seconds X 2 and hard tissues twice for 30 seconds at 3000 rpm/min for 30 seconds. To prevent heating between homogenisation steps, samples were cooled on wet ice to equilibrate to room temperature. Following homogenisation, tubes were left at room temperature for 5 minutes before 200 ul BCP was added. Upon addition of BCP, the tubes were vigorously shaken by hand for 30 seconds and incubated again at room temperature for 3 minutes. Samples were centrifuged at 12000 x g for 15 minutes at 4 °C and the upper aqueous phase was transferred to a fresh microfuge tube, taking care not to disturb or aspirate the interface. The RNeasy Micro Kit (#74004, QIAGEN) was used for purifying total RNA in the following steps: an equal volume of 70% ethanol was added into the microfuge tube before transferring the mixtures into the RNeasy Mini spin column. The loaded lysate was spun for 30 seconds at 10000 x g and the flow-through was discarded resulting in RNA molecules binding to the silica membrane and all contaminants being washed away. Wash buffer RW1 (700 µL) was added to the column and centrifuged at 10000 x g for 30 seconds, after which the effluent was discarded. Twice the columns were washed with

500 µL Buffer RPE, Each time the columns were spun for 30 seconds at 10000 x g to remove the wash. The residual liquid was removed by further centrifugation at 16000 x g for 2 minutes. After that, the column was placed in a new 1.5 mL microfuge tube and 50 µL pre-heated (45 °C) RNAse-free water was added. The column was incubated at room temperature for 2 minutes before centrifuging at 10000 x g for 1 minute to elute the RNA.

RNA integrity was measured using an Agilent 4200 Tapestation System (Agilent Genomics, Santa Clara, USA) with RNA Screen Tape assay (#5067-5577, Agilent), and the concentration was estimated using Qubit Fluorometer along with a Qubit RNA BR Assay kit (Q10210; Thermo Fisher Scientific).

2.3.2 Generating Library and Sequencing

RNA samples from three different dog breeds were pooled by tissue to maximise detection of tissue-specific transcripts (**Table 2.1**). In an effort to construct good quality Iso-Seq libraries, pooled RNA integrity number (RIN) values were >8.9 except for tracheal cartilage, whose RIN value was 7.2. To enrich for full-length transcripts, SMRT cell libraries were generated using TeloPrime Full-Length cDNA Amplification Kit V2 (#013.08, Lexogen), which selects the methylated G-capped transcripts in amplification. This is highly protected for full-length RNA molecules that are both capped and polyadenylated. A total of 8 SMRT cells were used to sequence libraries from six types of tissues: adrenal gland, cerebellum, liver, testis, tracheal cartilage, and scapular lymph node. To obtain these tissues, the same postmortem protocol and tissue preservation method was implemented on other dogs which were donated to the school. Both library construction and PacBio Sequel sequencing were done by Vienna BioCenter Core Facilities.

2.3.3 Raw Data Processing

Iso-seq raw data were processed with the PacBio official pipeline, IsoSeq v3, to generate clean transcript sequences for analyses that followed. First, the circular consensus subread program produced a consensus sequence for each read that had a complete adapter sequence. Second, the lima program identified and removed the 5' and

3' complementary DNA (cDNA) primers and oriented the reads from 5' to 3'. Third, the IsoSeq 'refine step' removed the polyA tail and artificial concatemers. Finally, the IsoSeq 'cluster step' performed conservative clustering of sequences to obtain polished transcripts.

2.4 Cap Analysis Gene Expression (CAGE) Sequencing and Analysis

In this thesis, the dog reference genome is the modified RosCfam1.0 where the Y-specific scaffolds (NC_051844.1, NW_024010443.1, and NW_024010444.1) were replaced by our dog Y chromosome sequences, RosY_1.0.

Cap Analysis of Gene Expression (CAGE) libraries were generated from seven tissues based on the Takahashi et al. protocol (**Table 2.1**) (230). Each tissue had three biological replicate RNA samples, each sample with a RIN value over 7.0. Libraries were created and multiplexing sequenced on Illumina HiSeq 2500 platform by the Wellcome Trust Edinburgh Clinical Research Facility. Details of CAGE-seq samples are included in the **Supplementary Table 2.1**.

For pre-processing, raw reads were demultiplexed with FastX toolkit version 0.014 (231), and adaptor sequences were removed using Tagdust (232). After removing adapters and barcodes, the cleaned reads had an average length of approximately 27 bp, and they were mapped to the modified ROS_Cfam_1.0 dog assembly using Bowtie2 v2.4.0 in very sensitive mode (-D 20 -R 3 -N 0 -L 20 -i S,1,0.50) (233). Multi-mapped reads were excluded from any following analysis. CAGEr package (234) was used to convert mapped BAM files to BigWig files, and the CAGEfightR package (235) normalised and clustered CAGE tags by importing BigWig files.

Besides the samples sequenced by us, additional CAGE data including dogs and wolves from the Dog Genome Annotation (DoGA) project were used. The data were analysed by Sabina Gansberger (Karolinska Institutet), and CAGE peaks were visualised with the ZENBU browser (https://fantom.gsc.riken.jp/zenbu/).

2.5 Polymerase Chain Reaction (PCR)

2.5.1 Reactions and Thermocycler Programme

Q5® Hot Start High-Fidelity 2X Master Mix (NEB, #M0515) was used for PCR experiments with standard reaction components (**Table 2.2**) and thermocycling conditions (**Table 2.3**). The annealing temperature was empirically determined by gradient PCR and the extension time was dependent on the size of the product being amplified. Primer design was facilitated using Primer3 (v4.1.0) web version and oligos were synthesised by ThermoFisher Scientific.

Table 2.2 Reaction system of Q5® Hot Start High-Fidelity 2X Master Mix. The composition and quantities of a KOD Xtreme™ reaction mixture.

Components	Volume
Q5 Hot Start High-Fidelity 2X Master Mix	12.5 uL
Forward Primer (10uM)	2 uL
Reverse Primer (10uM)	2 uL
gDNA	500 ng
H_2O	Up to 25

Table 2.3 Standard Thermocycler Programme. The temperature required for the annealing step and the duration of the extension phase are dependent on the specific reaction being performed.

Steps	Temperature	Time
Initial Denaturation	98°C	30s
30 Cycles	98°C	10s
	variable	60s
	72°C	60s/ kb
Final Extension	72°C	2min
Hold	10°C	---

PCR products were visualised with a 2% agarose gel. First, 100 mL 1X TAE was mixed with 2 g Ultrapure™ Agarose (16500500, Thermo Scientific) and heated in a

microwave (NN-E442W, Panasonic) for 2 minutes until dissolved completely. After the agarose solution cooled to about 50 °C, 10 μL 10,000X SYBR Safe DNA gel stain (Invitrogen, S33102) was added and mixed by swirling. After cooling to approximately 50 °C, the molten agarose was poured into a gel tray with the well comb in place and left to set until reaching room temperature (~30 minutes) or until the gel was completely solidified. PCR products were mixed with 6X loading buffer with a ratio of 5:1, loaded into the prepared agarose gel with known order, and 1 kb Plus DNA ladder (10787018, Fisher Scientific) was added in the last lane to serve as a size standard. The gel ran at 70V for 45 minutes in an electrophoresis tank filled with 1X TAE buffer. The gel was illuminated and imaged using a Gel Logic 200 Imaging System (Kodak).

2.5.2 Sanger Sequencing

To sequence target bands within the gel, the bands were cutout using a clean scalpel and purified using the QIAquick Gel Extraction Kit (28706X4, Qiagen) following the manufacturer's instructions. Prior to Sanger sequencing, the concentration and purity of collected DNA were tested by NanoDrop (Thermo Scientific). Purified DNA of 5 μl with a concentration of 1 ng/μl and 5 μM primers were mixed and sequenced with Mix2Seq Kits (Eurofins Genomics). Sequenced results were visualised by SnapGene software (www.snapgene.com).

2.6 Data Visualisation

In this thesis, bar graphs, box graphs, scatter plots, line graphs, and pie charts were drawn with 'ggplot2' and 'ggpubr' package in R 4.1.1 for linux, and GraphPad Prism version 5.0.0 for window. The Lucidchart (www.lucidchart.com) was used to create flowcharts, and BioRender (BioRender.com) was for schematic diagrams.

2.7 High-Performance Computer and Bioinformatic Tools

2.7.1 Eddie Computing Cluster

The University of Edinburgh's Eddie Mark 3 ("Eddie" for short) is a high-performance Linux computing cluster, which consists of over 7000 Intel® Xeon® cores with up to 3 TB of memory available on a single compute node. It can suitably increase computing efficiency by scheduling and running tasks in parallel. The cluster uses the Univa Gridengine batch system on Scientific Linux 7. All bioinformatic analyses presented here were conducted using Eddie.

2.7.2 Bioinformatic Tools List

All programs, tools, and packages being used in this thesis are listed as below. The scripts related to my thesis were uploaded to my repository, which can be accessed at https://github.com/WengangXbio/script_bio.

Table 2.4 Overview of computational tools or packages in this thesis.

Scope	Program	Usage	Homepage	Reference
Genome or transcriptome assembly tools	Falcon	Genome assembly (long-read)	https://pb-falcon.readthedocs.io/en/latest/	(177)
	Flye	Genome assembly (long-read)	https://github.com/fenderglass/Flye	(179)
	wtdbg2	Genome assembly (long-read)	https://github.com/ruanjue/wtdbg2	(180)
	Trinity	Transcriptome assembly	https://github.com/trinityrnaseq/trinityrnaseq/wiki	(202)
	ABySS 2.0	Genome assembly (short-read)	https://github.com/bcgsc/abyss	(236)
	Bionano Solve	Scaffolding contigs with optical mapping	N.A.	
Alignment tools	bwa-mem2	Mapping short reads to reference	https://github.com/bwa-mem2/bwa-mem2	(237)
	minimap2	Aligning DNA or RNA sequences to reference	https://github.com/lh3/minimap2	(238)
	HISAT2	Mapping DNA or RNA short reads to reference	http://daehwankimlab.github.io/hisat2/	(239)
	Bowtie2	Mapping short reads to reference	https://github.com/BenLangmead/bowtie2	(233)
	STAR	Mapping RNA-Seq data to reference	https://github.com/alexdobin/STAR	(240)
	BLAST	Aligning nucleotide or protein sequences to sequence databases	https://blast.ncbi.nlm.nih.gov/Blast.cgi	(241)

	MUMMER	Aligning between genome sequences	https://mummer.sourceforge.net/	(242)
	MEGA	Sequence alignment for DNA or protein sequences	https://www.megasoftware.net/	(243)
Genetic and genomic data manipulation tools	FastX toolkit	Preprocessing short-reads data	http://hannonlab.cshl.edu/fastx_toolkit/	
	Tagdust	Removing adapter and barcode sequences	https://tagdust.sourceforge.net/	(232)
	samtools	Reading, writing, editing, indexing, and viewing SAM/BAM files	http://www.htslib.org/doc/samtools.html	(244)
	bcftools (view, index and filter)	Reading, writing, filtering and indexing VCF files	http://www.htslib.org/doc/bcftools.html	(244)
	VCFtools	Summarising, filter, and merging VCF files	https://vcftools.sourceforge.net/	(245)
	PLINK	Summarising and quality control for genotype data	https://www.cog-genomics.org/plink2/	(246)
	fastp	Preprocessing short-reads data	https://github.com/OpenGene/fastp	(247)
	pairtools	Hi-C pairwise reads detection	https://pairtools.readthedocs.io/en/latest/	
	bedtools	Intersecting, merging, counting, and complementing on BAM, BED, GTF, and VCF files	https://bedtools.readthedocs.io/en/latest/	(248)
Variants detection and annotation	GATK	Variants discovery and quality control	https://gatk.broadinstitute.org/hc/en-us	(249)
	bcftools (pileup and call)	Calling short variants	http://www.htslib.org/doc/bcftools.html	(244)
	strelka2	Small variant caller	https://github.com/Illumina/strelka	(250)
	SnpEff v3.0	Genetic variants annotation	https://pcingola.github.io/SnpEff/	(251)

Genomic sequence analyser	RepeatMasker	Scanning for repetitive elements	https://www.repeatmasker.org/	
	EDTA	Annotating transposable elements	https://github.com/oushujun/EDTA	(252)
	Meryl	k-mer counter	https://github.com/marbl/meryl	(253)
	Merqury	Evaluating genome assemblies by k-mer	https://github.com/marbl/merqury	(253)
	gc_content.pl	GC content calculator	https://github.com/DamienFr/GC_content_in_sliding_window	
	CNVnator	Predicting genomic copy variations using short-read data	https://github.com/abyzovlab/CNVnator	(254)
	featureCounts	Calculating short-read depth for given intervals	http://subread.sourceforge.net	(255)
	mosdepth	Calculating short-read depth for chromosomes	https://github.com/brentp/mosdepth	(256)
Data visualisation	Dot	Viewing genome to genome alignments	https://github.com/marianattestad/dot	
	Circos	Visualising genome features in a circular layout	http://circos.ca/	(257)
	OMTools	Optical mapping data visualisation	https://github.com/TF-Chan-Lab/OMTools	(258)
	Galaxy HiCExplorer	Drawing heatmap for Hi-C data	https://hicexplorer.usegalaxy.eu/	(259)
	ZENBU	Visualising CAGE-Seq data	https://fantom.gsc.riken.jp/zenbu/	
	SnapGene	Viewing sanger sequence results	www.snapgene.com	
	Geneious Prime	Viewing DNA sequence alignment	https://www.geneious.com/	
Transcriptome	IsoSeq v3	Processing Iso-Seq raw reads to generate	https://github.com/PacificBiosciences/IsoSeq	

annotation		full-length cDNA sequence		
	cDNA_Cupcake	Annotating transcriptome based on Iso-Seq data	https://github.com/Magdoll/cDNA_Cupcake	
	Stringtie	Constructing transcriptome by RNA-Seq alignment	https://ccb.jhu.edu/software/stringtie/	(239)
	Spaln	Annotating gene models using protein sequences	https://github.com/ogotoh/spaln	(260)
	CAGEr	Converting BAM files into BigWig format for CAGE-Seq analyses	https://www.bioconductor.org/packages/release/bioc/html/CAGEr.html	(234)
	CAGEfightR	Calling peaks for CAGE-Seq data	https://bioconductor.org/packages/release/bioc/html/CAGEfightR.html	(235)
Evolutionary Analysis	KaKs_Calculator 2.0	Calculating substitution rates for coding genes	https://sourceforge.net/projects/kakscalculator2/	(261)
	AL2CO	Sequence conservation analysis	http://prodata.swmed.edu/al2co/al2co.php	(262)
	RAxML	Phylogeny and bootstrapping analyses	https://github.com/stamatak/standard-RAxML	(263)
	BEAST2	Constructing time-measured phylogeny	http://www.beast2.org/	(264)

CHAPTER 3: Dog Y Chromosome Assembly and Features

3.1 Introduction

Sex chromosomes were readily recognised due to their differentiation in size as early as 1941 (265). Metaphase karyotyping analysis revealed two positive Giemsa bands, indicating that the dog Y chromosome is metacentric (266,267). Notably, the Y chromosome is one of the five chromosomes in dogs that contains nucleolus organizer regions (NORs), as revealed by silver staining (268,269) and ribosomal gene (rDNA) probes (270). NORs are tandem arrays consisting of rDNA repeats that produce ribosomal RNA (rRNA) and are visualised under the Ag-AS staining (271). Like other mammalian Y chromosomes, the dog Y chromosome also contains pseudoautosomal region (PAR), X-degenerate regions, X-transposed regions, ampliconic regions, and centromeres. The dog Y chromosome was estimated as 27 Mb in length (26), 6.6 Mb sequences of which belong to PAR by computational analysis and Fluorescence In-Situ Hybridization (FISH) (272).

Tremendous efforts were made to draft the dog Y chromosome in order to support studies of male demographic history, sex development, and male-specific diseases. In the beginning, several DNA fragments unique to the dog Y chromosome were cloned and sequenced (273–276) including an *SRY*-related segment. After that, two versions of radiation hybrid maps were built to describe genetic linkage along the Y chromosome, containing 4 and 10 markers respectively (277,278). In 2006, ~24 Kb sequences belonging to the male-specific region of the Y chromosome (MSY) were identified by comparing male and female dog genomes. By the year 2013, 2.46 Mb MSY sequences from a Doberman dog genome were assembled by fluorescent *in situ* hybridisation (FISH) to order and orient sequencing data generated from 454-based short read sequencing bacterial of artificial chromosomes (BAC) clones that originated from the Y chromosome (55). In 2019, the MSY was further extended when a male Basenji dog genome was

assembled from data generated by the PacBio RSII platform, resulting in the release of a 3.06 Mb scaffold consisting of seven contigs (279). For the wolf Y chromosome, Smeds et al. compared male and female short read coverage and identified 120 Y-specific scaffolds summing to 4.7 Mb (280). Another wolf's Y chromosome sequences (HG994382.1) was generated along with a wolf genome (mCanLor1.2), which was assembled with PacBio long reads, 10x Genomics, and Dovetail Hi-C data. This assembly is deposited in the National Center for Biotechnology Information (NCBI), but has no annotation to date.

For other orders of mammals, the Y chromosome sequences have been assembled completely for species, such as humans, mice, chimpanzees, rhesus, and bulls (4,44,54,64,66), while there is no complete Y chromosome assembly for carnivorans. To date, our knowledge of dog Y chromosomes is limited due to the lack of sequence data and incomplete assembly. In this chapter, I describe my testing and implementation of an improved assembly pipeline of the dog Y chromosome. The resulting assembly, hereafter named "RosY_1.0", reveals unique features of DNA compared with autosomes and the X chromosome. Furthermore, the genomic structure of the dog Y chromosome was explored to demonstrate its novel evolution with respect to other mammalian species.

3.2 Materials and Methods

3.2.1 Workflow of the Y Chromosome Assembly

As detailed in **Section 2.2**, sequencing and scaffolding data from a male Labrador retriever was generated by PacBio Sequel I, Illumina HiSeq X, Bionano Saphyr optical mapping, and Dovetail Hi-C. A novel pipeline was developed to generate relevant contiguous and accurate MSY sequences by integrating all the sequencing data produced by the aforementioned platforms (**Figure 3.1**). First, the dog genome was assembled with Falcon (177), Flye (179), and wtdbg2 (180) using the PacBio data independently. The assembly process for these three assemblers assumes a haploid genome size of 2.5 Gb. With the aid of optical mapping data, fragmented contigs were scaffolded with Bionano Solve software. The Y-specific scaffolds were filtered out by aligning the assembled contigs with the known dog or wolf Y-specific sequence (KP081776, CM016470.1,

HG994382.1, CAJNRB020000035.1, CAJNRB020000036.1, and CAJNRB020000005.1) using the minimap2 (238). The Y-specific contigs of the Falcon assembler were selected as the backbone of the dog Y chromosome after the three assemblers' sequences were evaluated, and the Flye assembly was used to fill gaps and extend contig boundaries, which were unsolved by the Falcon. This step was achieved using quickmerge (281). A polishing step was conducted by Pilon (282) to correct nucleotide errors that were caused by the high error rate of long-reads sequencing. Finally, the polished scaffolds were sorted and orientated by Dovetail Hi-C data *in silico*. The Hi-C library construction was conducted in-house and scaffolding step was performed by Dovetail Genomic service with HiRise software.

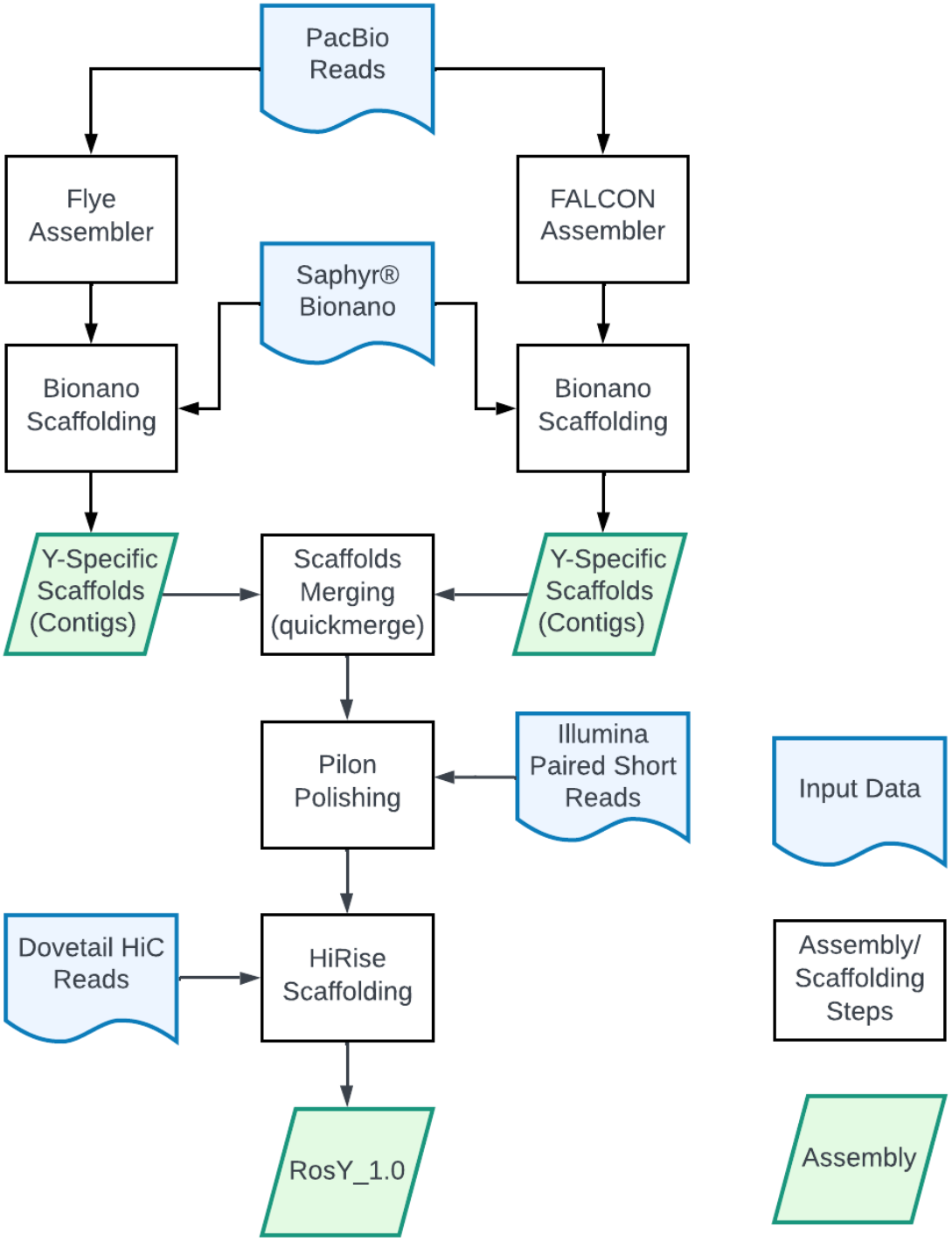


Figure 3.1. Dog Y chromosome assembly workflow.

3.2.2 Quality Assessment of RosY_1.0

Male-specific Illumina short reads (**Section 2.2.2**) were retained by mapping them against two female dog genome assemblies, Canfam3.1 and UU_Cfam_GSD. The

remaining unmapped reads were presumed to originate from the MSY. The quality and completeness of RosY_1.0 was assessed based on k-mer abundance (here k=19) using Meryl and Merqury (253). Two k-mer spectra were generated in a reference free manner by mapping whole genome and female-unmapped short reads on the RosCam_1.0 and RosY_1.0 respectively.

The accuracy of the RosY_1.0 at the nucleotide level was investigated by calling variants using the short read data (**Section 2.2.2**) (**Figure 3.2**). Theoretically, Y-specific short reads should be consistent with the RosY_1.0 sequence assuming short reads were sequenced to 100% accuracy and the RosY_1.0 was assembled without nucleotide error. Two approaches of variant calling, alignment-based and *de novo*-based, were implemented. The alignment-based method mapped short reads to the RosY_1.0 using bwa-mem2 (237) and variants were discovered using GATK's (249) germline short variant discovery pipeline. The *de novo*-based method assembled short reads using ABySS 2.0 (236) and detected Y-specific variants of using bcftools pileup and call tools (283).

3.2.3 Transposable Element Detection

Transposable elements (TEs) were annotated using a combination of two complementary tools: RepeatMasker 4.1.1 (https://www.repeatmasker.org/) and EDTA v2.0.0 (252). RepeatMasker scanned and compared against its TE libraries which represent known TEs identified in various model organisms. The latter relies on the benchmarking results of a collection of TE annotation methods and predicted unknown TEs with good control of the false discovery rate. Firstly, EDTA and RepeatMasker annotated TEs independently, which are called EDTA-TEs and RM-TEs respectively. Secondly, EDTA-TEs that overlapped with RM-TEs were pruned out, and the remaining EDTA-TEs were aligned using BLAST (241) against dog-specific TEs curated by Repbase (284), to categorise them as LINEs, SINEs, or others. Finally, overlapping or nested EDTA-TEs were only kept for the longest ones and TE copies with lengths lower than 100 bp were dismissed.

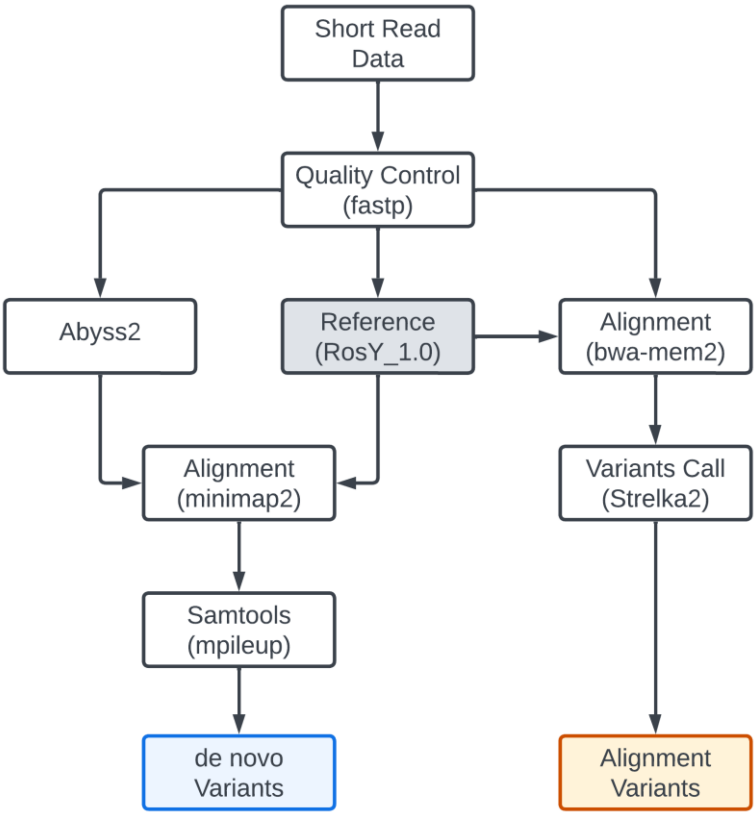


Figure 3.2. Assessment of the RosY_1.0 sequences in accuracy.

3.2.4 GC Content Calculation

GC content was calculated in using 1 Kb sliding windows and 500 bp step size. This analysis was finished by a python script (https://github.com/DamienFr/GC_content_in_sliding_window.git). To compare GC content among different sequences, a box plot was created, and a t-test was conducted using the ggpubr package in R.

3.2.5 Copy Number Variation Discovery

For discovering the complexity of the RosY_1.0, the assembled dog's Illumina short reads (**Section 2.2.2**) were aligned on the modified genome sequences using bwa-mem2 (237), and copy number variation (CNV) was estimated based on the read depth with the CNVnator (254) using a bin size of 150.

Additionally, I investigated the LINE1 array of MSY by analysing CNV in a total of 186 males and 29 females sourced from the Sequence Read Archive (SRA) (see **Supplementary Table 3.4** for details).

3.2.6 Similarity Sequence Detection and Syntenic Plotting

Similarity within the assembled dog genome (the RosCam_1.0), such as Y-X, Y-autosomes, and Y-Y, was analysed. Y chromosome sequences were used as a query in minimap2 alignment analysis and calculated based on a 5 Kb window every time with a sliding window of 2 Kb. Extracted sequences were aligned against the reference genome using minimap2 (238) with a parameter of -k19 -w19 -U50,500 -g10k -A1 -B4 -O6,26 -E2,1 -s200. The similarity level for each window was calculated by using the length of matches divided by 5 Kb.

In comparative genomics analysis, pairwise genome alignment between the RosY_1.0 and other Y chromosome assemblies was generated with MUMMER (285) and the Dot viewer (https://github.com/marianattestad/dot) was used to visualise the syntenic regions.

3.2.7 Bionano Data Visualisation

Bionano data were visualised by OMTools (258) to explore the complicated region in scaffold chrY2 including a 0.26 Mb gap. *In silico* enzyme label density was estimated based on the DNA sequences and matched with the labelling patterns generated by optical mapping. The labelling pattern within RosY_1.0 gap was established from a single Bionano scaffold whose aligned termini corresponded to sequence flanking the gap. The restriction sites within the "bridging" portion of the Bionano scaffold reflected the repetitive units of this region.

3.2.8 Hi-C Data Visualisation

Hi-C data were visualised in two steps where the pairtools (https://pairtools.readthedocs.io/en/latest/) conducted detection of pairwise Hi-C data and the Galaxy HiCExplorer (259) plotted pairwise contacts to create a heatmap. First, Hi-C

reads were mapped using bwa-mem2 with the option of -SP to align paired reads independently. Then, read contacts were extracted, sorted, and deduplicated, and pairs were further filtered by mapping quality in both ends (mapq > 0). After contact information files were converted into a cooler file format (286) (https://github.com/mirnylab/cooler), Galaxy HiCExplorer plotted the Hi-C heatmap with the log transformation of contact count values (259). Scripts related to this section are available within my repository.

3.2.9 Ultra-long Nanopore Sequencing

Ultra-high molecular weight (uHMW) genomic DNA from the Labrador retriever dog was isolated from fibroblast cells using the Monarch® HMW DNA Extraction Kit for Tissue (New England Biolabs, T3060). The extracted uHMW gDNA was subsequently sequenced using the PromethION 24 platform, employing the Ultra-Long DNA Sequencing Kit V14. Genomic DNA was extracted by Jeffrey Schoenebeck, and other procedures were conducted by Edinburgh Genomics at the University of Edinburgh.

In total, three flow cells were utilized for sequencing, and each flow cell yielded an average of 120.64 Gb of data. The generated data had an estimated N50 value of 58.24 Kb. This dataset was specifically employed to resolve the LINE_CF1 array and its boundaries on the Y chromosome.

3.3 Results

3.3.1 RosY_1.0 Statistics

Generally, 7.57% of nucleotides were corrected in the Pilon polishing step, indicating an accuracy of 92.43% for primary assembly based on PacBio raw reads. Data used to generate the ROS_Cfam_1.0, detailed in **Section 2.2**, was reanalysed to produce a high quality Y chromosome assembly. Using the PacBio data, three *de novo* assemblers were evaluated for their ability to produce contigs corresponding to the Y chromosome. Falcon, Flye, and wtdbg2 each independently generated contigs belonging to the Y chromosome whose summed lengths were 5.10, 5.28, and 4.84 Mb, respectively. Next, the

PacBio primary contigs were scaffolded with optical mapping data. Comparison among the assemblers indicates that Falcon produces the longest MSY contigs, while Flye generated the most contiguous scaffolds (**Table 3.1**). Because FALCON-Unzip was able to generate haploid-specific contigs ("haplotigs"), it enabled localizations of the transition region between the PAR and either the MSY or female-specific X region (FSX) in scaffold chrY1. Also, FALCON assembled longer contigs than other assemblers in the repetitive regions of the scaffold chrY2 (**Table 3.1**). Assembly by Flye was advantageous insofar as it produced fewer and longer contigs. Wtdgb2 produced the shortest sequences compared to the other two assemblers. It lacked contiguity like Flye sequences and also could not resolve the transition region like FALCON (**Table 3.1**). There were no discernible advantages of using wtdgb2.

Working in parallel, contigs produced by FALCON and Flye were scaffolded using optical mapping. Falcon-based data, specifically three Y-linked scaffolds, provided the foundational sequence of the Y chromosome, to which Flye scaffolds were used to patch gaps and extend the ends of the FALCON-based scaffolds (**Figure 3.3**). This step is achieved by quickmerge. The patched scaffolds were polished with Illumina short reads. The order and orientation of the scaffolds were guided by interaction patterns revealed by Hi-C data (**Figure 3.4**). The sequence that corresponded to the PAR was trimmed according to the interaction map between the X chromosome and the Y-specific scaffold (**Figure 3.5**). By leveraging the strengths of FALCON and Flye assemblers, a hybrid dog Y chromosome assembly was generated, which we called "RosY_1.0" (**Figure 3.1**). As a result, the total length of the RosY_1.0 was 6.78 Mb, consisting of three Y-specific scaffolds with three gaps (**Table 3.1**).

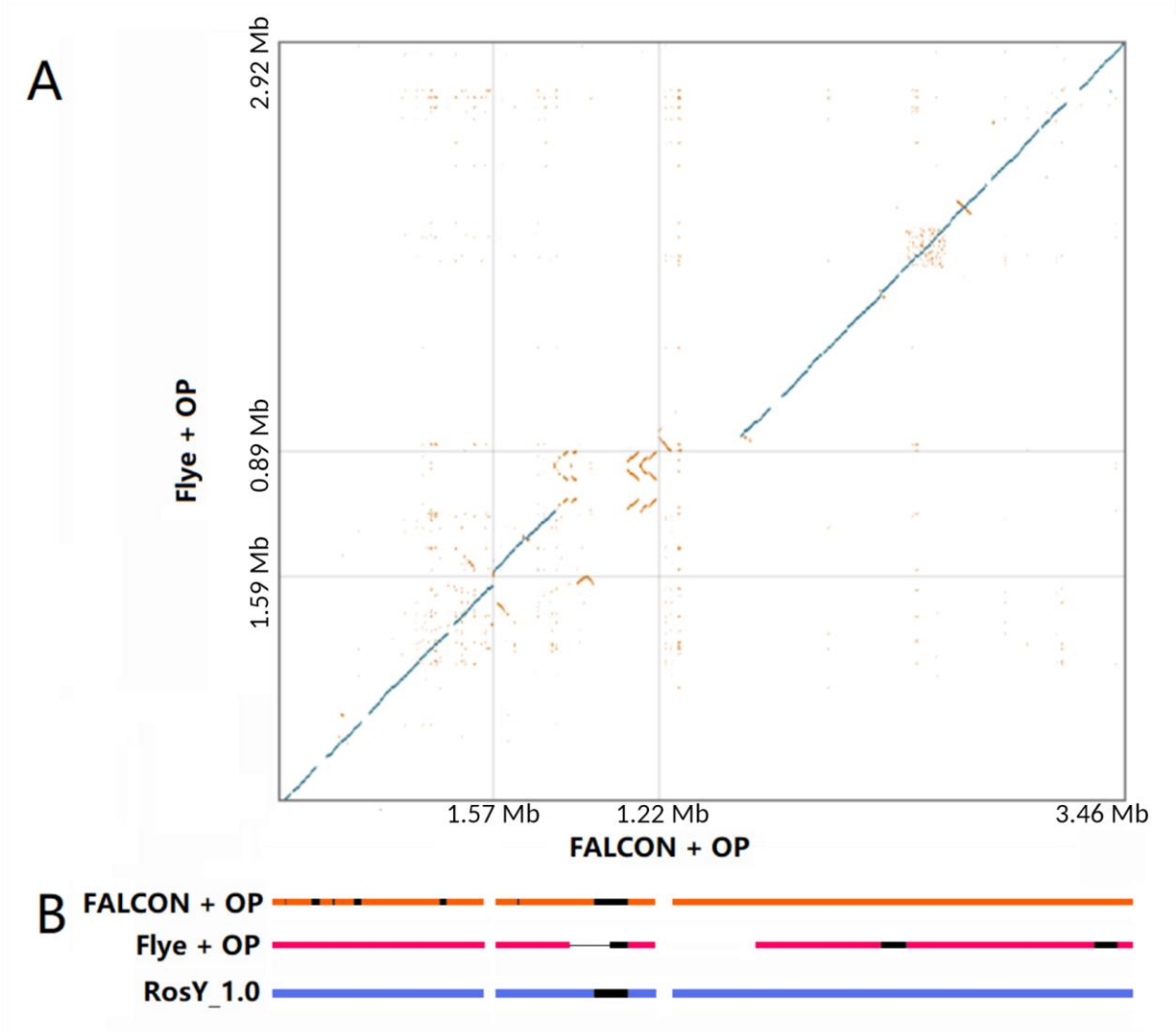


Figure 3.3. Hybridising the Falcon and Flye assemblies generated an improved dog Y chromosome. (A) The genome dot plot displayed collinearity between Flye and FALCON assemblies for the Y chromosome. The blue dot represented a unique alignment, and the orange represented a repetitive alignment. (B) Diagrammatic depiction of the Y chromosome assemblies. Black areas indicated gaps and the thick line in the 'Flye + OP' assembly referred to missing sequences compared with 'FALCON + OP'. The 'FALCON + OP' indicated assembled sequences generated with FALCON and scaffolded with optical mapping. OP, optical mapping.

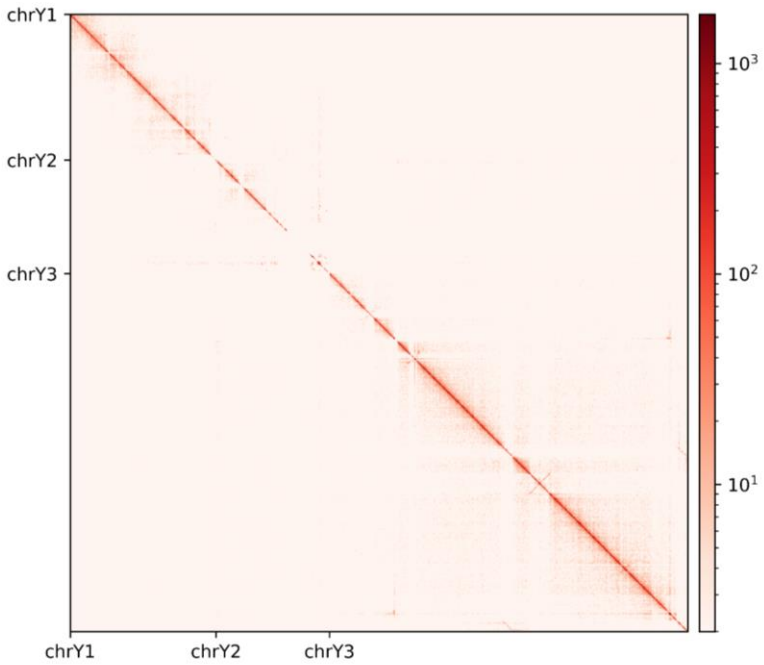


Figure 3.4. Hi-C interaction heatmap of Y-linked scaffolds. Three Y-specific scaffolds are constructed in order and orientation based on Hi-C interaction.

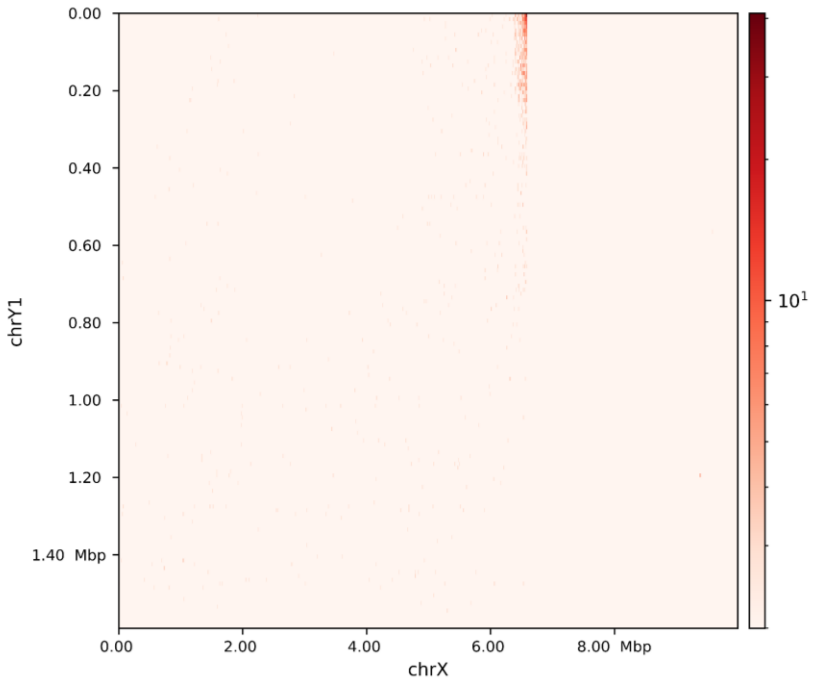


Figure 3.5. Hi-C interaction heatmap indicating the pseudoautosomal region. Heatmap shows Hi-C interaction between the X chromosome and Y-specific scaffold chY1. The enrichment of interactions at around 6.6 Mb of the X chromosome indicates that on the chrY1 scaffold, the sequences belonging to the pseudoautosomal region range from 0 to 6.6 Mb.

Table 3.1 Y chromosome assembly statistics.

Assembly	Contigs	Scaffolds	Length (Mb)	Gap	Gap length (Mb)	Scaffold 1		Scaffold 2		Scaffold 3	
						Length (Mb)	Gap	Length (Mb)	Gap	Length (Mb)	Gap
Falcon	21		5.10			1.38		0.96		2.06	
Flye	4		5.28			1.59		0.76		2.92	
wtdbg2	11		4.84			1.35		0.63		2.86	
Falcon + OP[1]		3	6.26	19	1.16	1.57	6	1.22	4	3.46	9
Flye + OP		3	5.41	1	0.13	1.59	0	0.89	1	2.92	0
wtdbg2 + OP		3	5.16	11	0.32	1.37	3	0.83	3	2.96	5
RosY_1.0		3	6.78	8	0.26	1.60	0	1.25	3	3.94	5

[1]Bionano optical mapping for scaffolding.

3.3.2 Assessment of the RosY_1.0 Assembly

3.3.2.1 Comparison with Y Chromosome

Comparisons between the RosY_1.0 and other Y chromosome assemblies showed high levels of synteny and collinearity (**Figure 3.6**). Basenji had fragmented contigs corresponding to scaffold chrY1 and chrY2 sequences and inversed sequences were observed for scaffold chrY3. For the Doberman Y chromosome assembly, there was no sequence that matched the scaffold chrY3 of the RosY_1.0, and inconsistent mapping was seen in scaffold chrY1 and chrY2. The wolf sequences displayed a near-perfect collinearity with the RosY_1.0 sequences and the alignment showed two copies of scaffold chrY3 sequences to be assembled in the wolf Y chromosome in a form of 'head-to-head' tiling. This alignment pattern indicated dogs and wolves' Y chromosomes may be different in sequences, or the collapsed assembly led to one copy of chrY3 sequences presented on the RosY_1.0.

3.3.2.2 K-mer Analysis

Plots of k-mer spectra enable visualisation of the distributions of sequence according to ploidy. Due to the RosY_1.0 hemizygosity, the single-copy k-mer multiplicity of MSY reads should be approximately half that of the autosomes' single-copy multiplicity of x=40 (**Figure 3.7A**). However, the single-copy spectrum of RosY_1.0 was bimodal, with k-mer multiplicity peaks at x=20 and 40 (**Figure 3.7B**). This observation suggests that one or more regions within the RosY_1.0 are collapsed (i.e. repeats with 2 copies are represented once with the assembly). Based on comparison with the wolf, the collapsed region corresponds with RosY_1.0's chrY3 scaffold (**Figure 3.6**). More subtly, RosY_1.0's two-copy k-mer spectra are also bimodal (blue spectrum, **Figure 3.7B**), indicating some repeat sequence is either assembled correctly or partially collapsed (i.e. repeats with >3 copies are represented as two copies). In the RosY_1.0, higher order spectra (k-mer copies >2) are also uniquely evident, as expected of k-mers that match three or four copy regions. Also of note, read-only k-mers, typically attributed to sequencing errors, occur infrequently among autosomal reads (grey) (x<10, **Figure 3.7A-**

B). However, in the RosY_1.0 a portion of read-only k-mers ranged between x=15-40 , as is indicative of sequences that were not incorporated into the RosY_1.0 assembly (**Figure 3.7**).

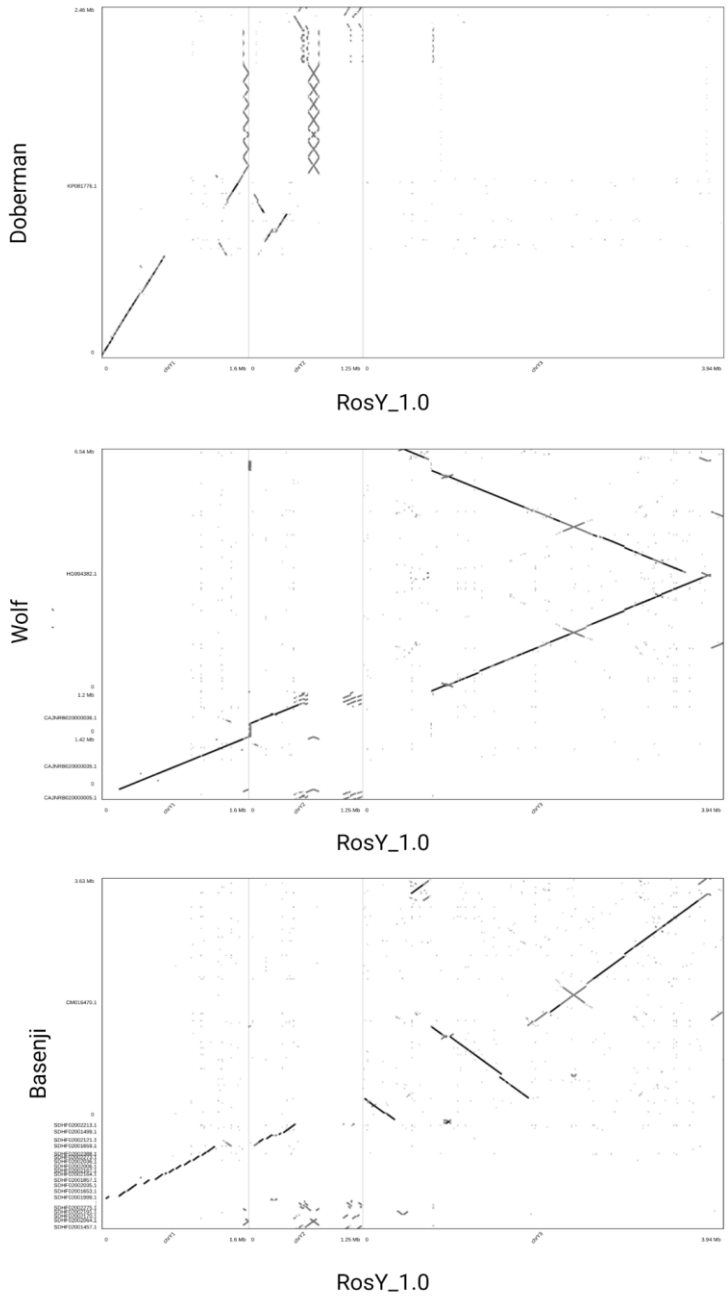


Figure 3.6. Alignment between the RosY_1.0 and previously generated Y sequences. Genomic dot plots for the RosY_1.0 (x-axes) compared with Doberman (KP081776) (top), Wolf (GCA_905319855.2) (middle), and Basenji (GCA_004886185.2) (bottom) Y-specific sequences (y-axes), respectively.

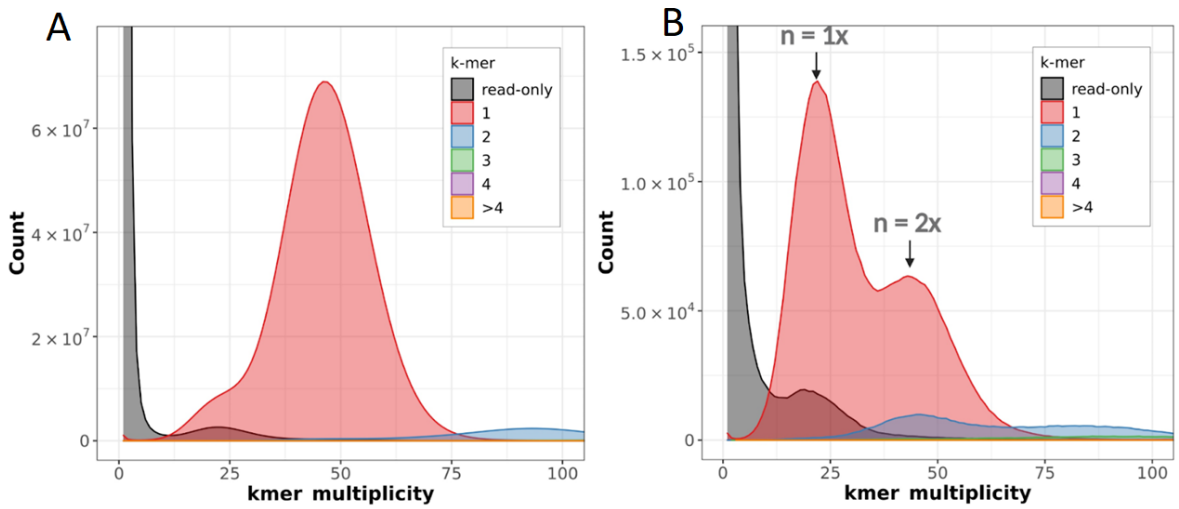


Figure 3.7. Assessment of the RosY_1.0 with a k-mer method. MERQURY spectra-cn plots comparing the whole genome pair-end reads to the (A) RosCfam1.0 and (B) Y-specific pair-end reads to the RosY_1.0.

3.3.2.3 Variants-based Assessment

De novo and reference-guided methods of variant calling using ROS_Cfam_1.0's whole-genome sequencing (WGS) enabled the quantification of long-read sequencing errors (**Section 3.2.2**, **Figure 3.2**). Overall, a total of 53 small insertion/deletions (INDELs) and 17 single-nucleotide variants (SNVs) were detected with at least one method, and 25 INDELs and 3 SNVs overlapped in both methods, which indicates the nucleotide error rate ranges from 0.02% to 0.05% (**Figure 3.8**). The actual error rate is likely lower; some discrepancies between the assembly and WGS reads are likely due to erroneous variant calls produced from the latter's short read data, particularly around homopolymers and simple repeats. Consistently, the indel discrepancies occur near homopolymers (83.1%), and simple repeats (16.9%) (**Supplementary Table 3.1, Supplementary Table 3.2**).

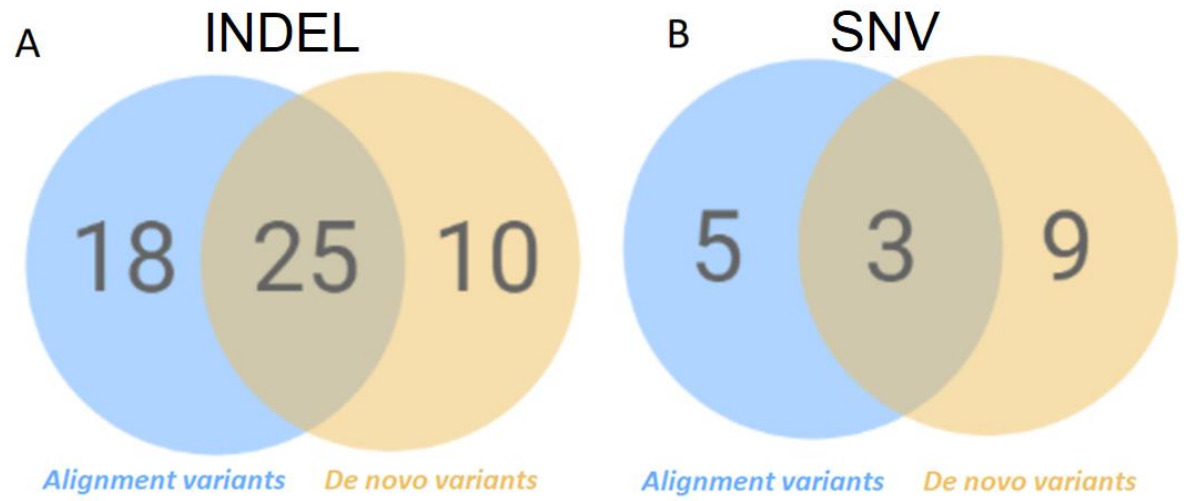


Figure 3.8. Assessment of the RosY_1.0 with a variants-based method. Venn diagrams of nucleotide errors detected by *de novo* (tan) and reference-guided (blue) approaches for (A) INDELs and (B) SNVs.

3.3.3 Overview of Y Chromosome Features

A circos plot was drawn to display an overview of the dog Y chromosome features and physical distributions (**Figure 3.9**). Read coverage showed the depth of short read sequencing by a 100 bp window and normalised with autosomal depth. As expected, the diploid PAR has double the read depth compared to scaffolds chrY1 and proximal chrY2. Scaffolds chrY1 and proximal chrY2 are single-copy regions of MSY, whereas the distal scaffold chrY2 had a depth of more than single-copy. For scaffold chrY3, the depth of most sequences was doubled that of haploid indicating these sequences were present twice (**Figure 3.9B**).

RosY_1.0 was largely composed of TE sequences (55%), a strong enrichment compared to PAR (26%) (**Figure 3.10A**). The enrichment of TEs on the RosY_1.0 was attributed to long interspersed nuclear elements (LINEs), which occupied 36% of the sequences. As seen, the proportions of LINE sequences in all three scaffolds of the RosY_1.0 were at the same level and higher than that of the PAR (**Figure 3.10B**). On the contrary, the densities of short interspersed nuclear elements (SINEs) in chrY2 and chrY3 were lower than those of the PAR, and chrY1 and the PAR exhibited no difference in density (**Figure 3.10C**). The TE sequences, which did not belong to either LINEs or SINEs,

differed among three scaffolds in density; chrY3 had the highest density in comparisons, and both chrY1 and chrY2 were lower than the PAR (**Figure 3.10D**).

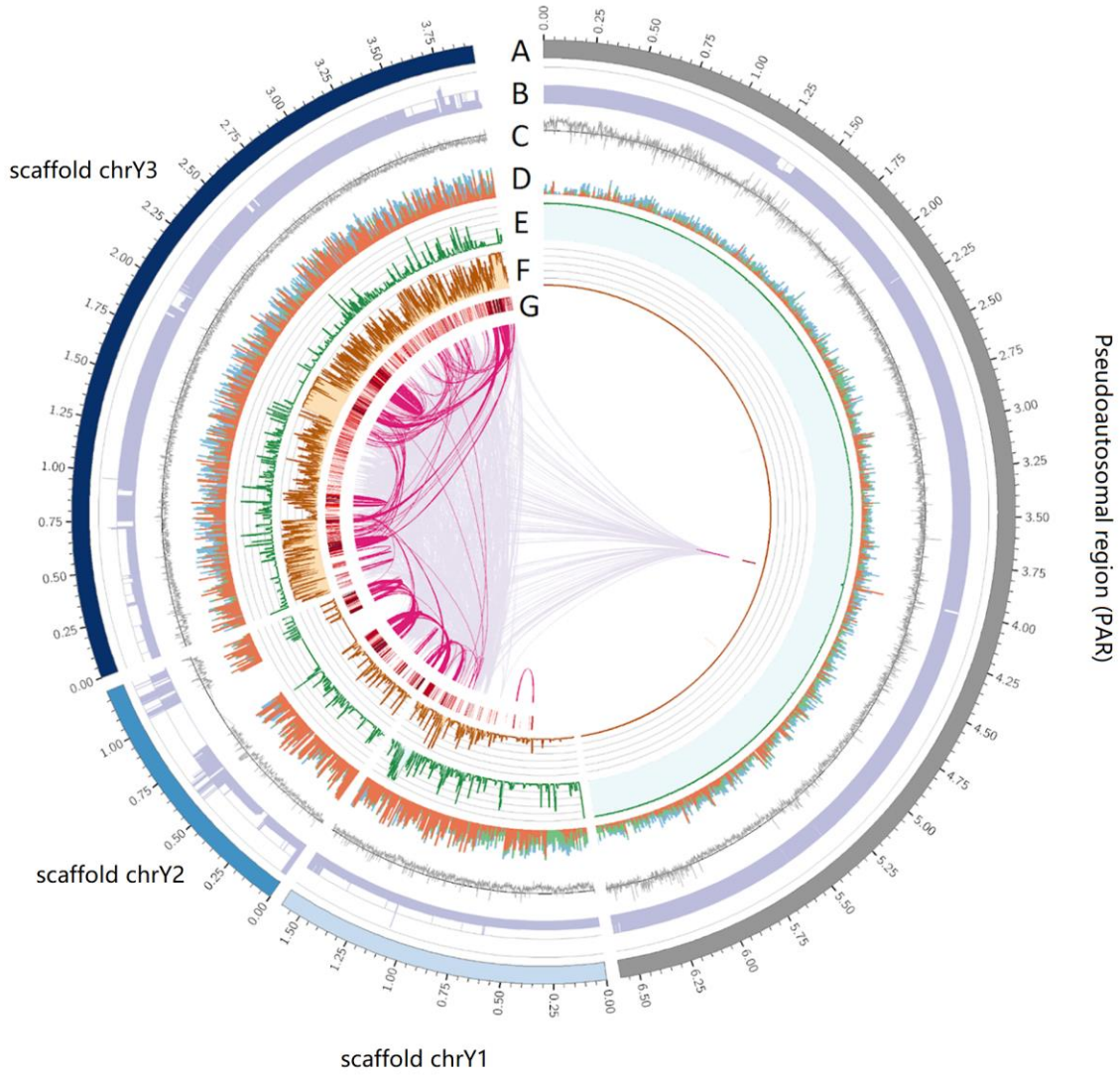


Figure 3.9. Circos plot depicting the dog Y chromosome genome assembly, RosY_1.0. (A) physical chromosome (Mb) of the dog Y chromosome, (B) genomic depth based on short reads, (C) GC content, (D) TE density coloured by category (orange: LINE; green: SINE; blue: others), (E) similarity with X chromosome, (F) similarity with autosomes, (G) self-similarity scale indicated by colour. The innermost layer shows inter-chromosomal synteny, including non-TE similarity (pink) and TE similarity (purple). PAR, pseudoautosomal region.

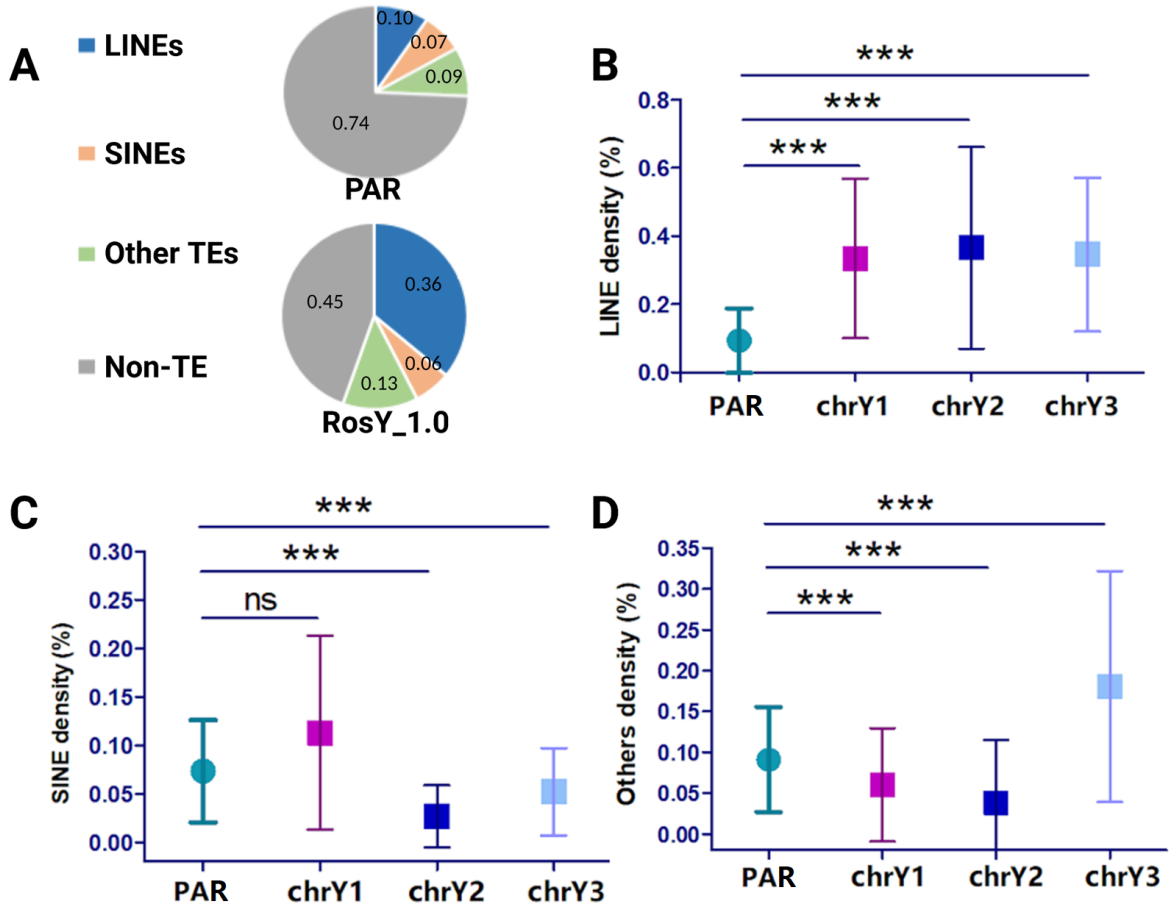


Figure 3.10. TE density classified by types. (A) Percentages of nucleotides containe in LINEs, SINEs, other TEs, and non-TE for the PAR and the RosY_1.0. The densities of LINE (B), SINE (C), and other TEs (D) are plotted in different regions (PAR, chrY1, chrY2, and chrY3). Compared to the PAR, the MSY is enriched for LINEs. For SINEs, the chrY2 and chrY3 sequences are lower than the PAB in density, and the chrY1 and PAR has no significant difference. The other TEs display fluctuating distribution in the MSY: the chrY3 is highest, followed by chrY1 and then chrY2. The PAR is lower than chrY3 for other TEs significantly but is higher than the chrY1 and chrY2. A t-test was used to examine the significance between groups. Asterisks refer to a difference between two groups in a statistic (*** indicates a P-value lower than $1x10^{-06}$).

Previous studies found that GC content was skewed towards high recombination regions in other species' PARs (287,288), and this finding was recapitulated in dogs: the GC content of the PAR was greater on average than all three MSY scaffolds with chrY1 displaying the lowest GC content (**Figure 3.11**).

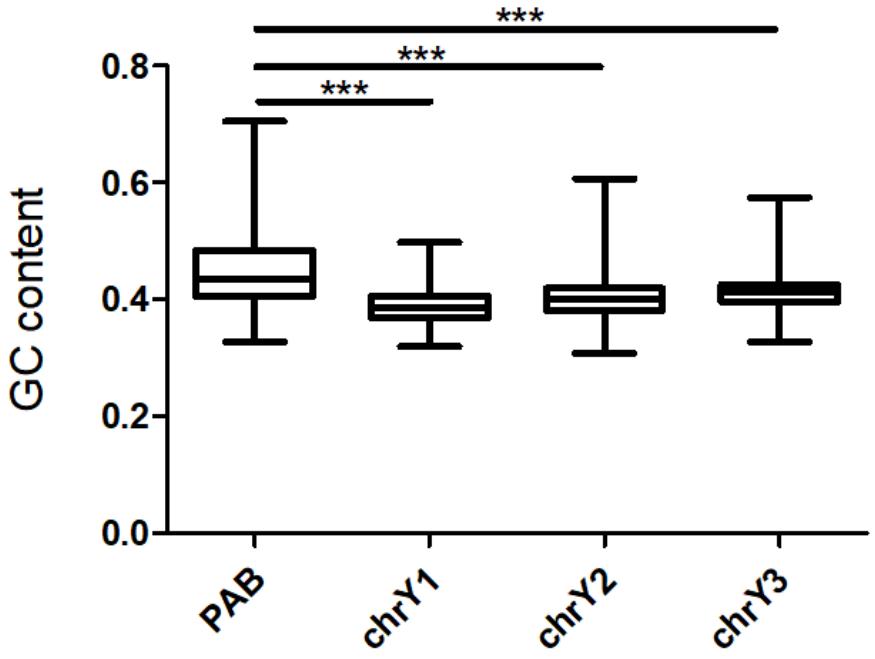


Figure 3.11. Box plot of GC content in different scaffolds. A t-test was used to compare GC content and showed the PAB is significantly higher than three MSY scaffolds in GC content.

Furthermore, the Y chromosome's interchromosomal sequence similarity with the X chromosome and autosomes was examined. The *OFD* translocation found at the distal end of scaffold chrY1 (~12.7 Kb) is highly similar to the X chromosome, as seen in earlier investigations (55). Other areas of notable interchromosomal similarity were found on the scaffold chrY3 (**Figure 3.9E**). This scaffold accumulated ~0.67 Mb of sequence with homology to the telomeres and centromeres of autosomes including CFA6, CFA15, CFA17, CFA19, CFA32, and CFA37 (**Figures 3.9F and 3.12, Supplementary Table 3.3**).

Intrachromosomal similarity was also assessed. During mammalian evolution, amplification events and TE insertions increased chromosomes' repeat features ubiquitously throughout the genome; it is particularly apparent within the MSY (4,62). To investigate the nature of the repeat content of the dog MSY, I highlighted non-TE repeated sequences by linking pairwise regions (**Figure 3.9G**). Multiple-copy gene loci were enriched with repeated segments indicating gene expansion events remained in massive ampliconic sequences on the dog Y chromosome (**Figure 3.13**).

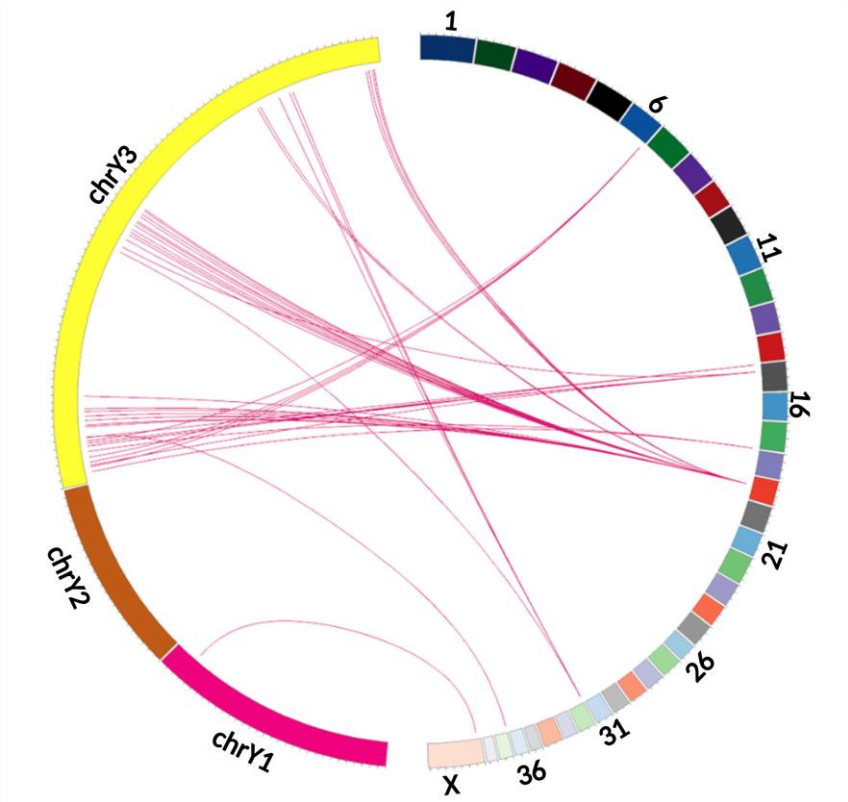


Figure 3.12. Syntenic sequences between the RosY_1.0 and autosomal sequences.

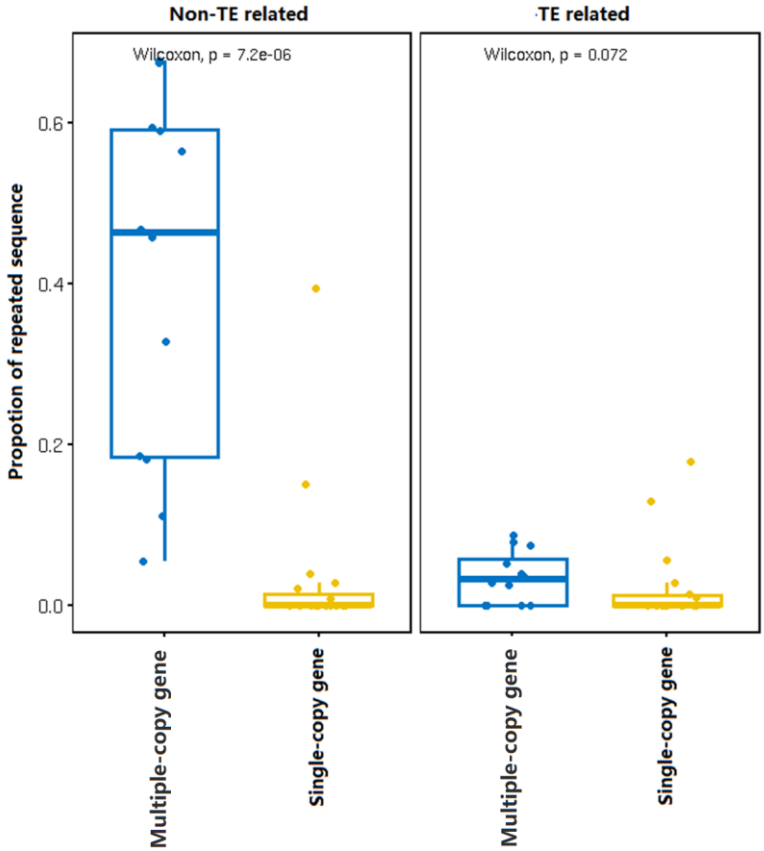


Figure 3.13. Non-TE self-similarity is significantly enriched by multiple-copy genes.

3.3.4 Structure and Complexity of the Dog Y Chromosome

3.3.4.1 Long-Range Gap in Scaffold chrY2

In scaffold chrY2, there was a 0.26 Mb gap, which was a potential repeated region missing in the RosY_1.0. This is because the depth was more than two in the flanking regions of the gap (**Figure 3.9B, Figure 3.14**); as well, the repeated feature was evident when it is close to the gap boundary (**Figure 3.9**). Therefore, we traced back the optical mapping data, which was contiguous over this region, and expected to find the sequence features in the gap region. By comparing restriction site distribution, four unique units were identified that are ampliconic in this region, either in the same or opposite directions (**Figure 3.14**). The gap with its flanking sequences represented palindrome sequences with several Kb in length (**Figure 3.15**).

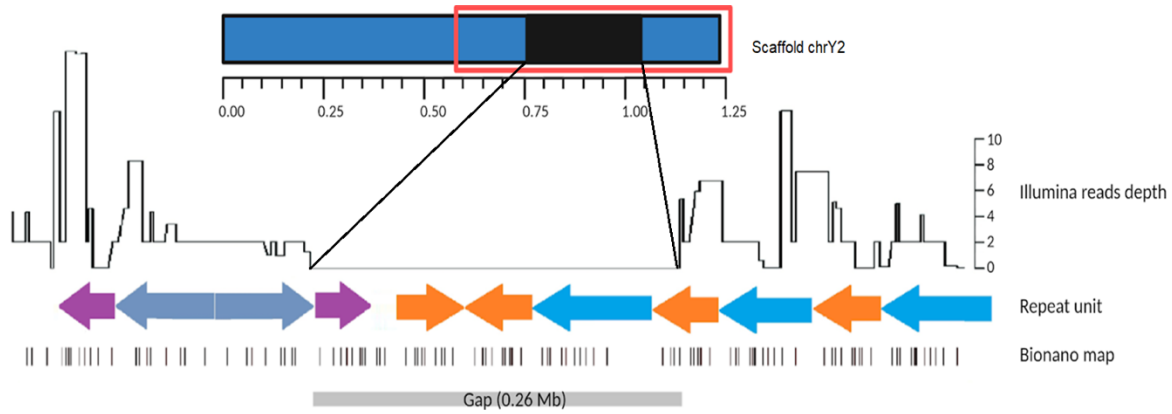


Figure 3.14. Complexity sequence in scaffold Y2. Within or flanking the gap, repeated and palindrome features are detected by the Bionano map. For the gap region, its repeat sequences were inferred by the occurrence of restriction sites whose patterns matched repeat units found within the neighbouring flanks.

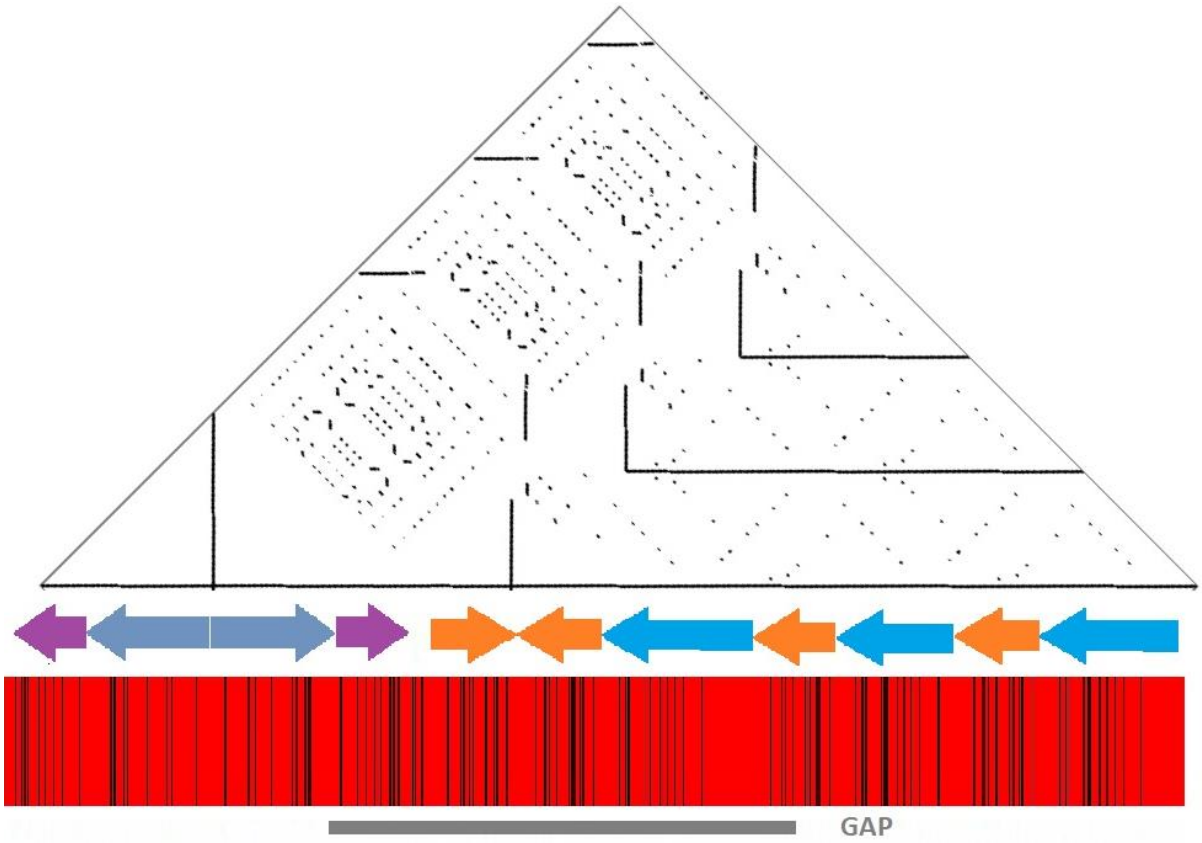


Figure 3.15. Self-similarity dot plot of the 0.26 Mb gap region in scaffold chrY2. Directly below the dot plot are distinct repeat units and orientation defined by coloured arrows. The red bar represents optical mapping data and black bars define restriction sites.

3.3.4.2 Y Chromosome-Specific LINE1 Array

There was a 28,800 bp region highlighted with 30x Illumina read coverage on the proximal flank of scaffold chrY2. This region consisted of a contiguous array of truncated LINE1_CF TEs. A singleton TE unit was 2,900 bp in length, truncating the 5' of ORF2 and missing ORF1 completely. In the RosY_1.0, nine LINE1_CFs were contiguously repeated and flanked on either side by a shorter LINE1_CF to form an array (**Figure 3.16A**). Estimated based on WGS data, read depth for all 29 female samples was negligible, proving the TE array is male-specific that belongs to Y chromosomes. And the depth for 186 male samples ranged between 15-99x with an average of 27x, indicating the length of this array was variable within dog populations (**Figure 3.16B, Supplementary Table 3.4**). Long-read data provided additional proof that the Y chromosome contains the TE array. Unlike the array with tandem LINE1_CF, normal LINE sequences were scattered across the dog genome. Assuming LINE1_CFs come from the Y chromosome, there should be a

difference in the number of LINE1_CFs for sequenced reads between males and females. Sequenced long-read data were scanned for counting the number of LINE1_CF sequences in each read. Indeed, the results demonstrated that males had notably more reads with the tandem pattern of LINE1_CFs than females. The longest read from a Labrador retriever had 66 LINE1_CF tandem repeats. Whereas female data did not show any reads with tandem LINE enriched (**Table 3.2**).

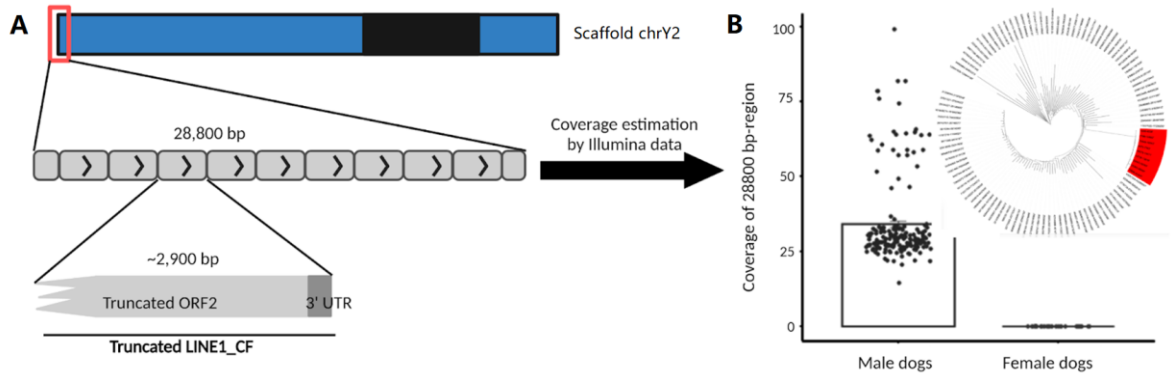


Figure 3.16. LINE1_CF array(s) in the dog Y chromosome genome. (A) Illustration of the LINE1_CF array on the RosY_1.0. One end of the chrY2 sequences has 9 repeated units, and each unit is composed of truncated LINE1_CF sequences. (B) Read depth of repeated units for males and females based on short reads data. There is no coverage for female dogs indicating the array was MSY-specific. Nine repeated sequences are clustered together (highlighted in red) in the phylogenetic tree, where 100 random LINE sequences across the whole genome are included.

Table 3.2. Long reads containing LINE1_CF tandem repeats are enriched in the male genome.

Breed	Sex	Platform	<=2	3~10	11+	Max	Accession
Labrador Retriever	Male	Sequel	1105	1828	160	18	PRJNA615959
Labrador Retriever	Male	Nanopore	23694	708	97	66	PRJNA610232
Basenji	Male	Sequel	1366	1481	108	18	SRR1130549
Basenji	Female	Nanopore	99631	158	0	5	SRR13077095
Boxer	Female	RSII	51	0	0	2	PRJNA13179
German shepherd	Female	Sequel	4	0	0	1	SRR13774717
German shepherd	Female	Nanopore	44371	81	0	4	SRR13774718

LINE1_CFs array was present incompletely in the RosY_1.0 as its normalised coverage was notably higher than that of the single copy region of the Y chromosome, and its one end was stopped as one of the boundaries of the scaffold chrY2 (**Figure 3.16A**). The assembled dog was further sequenced with ultra-long Nanopore technology to

investigate the LINE1_CFs array. As a result, no singleton read was able to span a full-length LINE1_CF array with two unique boundary sequences, and the longest read with tandem LINE1_CFs was 275 Kb in length including 94 LINE1_CF copies. To search for transition reads with LINE1_CF tandem repeats on one end and unique sequences on the other, the ultra-long reads and PacBio reads were screened. The bounds of the array were specified as four distinct sequences, each supported by 10, 8, 4, and 3 long reads, respectively. Therefore, the LINE1_CF repeats were present in two tandem arrays for the assembled dog, and arrays were estimated with a length of 0.86 Mb in total with about 300 copies based on the short reads coverage.

3.4 Discussion

RosY_1.0 is the most complete dog Y chromosome assembled to date but the repetitive feature of dog Y chromosome sequences hampered full-length assembly with gaps within and between scaffolds. The optical mapping demonstrated scaffold chrY2 was enriched with repetitive sequences, in the form of either tandem repeats or a palindrome. For the genome assembly using long reads from the PacBio Sequel I system, the repetitive regions were prone to be collapsed, leading to incomplete sequences. This is because, on one hand, the high error rate of long reads cannot distinguish between copies of highly repetitive sequences; on the other hand, no single long read can cover the full length of repetitive regions where a single copy was several kilobases in length. The same circumstances were seen for scaffold chrY3 with only one copy sequence being generated, which should be two copies existing in the dog Y chromosomes.

It has been previously suggested that the Y chromosome and low recombination regions accumulated interspersed repeats quickly to distinguish sex chromosomes (289). In agreement with this hypothesis, an enrichment of LINEs rather than SINEs or other TEs on the MSY of the dog Y chromosomes was seen compared with PAR, the recombination region. This is not a unique feature for mammals, e.g., the mouse Y chromosome was high in LINE sequences (4). In humans, the overall density of LINEs in MSY was higher than that of autosomes and the X chromosome (290). For non-

mammalian species, a high number of LINEs were observed on the W chromosome in tonguefish (291), and a high number of TEs on the Y chromosome of platyfish (292). In the African clawed frog, the abundance of TEs in the recombination region was significantly lower than that of the non-recombination region of the W chromosomes (293). Amassing TE sequences on the W/Y chromosome was strongly indicative of recombination deficiency (32). In turn, the enrichment of TEs may enlarge the non-recombination regions through insertions and inversions (294,295) and may lead to genetic decay and gene degeneration. This result indicated the insertion of LINEs played an important role in the formation of dog Y chromosomes.

SINE elements are distributed heterogeneously across the Y chromosomes. In dogs, the scaffold proximal to PAR displays a higher number of SINE insertions compared to the distal scaffolds. The dog's SINE density in the distal scaffolds is consistent with observations in chimpanzees, humans (66), and mice (4), where SINE presence is comparatively lower. An increased enrichment of SINEs is noticeable closer to PAR, implying that the inhibition of recombination between X and Y chromosomes might be linked to SINE insertions.

The existence of LINE1_CF tandem repeats on the dog Y chromosome was a unique feature to the other mammalian Y chromosomes. Although tandem repeats are very common within the Y chromosomes of studied mammals, the repetitive sequences were usually related to gene expansions. For example, amplicons contained gene families of *SLY*, *SSTY*, and *SRSY* in mice (223), of *HSFY*, *ZNF280BY*, and *TSPY* in bulls (64), of *YIR2*, *ETSTY*, and *UBA1Y* in horses (43), and of *TSPY* in humans (44). The truncated LINE1_CF was a non-coding sequence and its copy number was highly variable within the dog population indicating a novel evolutionary pattern of Y chromosome sequences.

The Y chromosome of the dog (27 Mb) is physically smaller than that of the human (57 Mb), mouse (95 Mb), and horse (45–50 Mb). By summing up the sequences assembled of 6.78 Mb, the unassembled LINE1_CF arrays of 0.86 Mb, the other copy of chrY3 sequences of 2.75 Mb, and the PAR of 6.6 Mb, a total of 16.99 Mb of Y chromosome regions were identified. The missing 10 Mb of sequence can be attributed to

ampliconic regions whose assembly is collapsed, as well as the missing centromere and NOR. The gaps between scaffolds could also contain ampliconic sequences whose lengths were not estimated in this study.

In this chapter, a hybrid assembly pipeline was developed to generate most parts of the haploid dog Y chromosome. These sequences are expected to provide fundamental knowledge for canine research and beyond in such areas as male fertility, sex determination, and development. The novelty of the dog Y chromosome sequences and genomic structure improve the understanding of mammalian sex chromosome evolution.

CHAPTER 4: Characterisation of Genes on the Dog Y Chromosome

4.1 Introduction

Typically, mammalian Y chromosomes have a limited number of distinct coding genes (**Figure 1.6**), however some genes can occur multiple times due to the accumulation of ampliconic sequences that contain gene families. Nonetheless, genes located on the Y chromosome can have critical functions. The importance of the Y chromosome was highlighted after the findings of the Y-linked genes being associated with gonadal sex reversal, Turner syndrome, spermatogenic failure, and graft rejection in the studies of individuals with incomplete Y chromosomes (296). The first finished human Y chromosome was annotated with 78 coding genes corresponding to 27 distinct gene families (44), followed by mouse Y chromosome with its 700 protein genes represented by 16 gene families (4). Representing carnivorans, the dog and cat MSY were annotated (55). According to Li *et al.* 2013, the dog Y chromosome was assembled with 2.5 Mb, containing 13 single-copy genes and 4 multiple-copy gene families. The families were represented as two copies of *BCORY*, seven copies of *SRY*, two copies of *CUL4BY*, and over 20 copies of *TSPY*. The cat MSY maintained 15 genes, 5 of which were multiple-copy genes, and there are 13 genes overlapped with the dog Y chromosomes'. The carnivore MSY displayed lineage-specific evolution as different gene family expansions and architectural changes were seen between the dog and cat (55). For example, *TETY2*, *HSFY*, and *CYorf15*, which were single-copy in the dog, exist as multiple-copy genes in cats. In cats, *FLJ36031Y* and *TETY1* translocated to the Y chromosome from autosomes, whereas these genes appear absent from the dog MSY (55). Though the differences between dog and cat Y chromosome might be attributed to incomplete assembly of their respective MSY, more recent phylogenetic analyses of Y chromosome genes identified reciprocally monophyletic clusters in Caniformia (dog-like carnivores) and Feliformia

(cat-like carnivores), suggesting that the Y-linked genes evolved independently after the split of these lineages (297).

Regarding dogs' multiple-copy genes, their copy number is consistent across several studies. Contrary to Li et al. s' estimate of seven *SRY* copies, which was based on quantitative polymerase chain reaction(qPCR) results from three dogs, others have estimated *SRY*'s copy number as three on average by both digital droplet PCR (ddPCR) and WGS read depth (55,280,298). For other carnivores, such as red foxes (228,298) and cats (55,299), single-copy *SRY* was reported on the Y chromosomes. Also, the estimated copy number of the *TSPY* gene in dogs was different between the read-depth method of over 100 copies (280) and the qPCR method of 25-35 copies (55).

With the emergence of the MSY genes, detection of strata is expected, however the gene composition and timing of strata across mammals should differ according to their evolutionary histories (**Section 1.1.2.3**). The cat was the only species for carnivores whose origin of Y chromosome genes was revisited by comparing X–Y homologous gene pairs (78). In agreement with humans, the earliest divergence times of stratum 1 genes support the origin of the *SRY*/*SOX3* and *CUL4BY*/*CUL4BX*, which occurred just prior to the eutherian–marsupial split. *AMELY* was the only gene that was assigned as stratum 4 indicated by a recent divergence time. Therefore, the knowledge of the origination of Y-linked genes can help us understand the evolution of sex chromosomes.

A major consequence of the stratification of Y genes is their inability to bidirectionally exchange variation through recombination. Having been cut off from their X-linked gametologs, the genes and their regulatory elements are less confined by the forces of purifying selection. Therefore, the evolution of Y genes can manifest through their gene expression. Previously, Y-linked genes were seen to be expressed in two ways: ubiquitous and testis-specific. Ubiquitous expressed Y-linked genes were interpreted to have retained their broad, dosage-sensitive functional roles. In contrast, the Y-linked testis-specific genes were interpreted to be functioning to support male fertility and sex development (31). Also, by comparing Y-linked genes with their X-linked gametologs, the *EIF1AY* had a fivefold increased expression than *EIF1AX* in the heart tissues and was

speculated as a potential gene that contributed to sex differences in heart diseases (87). To date, the expression of dog MSY genes, and more broadly carnivorans, was not studied due to limited data availability and gene annotation.

In this chapter, the evolution of dog MSY genes was explored using various metrics. MSY genes were annotated on the RosY_1.0 and integrated with publicly available datasets of RNA-Seq across dog tissues to characterise the expression pattern of MSY genes, which provided a foundation for understanding functions. The increased number of reference-quality genomes enabled divergence analysis that compared the dog to other species. Moreover, discovery of Y-linked polymorphisms provided new insights into the molecular basis of evolution, domestication, and adaptations on the dog Y chromosome.

4.2 Materials and Methods

4.2.1 Coding Gene Annotation

4.2.1.1 RNA-Seq Annotation

The RNA-Seq data for annotation were chosen from a wide range of tissues and canine breeds (**Supplementary Table 4.1**). Transcriptome reconstruction was accomplished with two computational strategies, alignment-based and *de novo* methods.

For the alignment-based strategy, RNA-Seq reads were mapped on the modified ROS_Cfam_1.0 directly with HISAT2 (239), and a transcriptome model was generated using StringTie (201).

In the *de novo* method, Trinity (202) used RNA-Seq data to construct transcripts. Then the assembled transcripts were aligned to the modified ROS_Cfam_1.0 using minimap2 (238), which included a splice model. The alignments with a mapping quality (mapQ) score greater than 55 were selected for the following annotation.

4.2.1.2 Protein-Based Annotation

Protein sequences were used to annotate coding genes that were not expressed in the analysed RNA-Seq and Iso-Seq data (**Section 2.3**). Coding gene sequences of Y chromosomes from humans, pigs, chimpanzees, and mice were downloaded from the UniProt FASTA database (https://www.uniprot.org/). Dog coding genes on the Y chromosome were collected from a previous study (55). Spliced alignment was conducted by the Spaln program (260) with "-M7" flag to allow detection of multi-copy genes.

4.2.1.3 Manual Gene Curation

Annotation tracks, including RNA-Seq annotation, Iso-Seq annotation, and protein-based annotation, were visualised simultaneously by IGV v2.3.90 (300). The gene model was merged manually with three steps. (i) Iso-Seq was the prioritising evidence of transcriptional expression and was recognised as the highest layer in the annotation. (ii) The second layer annotated genes that were expressed in the RNA-Seq, but absent in the Iso-Seq data. More specifically, only RNA-Seq annotated genes with intact open reading frames (ORFs) and homology with recognised genes of other mammals were used. (iii) Protein-based annotation, which served as layer three, was used to annotate genes that were not expressed in the RNA-Seq or Iso-Seq data. To avoid false annotation, the qualified gene model was required to have a comparable exon structure with its homology gene annotated in closed species. It should be highlighted that there is a chance that two neighbouring genes on the same strand will be mistakenly combined when RNA-Seq annotations are performed using the mapping-first or reference-free approaches. These errors were recognised when one gene model of RNA-Seq overlapped with two gene models from the protein-based annotation. To correct the error, the RNA-Seq-based model was split into two coding genes accordingly.

In the procedure above, isoforms for each gene were not taken into consideration since only Iso-Seq can construct isoforms with accuracy. Therefore, the transcript, with the longest ORF sequences, served as a representation for each gene.

The annotation of expressed pseudogenised and noncoding genes followed the same protocol as that of coding genes, and these genes' ORFs tended to be short and their putative peptides did not have any homology to known proteins. For the untranslated noncoding genes and pseudogenes, their annotation on the RosY_1.0 was lifted over from the RefSeq annotation of the RosCfam_1.0 using nucleotide BLAST (241).

4.2.2 Estimating Gene Copy Numbers

The Y chromosome genes were estimated for copy numbers *in silico* by counting the read depth within each gene locus. For genes that existed as multiple-copies on RosY_1.0, their copy numbers were counted by summing up the depth of the corresponding ampliconic loci. A total of 191 WGS samples with an average autosomal depth greater than 40x were selected. Their alignments on the modified RosCfam_1.0 were analysed using CNVnator with a bin size of 100 bp.

4.2.3 Gene Expression and Classification Analysis

94 RNA-Seq data from 23 tissues were downloaded including both adult and embryonic samples (**Supplementary Table 4.2**). Fastp v0.22 (301) was employed to process the raw RNA-Seq readings in order to eliminate low-quality reads and trim bad ends using the default parameters. Clean reads were mapped with STARv2.7.8 (302) and quantification of reads mapping to exons was done using featureCounts (255) in Subread (http://subread.sourceforge.net) tools. Gene expression was represented as transcripts per million (TPM), which normalizes read counts according to gene length (303). This step was completed by an in-house script.

A gene's expression level in a tissue was calculated from its median expression level among samples from that tissue. Log transformation of median expression was applied to hierarchical cluster Y chromosome genes using correlation distances by complete linkage method (87).

The tissue specificity index (τ) measured the tissue expression breadth (304). Assuming there were n tissues in the analysis, S_{max} and S_j presented the maximum expression level of the gene across all tissues. The τ was calculated as

$$\tau = \frac{\sum_{j=1}^{n}(1 - [\frac{log_2 S_j}{log_2 S_{max}}])}{n - 1}$$

4.2.4 Divergence Analysis

4.2.4.1 Annotation of Homologous Genes in Related Species

Homologous genes of selected species were used for calculating the rates of evolution for dog Y-specific coding genes. Two representative species of carnivorans were chosen to compare with dogs: the fox (a member of the Caninae family), and the cat (*Feliformia* suborder). Fox and dog were estimated to diverge around 12.2 million years ago (MYA), and for dog and cat, their split was around 55 MYA (305).

Orthologous genes of foxes and cats were annotated using a two-step method. Prior to the annotation stages, protein sequences of Y-specific genes were prepared, including the genes of the cat that have been already annotated (55,61,62) and the genes of the dog, which were presented in RosY_1.0. Cat's Y chromosome came from a male jungle cat genome assembly (FelChav1.0) and the Y chromosome sequences of the fox were assembled in the form of unplaced scaffolds (228). The first phase was to search for potential gene loci on the fox and cat genomes by TBLASTN, using the prepared protein sequences of Y-specific genes as a query. Next, the border of exons was defined by mapping the RNA-Seq data from the fox and cat genomes onto the corresponding fox and cat genomes. Notably, in the fox, when a gene was shown as spanning over two or more unplaced scaffolds, these scaffolds were linked together in the proper orientation and order according to the TBLASTN results.

Paralogous genes were located on the X chromosomes, and they were obtained from the RefSeq annotation of ROS_Cfam_1.0 for dogs, of ASM1834538v1 for foxes, and of felCat9.1_X for cats.

4.2.4.2 Calculating the Rates of Evolution

Coding sequences (CDS) were detected by ORFfinder (https://www.ncbi.nlm.nih.gov/orffinder/). The paralogous and orthologous genes' CDS

were aligned with ClustalW in MEGA7 (243). The evolutionary rate Ka, Ks, and Ka/Ks were calculated based on the γ-MYN method (306) in KaKs_Calculator 2.0 (261).

4.2.4.3 Construction of Phylogenetic Tree

To construct phylogenies of mammalian MSY genes, homologous sequences were downloaded from the NCBI database (https://www.ncbi.nlm.nih.gov/) for eutherian mammals, opossum, and platypus. For each MSY homolog, the coding DNA sequences were predicted by ORFfinder and aligned with ClustalW for codon sequences (243). Phylogenies were built with the Maximum Likelihood algorithm using a Tamura-Nei model and the reliability was tested by 100 bootstrap replicates (243).

4.2.4.4 Calculation of Intron Similarity Level

The similarity of intron sequences between X-linked and Y-linked genes was investigated based on the k-mer method. First, intron regions and TE sequences were annotated on the RosY_1.0. Then, the annotated TE intervals within introns were removed by the subtract function in bedtools (248), generating non-TE intronic sequences. Third, the non-TE intronic sequences were randomly extracted as 150-mers. Fourth, the intronic sequences of X-linked genes were extracted and regarded as subjects, and a BLAST search was done on the X-linked gene introns with the 150-mer as a query. BLAST searches were conducted with relaxed parameters (-evalue 1e-8 -word_size 7 -gapopen 5 -gapextend 2 -penalty -1 -reward 1) to allow divergent sequences to align. For each gene, the similarity of the intron was calculated as the length of the alignment divided by the total length of 150-mer.

4.2.5 Polymorphism Comparisons

4.2.5.1 Study Cohort

222 male samples, including 212 dogs and 10 wolves, were utilised for polymorphism analysis on Y chromosomes in the *Canis lupus* population (**Supplementary Table 4.3**). In order to maximise the detection of variations, sampled dogs included 198

recognised breeds dogs and 14 indigenous dogs and 10 wolves distributed widely over the world. The sequencing depth of selected samples was greater than 5X and the gender for each sample was inferred by sequencing depth ratio between chrY1 and autosomes (details to see the repo).

4.2.5.2 Variant Calling and Annotation

Illumina sequencing raw reads were pre-processed with fastp v0.22 (301) to prune out poor reads (average quality score < 20) and cut ends with low quality. Prepared reads were aligned to the modified RosCfam_1.0 assembly using BWA2 v2.2 (237). Single-nucleotide variants (SNVs) and small insertions/deletions (INDELs) were called within the pseudoautosomal region (PAR) interval of the X chromosome (NC_051843.1: 1-6,590,648) by the Genomic Analysis Toolkit (GATK) v4.1.7 (249,307). First, each BAM file was processed with Mark Duplicates and BQSR (Base Quality Score Recalibration) steps to generate analysis-ready BAM files. Second, the HaplotypeCaller program called variants to create individual GVCF files. Third, joint calling was performed simultaneously on all GVCF files to create a cohort VCF file of all samples. Finally, high quality raw variants were retained for use in subsequent analyses (SNV criteria: QD > 2.0, QUAL > 30.0, SOR < 3.0, FS < 60.0, MQ > 40.0, MQRankSum > -12.5 and ReadPosRankSum > -8.0; indel criteria: QD > 2.0, QUAL > 30.0, FS < 200.0, ReadPosRankSum > 20.0). The prediction of variants' functional effects was performed with SnpEff v3.0 (251).

A brief explanation of GATK criteria terms are as follows. QD (Quality by Depth) measures variant quality relative to the depth of sequencing coverage. QUAL (Quality) is a quality score indicating high confidence in a variant call. SOR (Strand Odds Ratio) checks if a variant is evenly supported by both DNA strands in balance. FS (Fisher Strand Bias) strand bias in variant calls suggesting minimal bias in strand distribution. MQ (Mapping Quality) represents high mapping quality of reads covering the variant position, reflecting confident mapping. MQRankSum checks if the mapping quality ranks of supporting reads are reasonable. ReadPosRankSum evaluates the position of the variant

within reads, and a value greater than the threshold suggests that the variant is not consistently found at the ends of reads, reducing the likelihood of sequencing artifacts.

4.2.5.3 Construction of Time-measured Phylogeny

BEAST analysis was performed by Jiaqi Yang (Max Planck Institutes) and Laurent Frantz (University of Oxford and Ludwig Maximilian University of Munich). Ancient sample data were obtained either by downloading from NCBI SRA (accession numbers: SAMEA104190273, SAMN04884534, SAMEA7538371, SAMN04884535, PRJEB13070) or were provided by the collaborator from the University of Oxford. Eleven directly radiocarbon-dated samples with genotype missingness <10% were included for tip-dating (**Supplementary Table 4.4**). All the reads were mapped to the genome using Bowtie2 with the "--very-sensitive-local" parameter. Strelka2 was used to call SNVs on the Y chromosome using its default settings but with the addition of the "--vcf" option (250) This option restricted Strelka's SNV calls to just biallelic polymorphisms made available by a reference panel of modern canids (**Section 4.2.4.5**). BEAST v2.51 was used to calibrate the substitution rate on the Y chromosome (264). Due to the generally low sequencing coverage of the Y chromosome in ancient samples, there was a concern that deamination could result in inaccurate genotypes. To address this, the analysis using the GTR + G model involved setting A-G and C-T substitution rates to zero. The MCMC step was then set to iterate 100 million times, with the initial 10% of runs being omitted from the analysis.

4.3 Results

4.3.1 Gene Content of RosY_1.0

The RosY_1.0 assembly was annotated, resulting in 23 coding genes and 43 pseudogenes or non-coding genes (**Figure 4.1, Supplementary Table 4.5**). Among the coding genes, four genes, *WWC3Y*, *AP1S2Y*, *TMSB4Y*, and *PRSSY*, were not previously reported for the dog Y chromosome. *TSPY*, *CUL4BY*, *BCORY*, and *UBE1Y* are ampliconic, representing cumulatively 6, 3, 2, and 2 copies, respectively, in RosY_1.0.

The copy number for coding genes was estimated *in silico* based on mapping depth of whole-genome sequencing data (**Figure 4.2**). Of the remaining 19 genes that were annotated as single-copy, 18 were validated as such by the *in silico* prediction. Although annotated once, *SRY* was estimated to have two or more copies in the dogs tested. More than 82% of dogs' Y chromosomes contained two copies of *SRY*, and the maximum was 4 copies, which was only detected in one New Guinea singing dog sample. Notably, the *SRY* locus occurs in a complex DNA structure, with 2.8 Kb of spacer sequence containing a coding region, flanked by 60 Kb of palindromic sequences of 99.7% similarity (**Figure 4.3**). *UBE1Y* and *BCORY* had two copies, the same as the RosY_1.0 presentation. The copy numbers of *CUL4BY* were estimated from 10 to 18 with a medium number of 12, and TSPYs' copies ranged between 51 and 90 with 67 in the medium (**Figure 4.2**).

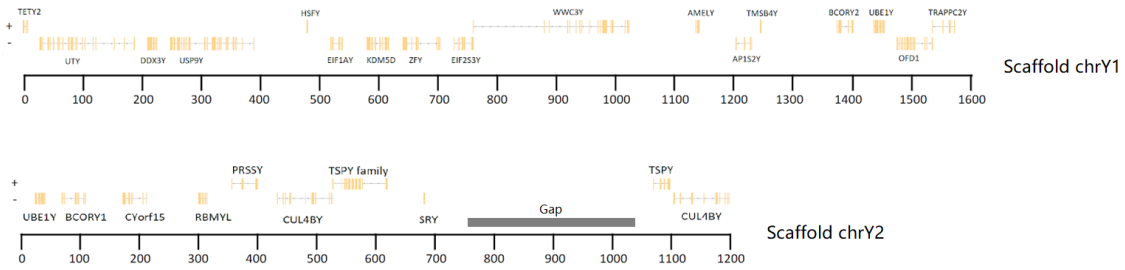


Figure 4.1. Distribution of annotated coding genes on the RosY_1.0. Scaffolds chrY3 is ignored due to the absence of genes.

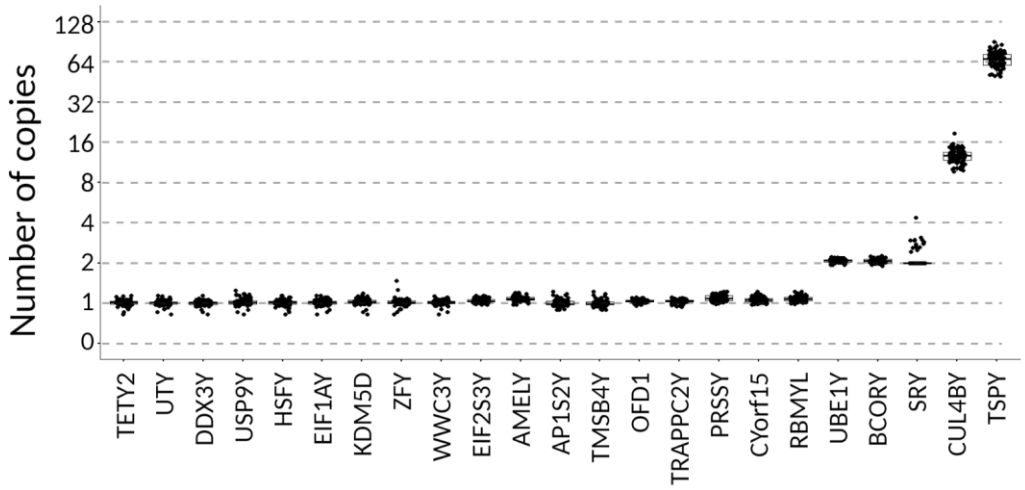


Figure 4.2. Estimated copy number of MSY genes based on WGS data read depth.

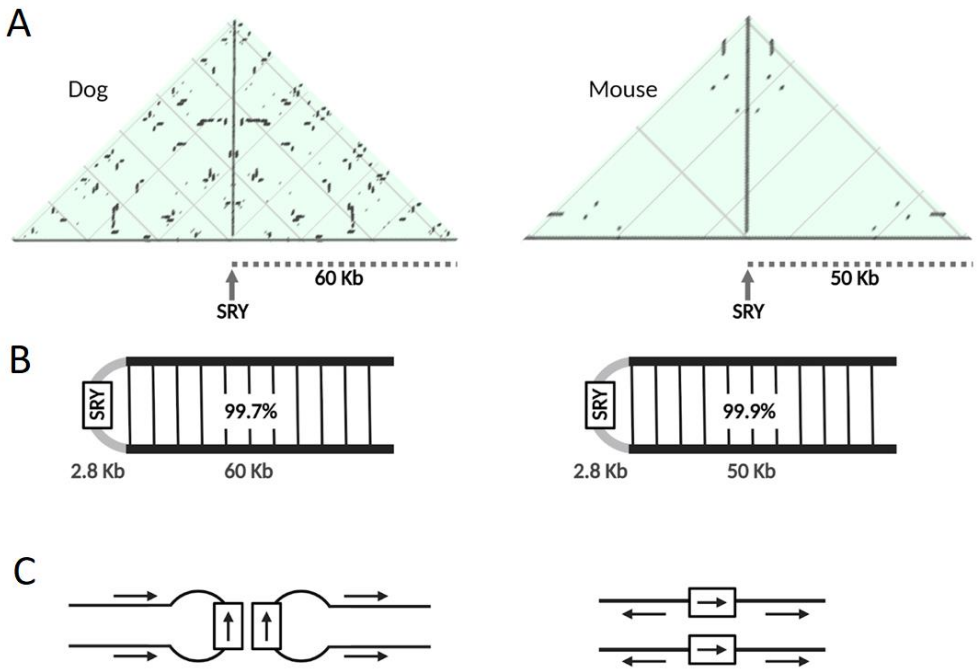


Figure 4.3. Dogs' *SRY* gene is embedded within a palindrome structure. (A) Self-similarity dot plot for *SRY* and its flanking regions in the dog and mouse. (B) The palindrome structure has formed a 2.8 Kb length unique sequence in both the dog's and mouse's *SRY* genes. (C) The multiple-copy *SRY* and its palindromic structures prevent decay by enabling gene conversion in two possible ways. One way is that two palindromes' arms recombine independently, and the gene conversion occurs between two palindromes' loops (left). And the other is that intrachromosomal recombination between two palindromes enables gene conversion within the spacers (right). Boxes represent *SRY* genes and arrows indicate palindromes' directions.

4.3.2 Origin of MSY Genes

The divergence of XY gametolog sequences was examined to establish evolutionary strata of dog's Y-linked genes. Synonymous substitutions (Ks) measure divergence time between XY gametologs assuming they evolved in neutral fashion. Consistent with studies of horses and humans, *HSFY*, *SRY*, *CUL4BY*, and *RBMLY* had distinct Ks values with a mean of 3.14. These genes represent therian's Stratum 1; as the oldest genes, they had a clear boundary with other genes in terms of their physical position (**Figure 4.4A**). The remaining 18 genes, which were assigned to the Strata 2/3 in human (1) and horse (43) studies, had a mean Ks value of 0.62, with an outlier of 3.90 for *TSPY*. TSPY's departure from other S2/3 genes was also observed in cattle, rats, marmosets, rhesus monkeys, chimpanzees, and humans (1). An explanation for these

observations is that *TSPY* diverged earlier than other Stratum 2/3 genes or that TSPX is actually a paralog (e.g. the X-linked ortholog of *TSPY* was lost from therian X chromosome).

For each of the dog's MSY genes, a phylogenetic relationship was constructed based on the homologous coding sequences from a range of mammals coding sequence. The phylogeny of X-linked and Y-linked genes could be classified into monophyletic (**Figure 4.4B**) and polyphyletic categories (**Figure 4.4C**). Of the dog's 23 MSY genes, 16 genes displayed a monophyletic pattern with other mammalian Y homologs, supporting a single origin of mammalian MSY genes (**Supplementary Figure 4.1**). Seven MSY genes were polyphyletically clustered with the Y homologs of other species, suggesting parallel evolution of these gametologs or multiple independent origins in mammals (**Supplementary Figure 4.2**). Among these, *ZFY* and *EIF2S3Y* generated a single origin for *Scrotifera* species to form a sister group with their X-linked gametologs. *TRAPPC2Y*, *TMSB4Y,* and *OFD1* originated at the common ancestor of canids, such that XY gametologs separated after the formation of *Canidae*. Besides, the topology showed that *WWC3X/Y* and *AMELX/Y* derived from the common ancestor of *Caniformia* and *Carnivora* respectively.

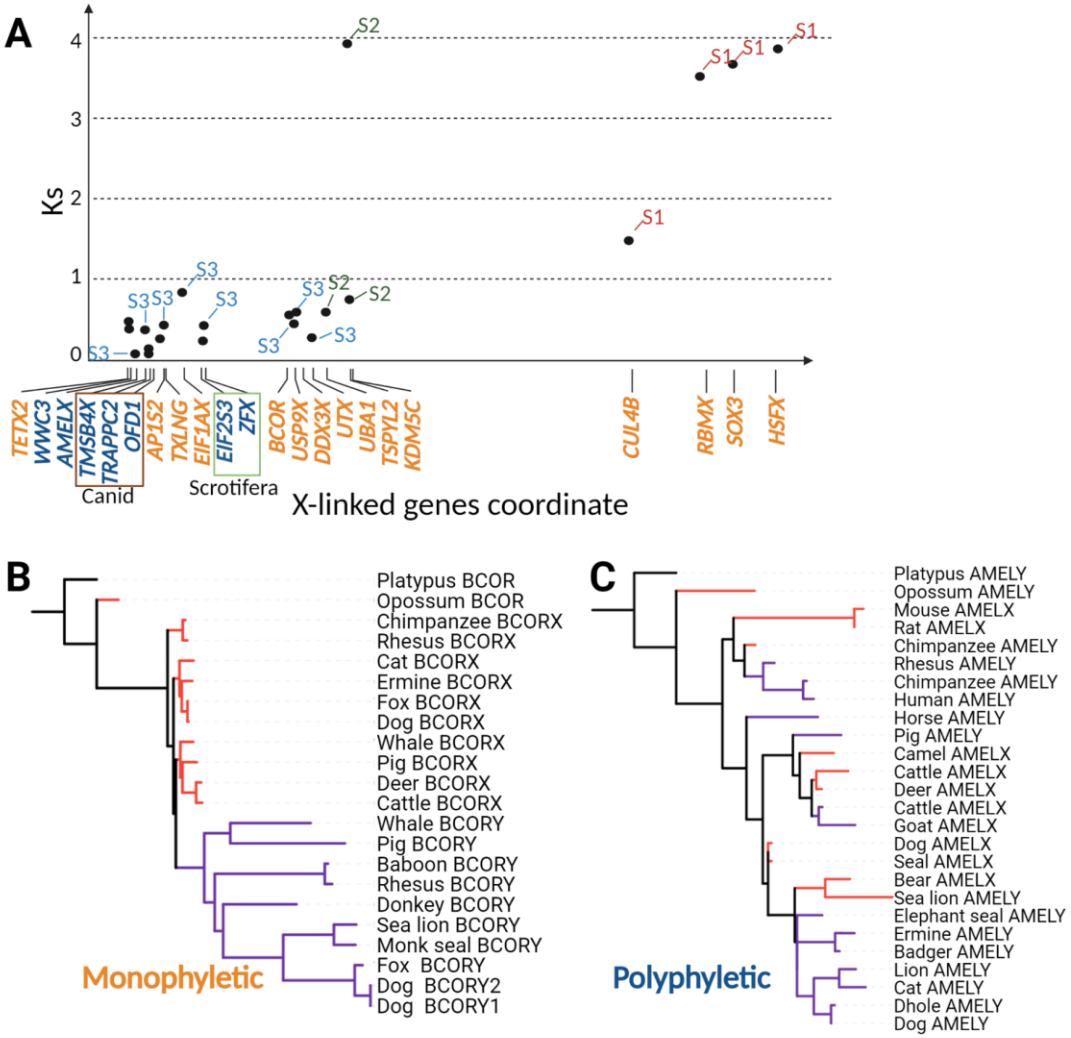


Figure 4.4. Evolutionary strata of MSY genes. (A) Ks values between X and Y gametologs are ordered by the physical coordinates on the dog X chromosome. Dots are labelled with evolutionary strata (S1, S2, and S3) according to previous studies in humans. Gene names are coloured based on the topology classification, and brown and green boxed genes indicated these are derived from Canid and Scrotifera ancestors, respectively (referring to **Figure 4.5** and **Supplementary Figures 4.1 and 4.2**). (B) An example of the monophyletic MSY gene *BCORY*. (C) An example of the polyphyletic MSY gene *AMELY*.

The polyphyletic clades of MSY genes might be caused by the process of gene conversion between XY gametologs. The topology can infer the time when gene conversion occurred. Theoretically, the earlier that gene conversion occurred, the fewer homologous sequences are retained between XY gametologs' respective introns. To validate this, the intron similarity for each pair of homology genes was calculated (**Figure 4.5**). Consistently, all monophyletic genes, which originated ancestrally without

conversion events between sex chromosomes, displayed low intron similarity compared to polyphyletic genes. One exception was *TETY2*, a gene that is located within the pseudoautosomal boundary (PAB) and whose sequences are only partially homologous to the X-linked PAB due to an unusual rearrangement in this region (**Section 5.3.1**). It is known that canid-originated genes had the most recent conversion event and two of three canids-originated genes, *TMSB4Y* and *OFD1*, were observed to have the highest levels of intron similarity. *Scrotifera*-originated genes (*ZFY* and *EIF2S3Y*) were the earliest converted genes in this study showing a slightly higher similarity level than monophyletic genes, but lower than other polyphyletic genes. However, the *Carnivora*-originated gene, *AMELY*, was higher than the *Caniformia*-originated gene, *WWC3Y*, in intron similarity and even at the same level with another canid-originated gene, *TRAPPC2Y*. Besides, the gene conversion events were supposed to occur as a block of genes. As seen, those genes with the same origination were close to each other on the dog Y chromosomes (**Figure 4.4**).

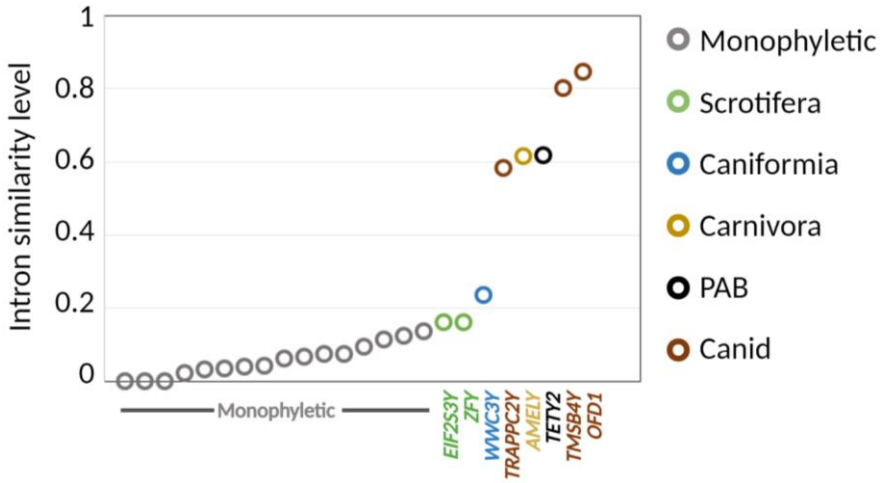


Figure 4.5. Intron similarity and origin of genes. Intron similarity for MSY genes is plotted in order. Each dot represents each gene coloured by the origination. Monophyletic, the gene originated ancestrally without gene conversion; Scrotifera, the gene with a gene conversion at the ancestor of *Scrotifera*; Caniformia, the gene with a gene conversion at the ancestor of *Caniformia*; Carnivora, the gene with a gene conversion at the ancestor of *Carnivora*; PAB, the gene located within the pseudoautosomal boundary; Canid, the gene with a gene conversion at the ancestor of *Canid*.

4.3.3 Classification of MSY Genes

The divergence of XY gametologs can manifest through their gene activity and functions. To test for this, gene expression data were examined. A total of 95 bulk-tissue RNA-seq samples from male dogs were collected from public archives, spanning 22 tissues and two developmental stages (adult and embryo). We quantified the expression abundance of MSY genes based on the alignment of short-reads. For multiple-copy genes, their expression was estimated by summing up the expression of all its corresponding single copies. All 23 MSY genes or gene families were observed as broad, testis-specific, or with no expression in this study. Intuitively, broadly expressed genes can be subgrouped into high and low groups with at least half of analysed tissues having a TPM over 10 for the high group and a TPM lower than 10 for the low group (**Figure 4.6A**). *SRY* and *AMELY* were excluded from classification, as there were no embryonic data sets from male dogs to support their expression. Consistent with this observation, hierarchical analysis clustered the expressed genes into three groups that were named ubiquitous, low-expression, and testis-specific for the following studies (**Figure 4.6A**, **Supplementary Table 4.6**). Tissue specificity index (τ) and expression abundance (TPM) supported these groups. τ of testis-specific genes was significantly higher than that of ubiquitous or low expression genes (P=0.009 and P=0.000, respectively), whereas ubiquitous genes did not appear to be expressed more broadly than low expression genes statistically (P=0.728) (**Figure 4.6B**). In terms of expression abundance, on average the ubiquitous genes are expressed higher than low-expression genes and higher than testis-specific genes across all the tissues that were tested (**Figure 4.6B**).

The relationship between the XY gametologs was investigated in terms of their relative expression (**Figures 4.6C, 4.7**). The gene expression of all the ubiquitous and low-expression Y-linked gametologs were strongly correlated with their X-linked homologs, especially *OFD1*, *EIF2S3Y*, *KDM5D*, *USP9Y*, *UTY*, and *ZFY* (r^2>0.8). Testis-specific genes and their paired X-linked genes showed a substantially lower expression correlation compared with ubiquitous and low expression genes, and *TSPY*, *BCORY*, and *HSFY* were

not coexpressed with their paired genes significantly (P>0.05). *PRSSLY* and *TETY2* had no X-linked gametologs, therefore they were omitted from the analysis.

Quantitative differences in XY gametologs were characterised by comparing the abundance of gene expression tissue by tissue (**Figures 4.6D, 4.8**). Four ubiquitous genes (*EIF2S3Y*, *EIF1AY*, *USP9Y*, and *UTY*) were expressed equally with their X-linked gametologs in most tissues. Three ubiquitous genes (*TMSB4Y*, *DDX3Y*, and *KDM5D*) and all five low expression genes showed higher relative expression in their X-linked gametologs in most tissues, with an X/Y ratio of over 2:1. In another gametolog pair, *ZFX/ZFY*, the Y-linked gene was expressed more than twice than its X-linked gametolog. The testis-specific genes had virtually no expression in other tissues, so the X/Y ratio was extremely high when their X-linked gametologs were expressed. Even though in testis, *TSPY*, *RBMYL*, and *UBE1Y* were still expressed at lower levels than their X-linked gametologs.

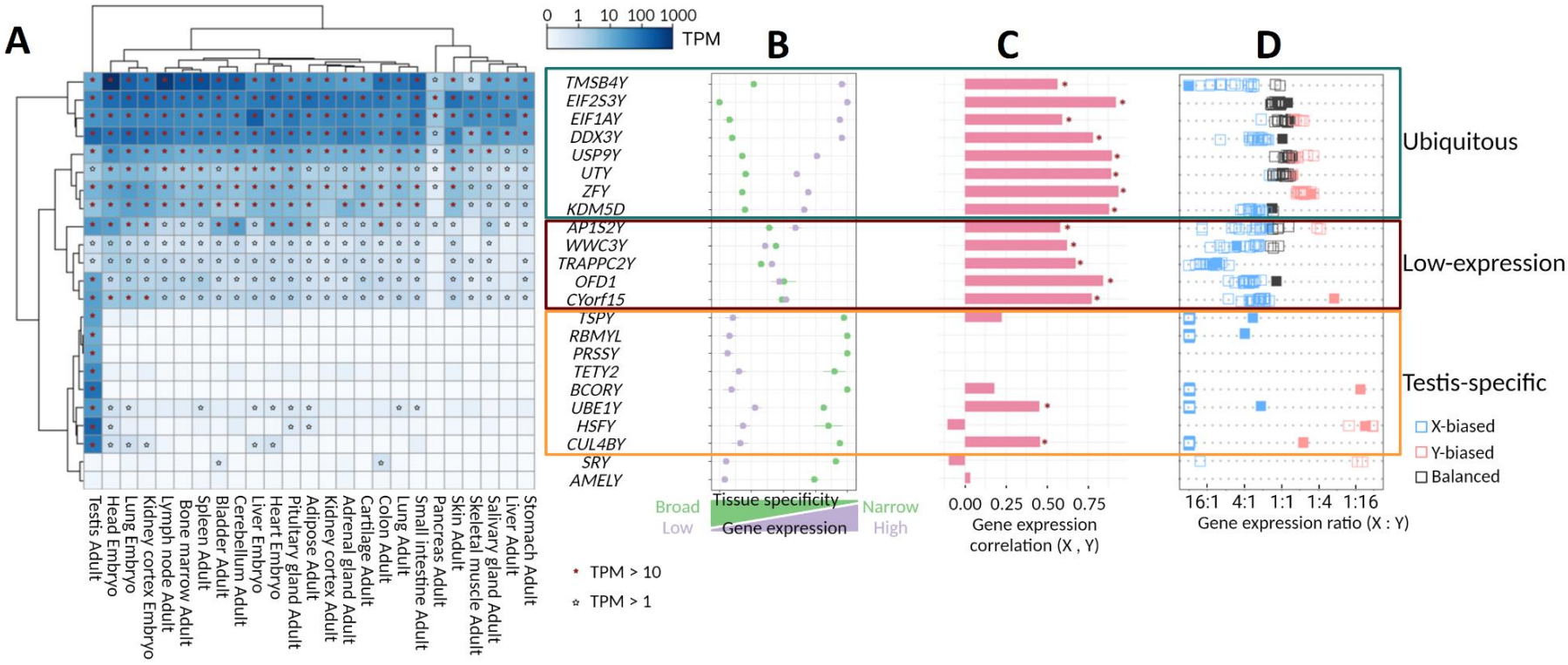


Figure 4.6. The expression, classification, and characterization of MSY genes. (A) The expression of MSY genes and gene families at the median level of tested samples across tissues. The row and column order is determined by hierarchical clustering. The solid star symbol means a high expression with a TPM over 10, and the open star symbol donates a low expression between 1 to 10. Asterisks on the right of the gene symbol indicate multiple copy genes. (B) Tissue specificity and average expression level for each gene. (C) The correlation level of expression between XY gametologs. Asterisks signify significant correlations. (D) The ratio of expression abundance between XY gametologs coloured by the expression balance. Each square represents one tissue and solid squares donate testis. (A high resolution is available on https://github.com/WengangXbio/script_bio/blob/main/Figure%204.6.png)

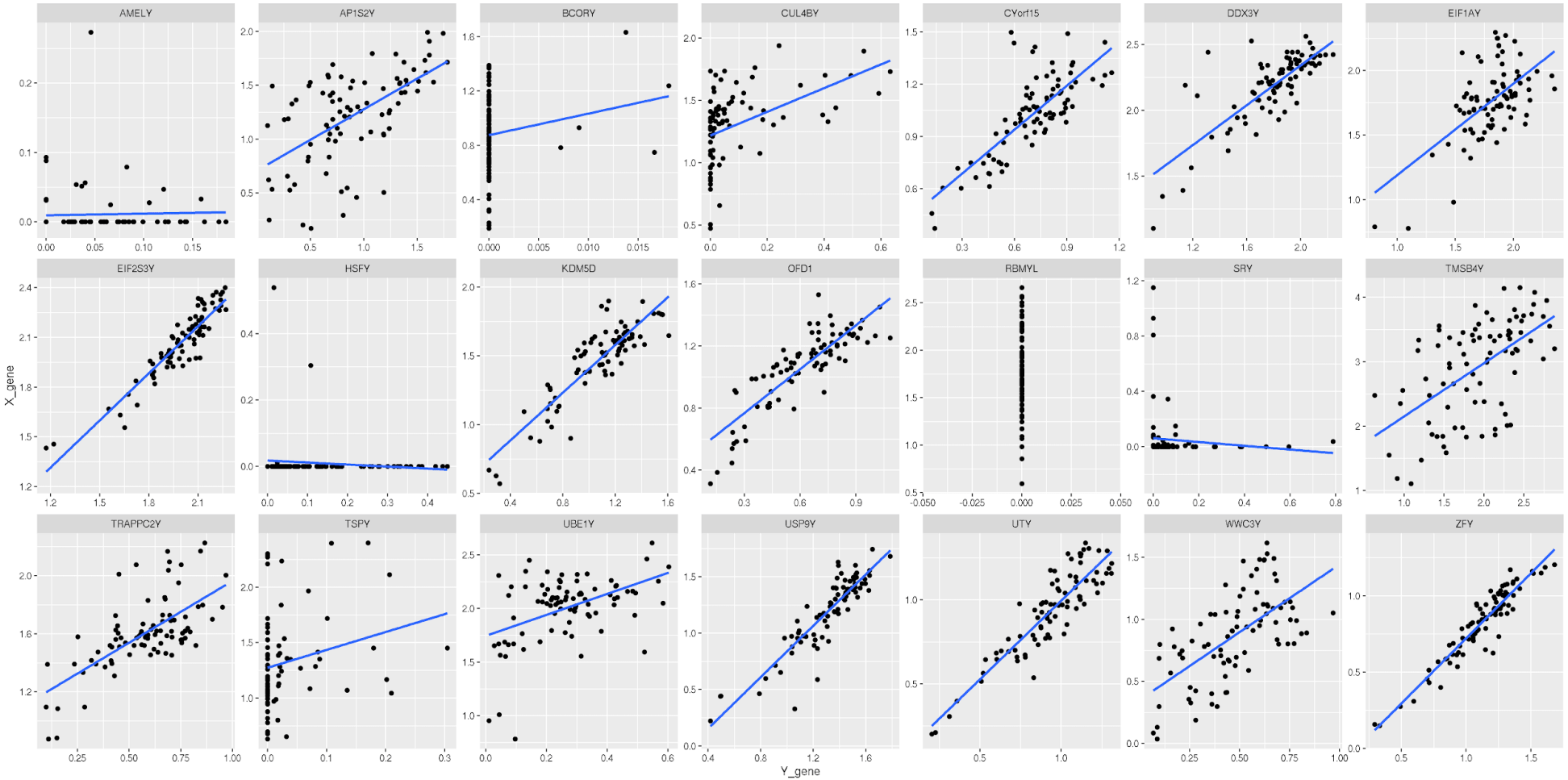


Figure 4.7. The coexpression of MSY genes with their X-linked homologs. The correlation between Y-linked (x-axis) genes and X-linked (y-axis) genes in expression is shown in the scatter plot. Each dot presents one sample from a wide range of tissues after excluding testis samples.

Figure 4.8. Expression comparison between XY gametologs. The comparison in expression between X-linked (pink bars) and Y-linked (green bars) gametologs. Each pair of bars shows the median expression level of one tissue with an adjustment of the expression of the X-linked gene to 1. A high resolution is available on https://github.com/WengangXbio/script_bio/blob/main/Figure%204.8.png

4.3.4 Divergence Analysis Revealing Gene Evolution

The long-term selection pressure on MSY genes was measured using the ratio of nonsynonymous to synonymous substitution rates (Ka/Ks) among paralogous and orthologous gametologs (**Figure 4.9A**). When the Ka/Ks ratio approximates "1", selection pressures are interpreted as neutral. A ratio <1 is indicative purifying selection, whereas a

ratio >1 is associated with positive selection (308). This analysis included paralog comparisons between X-linked and Y-linked homologs for dogs, foxes, and cats individually, and ortholog comparisons for Y-linked genes in dog-cat, dog-fox, and fox-cat. Before the selection analysis, in terms of expression, cats' and foxes' MSY genes were assumed to be grouped in the same way as dogs. This is well-justified, as even humans – whose last common ancestor with dogs dates back approximately 100 MYA, share the same expression profiles (i.e. *ZFY*, *USP9Y*, *UTY*, and *EIF1AY* were ubiquitous, and *TSPY* and *RBMY* were testis-specific) (2). Divergence analysis results were presented according to the gene classification (e.g. ubiquitous, low-expression, and testis-specific)

In both the paralogous and orthologous comparisons, ubiquitously expressed Y-linked genes had the lowest Ka/Ks ratio on average in all three species analysed (**Figures 4.9B,C**). In the paralogous gene comparison, the Ka/Ks ratios of testis-specific and low expression genes in fox and dog was significantly higher than the ratios of species' respective ubiquitously expressed genes (**Figures 4.9B, Supplementary Table 4.7**). In the orthology gene comparison, testis-specific and low-expression genes' Ka/Ks ratios were statistically higher compared to those of ubiquitously expressed genes (**Figure 4.9C, Supplementary Table 4.8**). Of note, in cats only one low-expression gene (CYorf15) was annotated, therefore low-expression group comparisons were statistically intractable; nonetheless the Ka/Ks ratio of CYorf15 was higher than both testis-specific and ubiquitous groups of genes. The aforementioned observations suggest that ubiquitously expressed genes for all species experience higher levels of purifying selection than testis- and low-expressed Y-linked genes. Among orthologs, Y-linked testis- and low-expressed genes experience stronger levels of purifying expression that manifests in a lineage-specific manner.

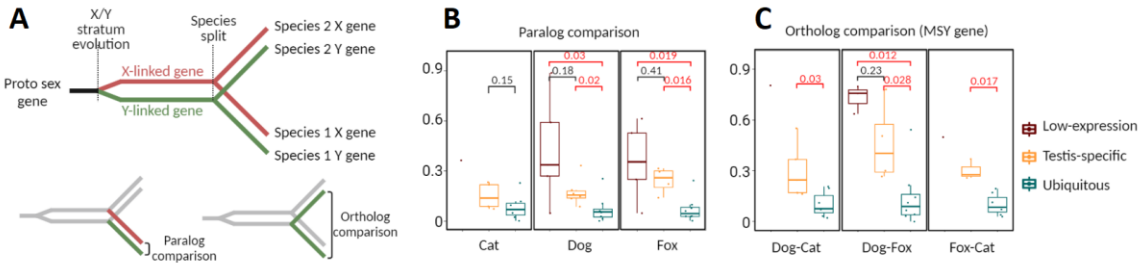


Figure 4.9. Divergence analysis (Ka/Ks) of paralog and ortholog comparison for MSY genes in carnivorans. (A) Illustration of paralogous and orthologous gene divergence. (B) Paralog comparison, Ka/Ks ratio between X-linked and Y-linked homologs for dogs, foxes, and cats. (C) Ortholog comparison, sequence divergence between orthologous genes (dog-cat, dog-fox, and fox-cat) for MSY genes. A t-test was performed in the ggpubr package in R.

4.3.5 MSY Gene Selection within Dog Population

Polymorphism data within the dog population was investigated, which reflects gene selection over tens of thousands of years. Technically, gene conversion and translocation events allow short reads from MSY genes to be multimapped with reads from other chromosomes, which leads to heterozygosity variants being called on haploid Y sites. Also, ampliconic MSY genes are prone to be called as heterozygotes when some of their copies had *de novo* mutations. To test the reliability of genotyping the MSY genes across dog populations, 222 sequenced dogs were genotyped for coding regions on the Y chromosomes. This effort revealed that *USP9Y*, *CUL4BY*, and *TSPY* genes were enriched with large amounts of heterozygous calls. Therefore, they were removed from polymorphism analysis (**Supplementary Table 4.9**).

Next, genic Y chromosome variants for the remaining Y-linked genes were annotated. Overall, a total of 60 missense variants and 46 synonymous variants were detected in 17 and 15 MSY genes, respectively (**Figure 4.10A-B, Supplementary Table 4.10**). Notably, five genes (*BCORY1*, *KDM5D*, *OFD1*, *UBE1Y*, and *WWC3Y*) accumulated more missense variants (68.3%) than others (42.5%), and all popular missense sites (MAF>0.1) came from these five genes. When the selection was neutral, the proportion of missense and synonymous variants were estimated by simulating variants in the CDS region randomly and bootstrapping indicated 77.2% of missense variants theoretically (**Figure 4.10C**). An overall proportion of 56.6% indicates a negative selection eliminating

missense variants historically. By tracing the genotypes of missense mutations on the Y haplotype phylogeny tree, 8 variants formed a monophyletic clade indicating their single origins (**Figure 4.11**). All the mutations demonstrated a MAF exceeding 0.1 in both the dogs and wolves examined in this thesis and the BEAST result showed they existed and spread in the dogs' Y chromosome 6.2 K - 54.6 K years ago. Seven of them were biallelic and only one was triallelic. Notably, these ancestral variants came from five genes (*BCORY1*, *KDM5D*, *OFD1*, *UBE1Y*, and *WWC3Y*), four of which were observed as an accumulation of more missense variants than others (**Figure 4.12**).

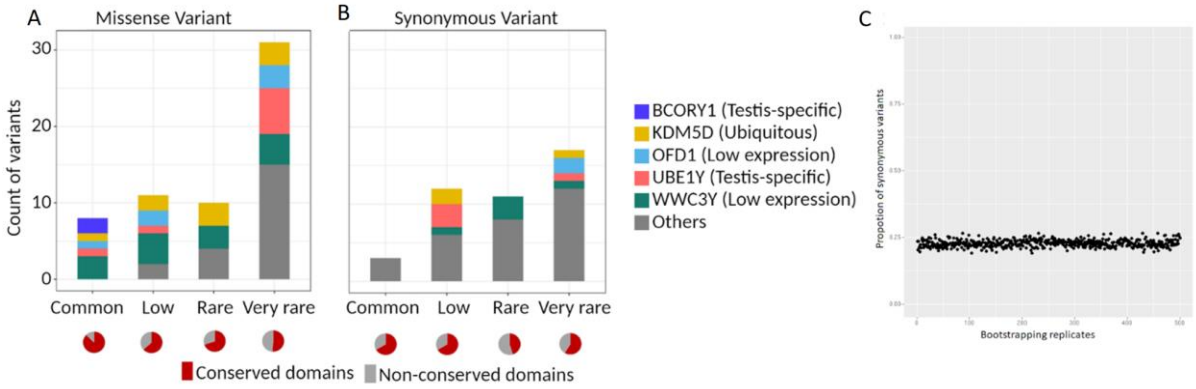


Figure 4.10. MSY gene missense and synonymous variants within *Canis lupus*. Missense variant (A) and synonymous variants (B) are grouped based on the minor allele frequency coloured highlighted genes and all others (grey). The pie chart below each bar indicates the proportion of variants that were located within functional domains. (C) The theoretical proportion of missense and synonymous variants was estimated by bootstrapping replicates.

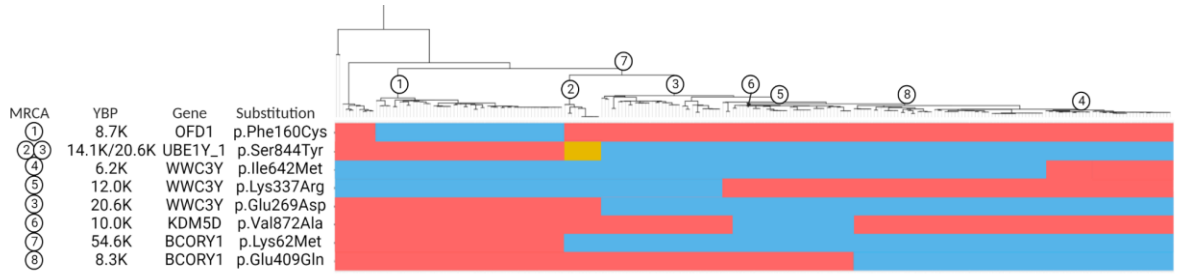


Figure 4.11. The most recent common ancestor of ancestral missense variants. The genotypes of selected individuals are coloured to distinguish their origin. MRCA is shown on the phylogenetic tree, which is built based on the Y chromosome markers. MRCA, most recent common ancestor; YBP, years before present. (A high resolution is available on https://github.com/WengangXbio/script_bio/blob/main/Figure%204.11.png)

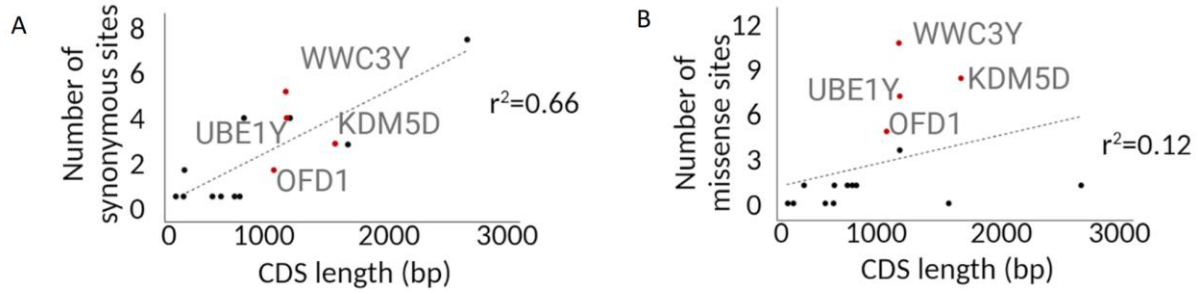


Figure 4.12. Enrichment of missense variants on the *UBE1Y*, *OFD1*, *KDM5D*, and *WWC3Y*. Scatter plots show relationships between CDs length and synonymous sites (A) and missense sites (B). Each dot represents one MSY gene in dogs.

4.4 Discussion

The number of gene copies was estimated based on the coverage of whole genome sequencing. Although the estimation of *TSPY*'s copy number was in disagreement with previous reports (55,280), *TSPY* presented the most copies on the dog Y chromosome, a phenomena that also occurs in mice (4), bulls (64), and horses (43). With representation from *Carnivora*, it appears that *TSPY* contributed to a massive expansion of ampliconic sequence on the Y chromosomes of multiple mammalian lineages. *SRY* showed variable copies based on *Canis lupus* samples ranging from 2 to 4. In this study, most male dogs were predicted to carry two copies of *SRY*, whereas ddPCR (298) and the coverage-based methods indicated an average of three copies of the *SRY* in the dog population (280). Until now, the *SRY* genes' physical positioning within palindrome sequences were only reported in mice (100). In mice, the palindrome exhibited a length of 50 Kb on one side, while in dogs, it was 60 Kb. Both species shared a distinctive feature of a 2.8 Kb unique sequence separation at the centre. In dogs, it is hypothesised that the multiple-copy of *SRY* and palindromes embedded allowed it to prevent gene decay by intrachromosomal conversion (**Figure 4.3C**).

As widely acknowledged, mammalian sex chromosomes evolved from a pair of autosomes. While 98% of ancestral genes were retained on the X chromosomes (75), only 3% survived on the Y chromosomes (44,309). The remarkable longevity of Y-linked genes can be explained by either the demand to preserve the dosage sensitivity of regulatory genes shared between both the X and Y chromosomes, or alternatively, the acquisition of

male-specific function(s) that confer some sort of advantage (1,310). In my transcriptomic analysis, the "ubiquitous" classified genes were co-expressed in tissues and in balanced quantities with their X-linked gametologs. This observation suggests that Y-linked ubiquitous gametologs are expressed with dosage equivalence between XY males and XX females. Their widespread expression suggests that ubiquitous genes might retain a wide range of functions and the cis-acting regulatory elements that direct their ubiquitous gene expression were conserved through Y chromosome evolution. UniProt annotations indicated the Y-linked genes were retained with housekeeping functions (1), and a few studies confirmed that some ubiquitous genes played a role in broad regulations of transcription and translation. For example, the DDX3Y protein regulates the beginning of cyclin E1's translation, which is necessary for the cell cycle progression from the G1 to the S phase (311). The UTY protein may function as a chaperone and is implicated in protein-protein interactions (312). The USP9Y protein controls protein turnover by inhibiting proteasome-mediated protein breakdown via removing ubiquitin from protein-ubiquitin conjugates (313).

Hence, it also makes sense that these genes have retained their functional roles through strong purifying selection that would purge detrimental mutations. The orthologous comparisons of Y-linked genes between *Carnivora* species displayed that ubiquitous genes were the most conserved among the three gene expression groups, suggesting less differentiation in function than the other two groups of genes.

A slightly lower, but significant, correlation and an unbalanced gametolog expression was observed among low-expression genes. This implies these genes may still share the same ancestral regulatory elements with their X-linked genes, but that these genes' functional importance is diminished at least in the context of providing dosage compensation in heterogametic males. Compared to ubiquitously expressed Y gametologs, low-expression genes had relatively relaxed purifying selection, suggesting a weak preservation of ancestral functions. The expression profiles of low-expression genes blurred an otherwise sharp bipartition defined by ubiquitous and testis-specific genes, suggesting they were likely under a transition stage deriving from ancestral functions to

new functions or dying. For example, *WWC3Y* and *TRAPPC2Y* were expressed widely, but X-biased in most tissues, indicating their somatic functions might be degenerating through being replaced by X-linked genes. *OFD1* and *CYorf15* were co-expressed in balance or even in favour of Y-linked gametologs only in the testis, indicating these two genes might be in transition to becoming testis-specific. Unlike the ubiquitous and testis-specific genes that evolved convergently within *Carnivora* species or even between distant taxa like dog-human that separated over 100 million years ago, most of the low-expressed genes were lineage-specific (e.g. *TRAPPC2Y* is dog-specific, *AP1S2Y* and *WWC3Y* only exist elsewhere in horses, etc). Together with the observation of a high rate of evolution in ortholog comparisons (**Figure 4.9B**), the low-expressed genes were suggested to be not conserved in functions among *Carnivora*.

The testis-specific genes were expressed exclusively or predominantly in this tissue. They had a substantial sequence divergence from their X-linked gametologs, suggesting they have functionally diverged from their X-linked genes. As shown previously, testis-specific genes such as *TSPY*, *RBMY*, and *UBE1Y* took functions in spermatogenesis (120,132,135). It is suggested that once diverted to testis-specific functions, that the process of amplification for testis-specific genes preserved their functions by intrachromosomal gene conversion (1). In this study, four of eight testis-specific genes (*TSPY*, *UBE1Y*, *CUL4BY*, and *BCORY*) occur as multiple-copies within the dog Y chromosome. It is tempting to speculate that adopted testis-specific genes are required for male fertility and sex development, so interchromosomal gene conversion with X-linked gametologs can lead to infertility or lethality. Coincidently, all testis-specific genes displayed a monophyletic clade revealing their single origination during mammalian evolution rather than gene conversion between XY gametologs. In contrast, all six ubiquitous and low-expression genes were polyphyletic, indicating the functions of these genes were historically interchangeable between XY gametologs.

Divergence analysis among *Carnivora* carried the footprints of evolutionary events over tens of millions of years. Orthologous divergence analysis demonstrated that testis-specific genes evolved at a rapid rate within carnivorans, whereas ubiquitous genes were

functionally conserved with a relevant slow divergence, as known, the character of rapid evolution was the hallmark of reproduction-related proteins (314).

In polymorphism analysis, nine missense mutations from five MSY genes (*KDM5D*, *OFD1*, *UBE1Y*, *WWC3Y*, and *BCORY1*) were already fixed in the *Canis lupus* population. It is plausible that genetic drift, Muller's ratchet, hitchhiking effect, or Hill-Robertson interference fixed these variants due to the lack of recombination for the MSY regions. Alternatively, these mutations might be selected for male fitness by developing new functions, as the location of these missense variants appears to occur within proteins' conserved domains (**Figure 4.10**). Of the five genes, four (*KDM5D*, *OFD1*, *UBE1Y*, and *WWC3Y*) showed tolerance for deleterious mutations, which have increased in load relative to synonymous variants (**Figure 4.12**). Taken together, these genes are implied to be less important in functions or they are likely evolving for new functions by accumulating favourable mutations.

In this chapter, the dog Y-specific genes were dissected, including their copy numbers, expression patterns, evolutionary rate, and polymorphisms within dog populations. These genes were divided into three expression categories theorised to correspond to different functional properties. Also, the *SRY* gene had variable copy numbers in dog Y chromosomes, and the palindrome flanking the *SRY* preserved its function from the accumulation of deleterious mutations. Historical events of gene conversion, likely to occur as groups of neighbouring genes at different time points, illustrated the dynamic evolution of dogs' Y chromosomes. The study of *Canidae* representatives, dogs, complements and increases our understanding of the evolution of mammalian Y chromosome genes.

CHAPTER 5: Pseudoautosomal Boundary Origins and Recombination Suppression

5.1 Introduction

The pseudoautosomal region (PAR) is a region of pairing, synapsis, and crossing-over between the X and the Y chromosomes. These chromosomes' PAR sequences share 96%-100% homology and occur on the terminal ends of the chromosomes. The pseudoautosomal boundary (PAB) is the border area where sequence homology reduces, recombination is suppressed, and sex chromosome-specific regions emerge that differentiate the X- and Y-specific chromosomal regions.

PABs were investigated in a few mammalian species. In general, PABs were detected with insertions of repeat sequences, such as primates (*Alu* elements), mice (variable mobile elements), cattle/ sheep (*Bov-tA* elements), horse (mobile elements) and pig (*tRNAGlu* elements) (315). Their gene contents were featured as containing a protein-coding gene on one sex chromosome and a truncated gene on the other. For example, in primates, the *XG* gene spanned the PAB of the X chromosome, but was truncated on the Y-linked PAB (316–318). For mice, the PAB sequence was located in an intron of the *Mid1* gene on the X chromosome, and the *Mid1* on the Y chromosome was truncated (319,320). The cattle and sheep share the same PAB sequences with a *GPR143* gene on the X chromosome and a pseudogenised *GPR143* gene on the Y chromosome (321). The horse PAB was found to be spanned with a coding gene called *XKR3* (43,322). In pigs, the PAB was in the intron of the *SHROOM2* gene, and the *SHROOM2* on the Y-linked PAB was truncated (323).

A recent study of house mice reported that the PAB changed in different subspecies independently, while structural diversity at the PAB was detected in wild house mice (324). A study of primates indicated the PAB of strepsirrhines remained unchanged during evolution, while the PAB moved forward with recombination stopping and the PAR shortened in haplorrhines (325).

A novel gene, *TETY2*, was first detected on the cat's Y chromosome (61), and it was located within the PAB of both dogs and cats (55). The *CLDN34* gene shared the first exon with *TETY2* and because of this, *CLDN34* was regarded as the X-linked gametolog of *TETY2*. Hence, *TETY2* is called *CLDN34Y* in previous studies (55,62).

Despite the dog's X and Y chromosome draft genome (55), our current knowledge regarding the PAB of dogs is limited. In this chapter, the novelty and characteristics of the dog PAB was explored, and its origin was inferred by integrating the PAB sequences from closely related species. The understanding of the PAB will facilitate future work regarding the motivation of sex chromosome dispersal, the mechanisms of recombination suppression, and sexual dimorphism.

5.2 Materials and Methods

5.2.1 Defining the Pseudoautosomal Boundary

In the Falcon assembly step, two primary contigs that potentially covered the PAB were identified. One spanned from the PAR to the female-specific X (FSX) region, and the other one from the PAR to the male-specific Y (MSY) region. The similarity between the two contigs was calculated with BLAST and visualised using the ggplot2 package in R 4.1.1.

The Y-linked sequences around the PAB were subtracted by a window-based method with a window interval of 200 bp and a sliding step of 50 bp. The extracted sequences were searched against X-linked sequences using BLAST with default settings. The similarity of each window was calculated by dividing the length of identical matches by 200 bp.

5.2.2 Phylogenetic Tree for the PAB Sequences in *Canidae*

Sequences homologous to the dog PAB were searched against canidae nucleotide collections and assemblies using the online version of the BLAST tool (https://blast.ncbi.nlm.nih.gov/). The subject sequences were assigned as X-linked or Y-linked according to the scaffolds they belonged to. Sequences that originated from

unplaced contigs or scaffolds were determined to be X- and Y-linked based on their flanking genes.

The multiple alignments of detected *Canidae* PAB sequences were generated using MEGA7 (243) with the MUSCLE algorithm (292). A phylogeny was constructed using the maximum likelihood method based on the Tamura-Nei model with 100 bootstrapping replications. The phylogenetic tree was visualised with iTOL v5 (327).

5.2.3 Sex-Specific Variants Analysis

5.2.3.1 Study Cohort

Male and female dog and wolf Illumina short read data, taken from the public domain, were used in this analysis. Dog data had to meet the following criteria for inclusion: (1) Male and female dogs of the same breeds were used to exclude breed-specific variants, (2) To guarantee a homogenous genetic background, dogs could not cluster ambiguously near wolves or between wolves and dogs in principal component analysis (PCA), (3) genome-wide sequencing depths were >25x, (4) the FSX:autosome depth ratio approximated 1 for females, and (5) the MSY:autosome depth ratio was ~0.5 for males. Having met these criteria, 328 male and 282 female samples were included in this analysis from 97 breeds (**Supplementary Table 5.1**).

5.2.3.2 PAR Variant Calling

Raw reads were pre-processed with fastp v0.22 (301) to prune out low quality reads (average quality score < 20) and trim both ends with low quality. Prepared reads were aligned on the modified RosCfam_1.0 using BWA-mem2 v2.2 (237), and the sequencing depth of autosomes, PAR, FSX, and MSY were estimated by Mosdepth v0.3.3 (256).

Single-nucleotide variants (SNVs) and small insertion/deletion (INDEL) were called within the PAB interval of the X chromosome (NC_051843.1: 1-6,590,648) by the Genomic Analysis Toolkit (GATK) v4.1.7 (249,307). First, each BAM file was processed with Mark Duplicates and RBQS steps to generate analysis-ready BAM. Second, the

HaplotypeCaller program called variants based on the reference genome. Third, a joint-call was performed on all GVCF files to generate a genotype callset for the cohort of dogs. Finally, good quality raw variants were included for the following analysis (SNV criteria: QD > 2.0, QUAL > 30.0, SOR < 3.0, FS < 60.0, MQ > 40.0, MQRankSum > -12.5 and ReadPosRankSum > -8.0; INDEL criteria: QD > 2.0, QUAL > 30.0, FS < 200.0, ReadPosRankSum > 20.0).

5.2.4 Validation of SINE Insertions in the PAB

The PAB is a 5 Kb length sequence between the PAB and the MSY/ FSX regions. Within these regions, two SINE sequences of 200bp each were observed only on the Y-linked PAB. To investigate the evolutionary timing that these two SINEs were inserted into the Y chromosome, three methods were used.

The polymerase chain reactions (PCR) used gDNA from 12 in-house dogs (6 males and 6 females) and 10 male wild canid samples from the National Museum of Scotland (2 dholes, 2 African hunting dogs, 1 golden jackal, 2 bush dogs, 2 maned wolves, and 1 wolf).

For the short read method, NGS data from 8 female and 8 male samples of wild canids were downloaded from the Sequence Read Archive (SRA). The long read method used the in-house PacBio long reads of the Labrador retriever and those of a basenji (SRR11305493).

5.2.4.1 PCR Method

The principle of PCR validation is when the PAB segments containing SINEs are amplified, PCR products of X chromosomes are ~200 bp shorter than that of Y chromosomes. In theory, two bands should be observed in gel electrophoresis in males, and only one smaller band for females if a single SINE element was inserted in the Y-linked PAB. Each paired primer was designed on the flanking regions of each tested SINE. To amplify both X and Y chromosome PABs in the same PCR reaction, primers were designed on the regions where sequences were fully identical between X and Y chromosomes. Primer sequences and annealing temperature are shown in **Supplementary**

Table 5.2, and the thermocycler system and program are present in **Table 2.1** and **Table 2.2**.

5.2.4.2 Short Read Method

Single Illumina reads are normally shorter than 250 bp, hence it is difficult to assemble the PAB sequences, especially where X and Y chromosomes are similar to each other. The short read method was developed to test the presence of SINEs in short-read sequenced samples. All reads were firstly mapped against the modified RosCfam_1.0, and the reads that aligned within the PAB of X and Y chromosomes were extracted for further analysis. Next, these extracted reads were remapped to the Y-linked PAB using customised parameters of minimap2 (--splice -O6,24 -B2 -G500) (238). The "--splice" model enabled X-derived reads "splice-align" by skipping inserted SINEs, the penalty parameters (-O6,24 -B2) tolerated more mismatches therefore allowing the X-derived reads to be mapped on the Y-linked PAB sequences, and the maximum gap option (-G500) was set to detect the existence of SINEs efficiently. By visualising alignment files with IGV, female samples were expected to see that two SINE intervals were skipped by all reads, and male samples' reads should be skipping the SINE loci in roughly half and the remaining passing through.

5.2.4.3 Long Read Method

The long read-based method can either provide validation from long read alignment or from long reads assembled contigs. For the alignment approach, long reads were mapped to the PAB sequence of the Y chromosome with minimap using default settings (-ax map-pb). The existence or absence of SINE insertions was seen according to the alignment viewed in IGV v2.3.90 (300). In the assembly-based approach, the X-linked and Y-linked PAB sequences were obtained from previously assembled contigs. The assembled contigs were aligned against the Y-linked PAB to reveal the presence of SINE insertions.

5.3 Results

5.3.1 Defining the PAB and Gene Contents

In dogs, *SHROOM2* (XM_038449369.1), *CLDN34* (XM_038449372.1), and *TETY2* (JX964858.1) were described as the last PAR gene, the first FSX gene, and the first MSY gene, respectively (55,328). To study the dog PAB, the three genes were searched within the RosCfam_1.0 assembly, resulting in two regions on the X chromosome (NC_051843.1) and an unplaced scaffold (NW_024010443.1). By BLAST searching with the identified sequences as queries, the contig 000111F and 000963F of the Falcon primary assembly were defined, which contained X-linked PAB and Y-linked PAB, respectively (**Figure 5.1**). The identification of two contigs that exhibited dog PAB characteristics enabled the evolution of canids' PABs to be explored.

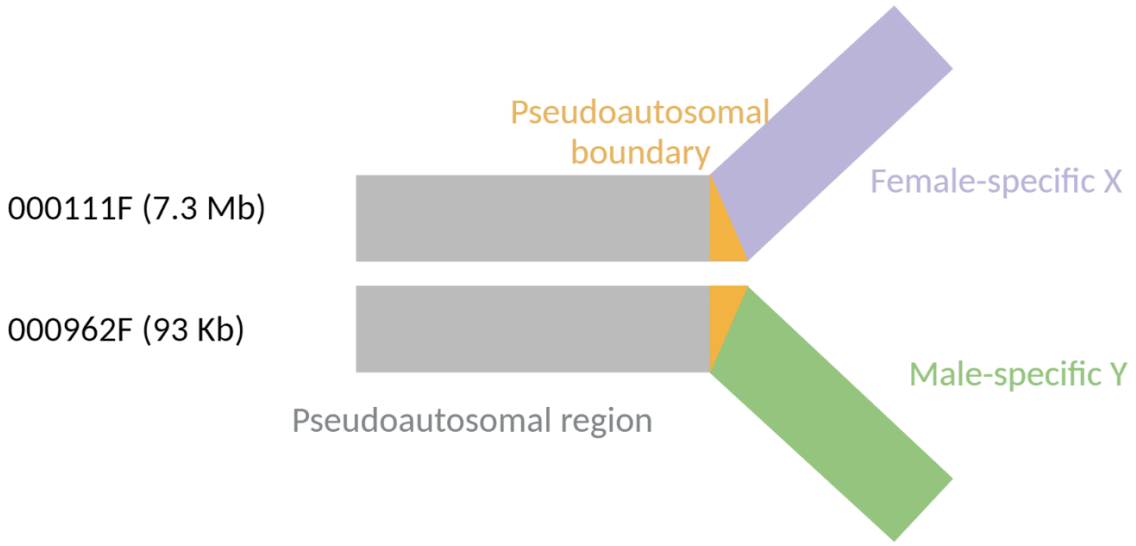


Figure 5.1. Schematic of the definition of the PAB in RosCfam_1.0 assembly. Falcon generated two primary contigs that belong to the X and Y chromosomes and PAB. The recombination region (i.e. PAR) is coloured grey, the transitional region is PAB (yellow), and the blue and green regions are Female-specific X and the Male-specific Y regions respectively.

From the beginning of the contigs, 5 Kb of nearly 99% identical sequence was identified; this was interpreted to correspond to the distal PAR. Following this stretch, homology between the contigs oscillated for ~5 Kb and then precipitously dropped to nearly zero (**Figure 5.2**). The dog PAB was located downstream of *SHROOM2* and

contains the 5' untranslated region (UTR) of *CLDN34*. The X-linked *CLDN34* was transcribed across various tissues including the testis. *CLDN34* expression originated from two transcriptional start sites (TSS) (**Figure 5.3A**). In contrast, the Y-linked *CLDN34* was truncated, consisting of four untranscribed exons (**Figure 5.2**, **Figure 5.4**). Downstream of the Y-linked *CLDN34* exons, on the same strand, was a copy of *TETY2*. Unexpectedly, the first exon of *TETY2* overlapped with the last exon of Y-linked *CLDN34* (positionally this Y-linked CLDN34 exon corresponds with the fourth exon of X-linked *CLDN34*). RefSeq annotation showed exon 4 was the initial exon of a *CLDN34* transcript (isoform 3 in **Figures 5.4A and 5.5A**) and was exclusive from other isoforms. Moreover, exons 2 and 4 of *TETY2* had homologous, but untranscribed sequence, on the X chromosome. These were located within the PAB and 24 Kb from the PAB, respectively (**Figure 5.2**).

To further explore if *TETY2* and *CLDN34* were ancestrally related, TBLASTN searches against monotremes with the *TETY2* and *CLDN34* protein sequences as queries were conducted. As a result, *TETY2* had an orthologous gene (*LOC103170854*) that was a different gene from *CLDN34* in the platypus. Moreover, coding *TETY2* only survived mammalian Y chromosome evolution in marsupials, afrotherians, and carnivorans and the independent evolution of both genes was seen in mammals (**Figure 5.6**).

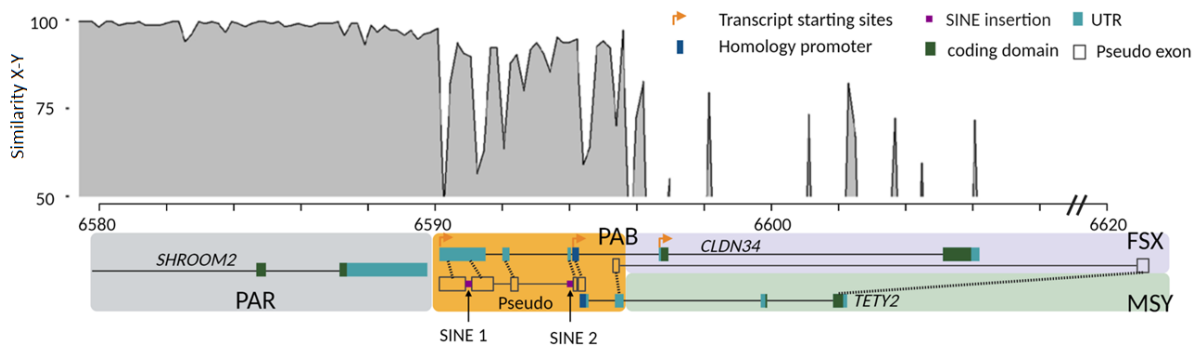


Figure 5.2. Similarity and gene contents in PAB and flanking regions. The similarity between X and Y chromosomes oscillated within the PAB and dropped to 0 in the chromosome-specific regions. *CLDN34* on the Y chromosome has been pseudogenised.

Two SINE repeats were inserted into the pseudo-exon and pseudo-intron, respectively. *TETY2* survived on the Y chromosome, however within the X-linked PAB, this gene is pseudogenised. Besides, the first exon of the *TETY2* had homologous sequences with an exon of *CLDN34* (blue).

Figure 5.3. Visualisation of CAGE-Seq data in the PAB. (A) CAGE-Seq peaks were called on X-linked PAB and visualised with ZENBU browser. A broad peak is observed corresponding to isoform 1 and isoform 2 in other tissues, and sharp peaks matching isoform 3 and isoform 4 are only seen in testes samples. Isoforms refer to **Figure 5.5A**. (B) CAGE-Seq peaks are called on Y-linked PAB. A sharp peak is called at the starting site of *TETY2*.

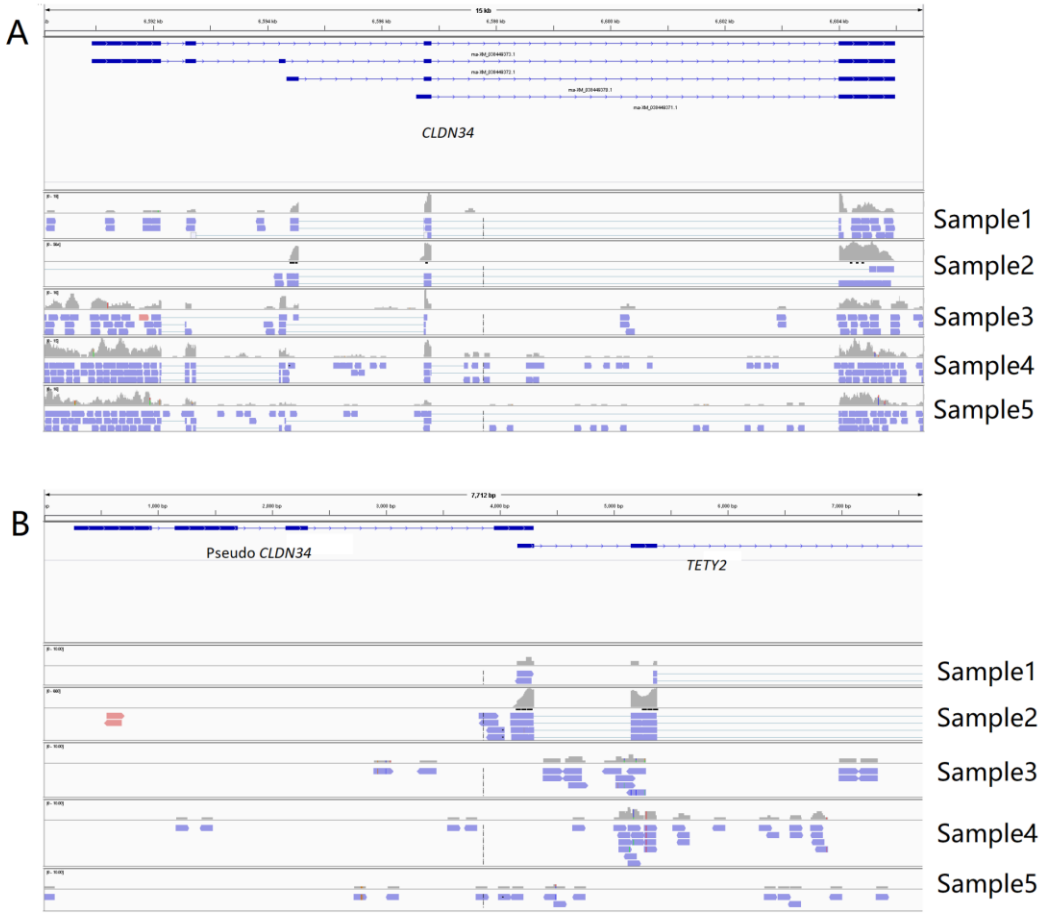


Figure 5.4. RNA-Seq alignment in the PAB. Visualization of RNA-Seq data in the PAB of X chromosome (A) and Y chromosome (B). Panels are screenshots taken from IGV. Gene models were displayed on the top for each.

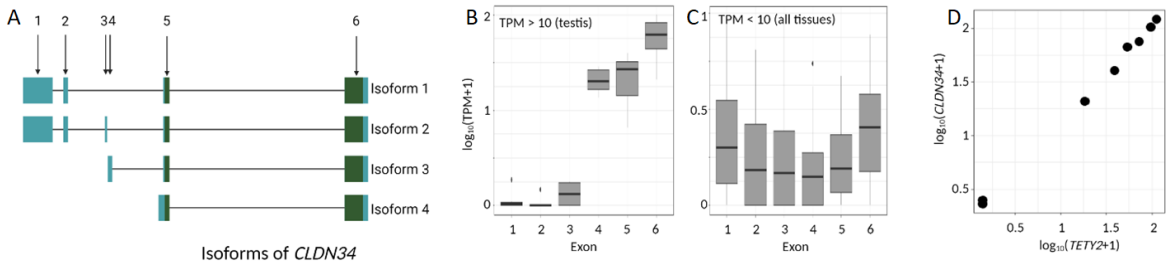


Figure 5.5. Expression of the *CLDN34* and *TETY2* in 94 male dogs. (A) *CLDN34* is annotated as four isoforms. Note that *CLDN34*'s exon 4 occurs next to a testes-specific TSS/promoter (see in **Figure 5.4**) and that this exon is also utilised by Y-linked *TETY2* (see in **Figure 5.3**). (B) Expression of *CLDN34* by exons in *TETY2* high-expression samples. In testis, where the *TETY2* is expressed with a TPM of more than 10, the last three exons have a high TPM indicating that isoform 3 is the dominant isoform expressed in this tissue. (C) Expression of *CLDN34* by exons in *TETY2* low-expression samples. In all other tissues except testis, *TETY2* TPM counts are low. Here exon 1 has the highest TPM, suggesting that isoforms 1 and 2 are dominant. (D) Co-expression of *TETY2* and *CLDN34* in testis. A scatter dot plot between the TPM of *TETY2* and *CLDN34* shows a high correlation of expression between the two genes.

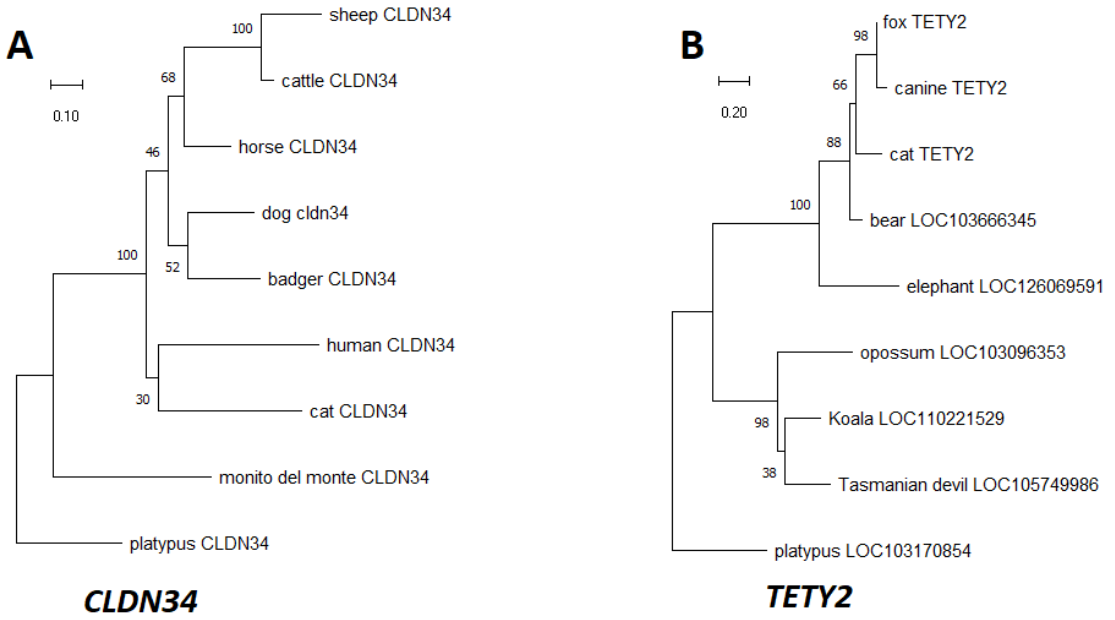


Figure 5.6. Independence of the *CLDN34* and *TETY2* by phylogeny. Phylogenetic tree of mammal species based on protein sequences for the *CLDN34* (A) and *TETY2* (B).

5.3.2 Duality of *CLDN34* Expression

Despite their different evolutionary histories, X-linked *CLDN34* and Y-linked *TETY2* appear to utilise an exon in common. This unusual phenomenon was investigated using transcriptional data. Gene abundance was quantified based on 94 RNA-seq data from 22 tissues from male dogs. Complementing this data were CAGE-Seq reads (**Section 2.4**) which helped to define transcriptional start sites (TSSs). *TETY2* was a tissue-specific gene that was only expressed in the testis with a sharp TSS (**Figure 5.3B**). On the contrary, *CLDN34* exhibited a duality in its expression. On one hand, *CLDN34*'s isoform 3 appeared tissue-specific, with a sharp peak of TSS that was only detected in the testis (**Figure 5.3B**). In testis, where the *TETY2* was expressed with a TPM of greater than 10, the last three exons had a high TPM indicating that isoform 3 was predominantly expressed (**Figure 5.5B, Supplementary Table 5.3**). On the other hand, it was ubiquitously expressed across all tissues with a broad shape TSS corresponding to isoforms 1 and 2 (**Figure 5.3A**). When looking at samples with low *TETY2* expression (TPM < 10), exon 1 had the highest TPM,

suggesting isoforms 1 and 2 are dominant (**Figure 5.5C, Supplementary Table 5.3**). Comparing testes with other tissues, we observed that the expression level of isoform 3 in testis was more than one order of magnitude higher than that of isoforms 1 and 2 in other tissues (**Figure 5.7**). Furthermore, the expression of *TETY2* and *CLDN34* was highly correlated for each testis sample (**Figures 5.5D and 5.7**), whereby other tissues did not present such a trend (**Figure 5.7**).

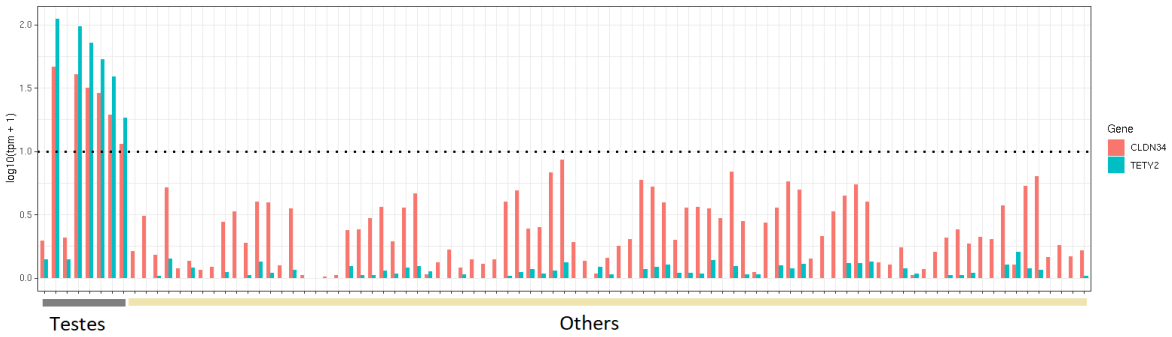


Figure 5.7. Expression of the *CLDN34* and *TETY2*. A wide range of tissues were quantified by RNA-Seq data. Bar plots indicate the TPM of *CLDN34* (red) and *TETY2* (blue). Tissues included the tistes, adrenal gland, stomach, heart, skin, cerebellum, lung, kidney cortex, colon, pituitary gland, head, salivary gland, liver, spleen, cartilage, pancreas, small intestine, bladder, bone marrow, lymph node, adipose, and skeletal muscle.

In addition, our interest further expanded to the expression pattern of *CLDN34* in female samples. The expression of *TETY2* and *CLDN34* was estimated from 7 tissues of female dogs (**Supplementary Table 5.4**). First, an absence of reads mapping to *TETY2* exons in female samples indicates the unique mapping of RNA-seq reads within the PAB between X and Y chromosomes, which can exclude the possibility of false discovery of correlation due to multiple alignments or misalignment. Second, exon 4 showed significantly lower expression than exon 1 indicating isoforms 1 and 2 were predominantly expressed for female samples (**Figure 5.8**). Third, *CLDN34* was ubiquitously expressed in females as well.

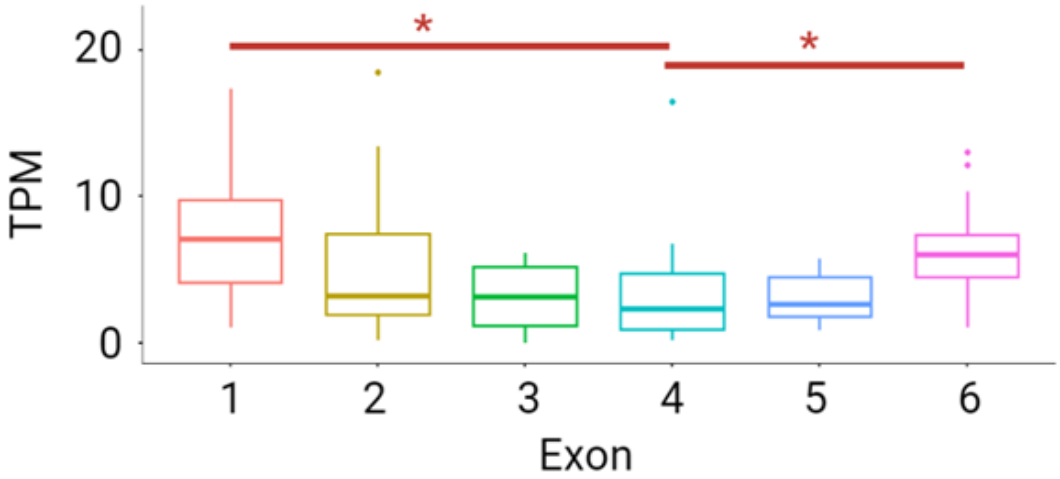


Figure 5.8. Expression of the *CLDN34* for each exon in female samples based on RNA-Seq. Exon 1 and exon 6 have significantly higher expression levels than exon 4 indicating that the dominant expressions were from isoform 1 and isoform 2.

5.3.3 Recombination Ceases and the PAB Origins

Across mammals, the PABs' loss of homology delineates where the X and Y chromosomes progressively lose their ability to recombine during meiosis. Beyond the PAB are the MSY and FSX, regions where meiotic exchange is unlikely. To investigate how the recombination rate changed around the PAB, sex-specific variants were counted across the PAR. The rationale is that assuming two mutations occurred on a Y-linked PAR, one is distal to the PAB and the other one is proximal to the PAB. If the meiotic recombination is suppressed around the PAB, the proximal variant is only observed as male-specific which is linked with the Y chromosome. In other words, the distal variants had more chance to jump on the X chromosome due to more frequent crossover events between sex chromosomes (**Figure 5.9A**). After a long-term recombination suppression since the PAB formation, the proximal PAB of Y chromosomes should accumulate more male-specific variants than distal regions in the dog population (**Figure 5.9B**). In agreement with this hypothesis, male-specific variants were enriched when close to the PAB. The minor allele frequency (MAF) of specific variants increased with proximity to the PAB, some of which were fixed with a MAF of 0.5 at the region 6 Kb away from the PAB displaying a complete linkage disequilibrium with Y chromosomes (**Figure 5.10**).

Comparative analysis of the PAB with more related species is required to pursue the origin of the dog PAB into even earlier Caniformia ancestors. Homologous sequences of the dog PAB in the Arctoidea were only detected on the X chromosomes. Both the X and Y chromosomes in the Canidae species, including the fox, maned wolf, African wild dog, and grey wolf, showed homologous sequences to the dog's PAB sequences. As a result, the maximum likelihood tree based on the identified sequences within the Caniformia species was constructed, revealing the Y-linked PAB of Canidae to be generated after the split of the Canidae and Arctoidea (**Figure 5.11**). This result also suggested a different PAB existed for Arctoidea's sex chromosomes.

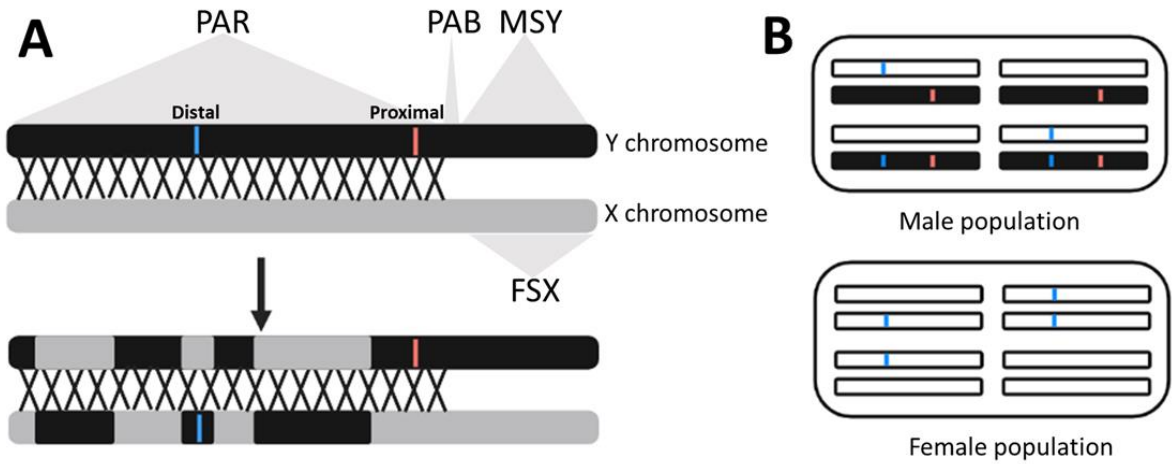


Figure 5.9. Rationale of sex-specific variants analysis. (A) Assuming two mutations occurred on a Y chromosome PAR simultaneously, one distal to PAB and the other proximal. If the meiotic recombination is suppressed around PAB, the proximal variant is only observed as male-specific which is linked with the Y chromosome. The meiotic recombination enables X chromosomes to carry Y chromosome-derived variants after many generations of delivery, but this also depends on the efficiency of recombination. (B) By investigating genotypes of the population, the region with more male-specific variants is indicated as recombination suppression between X and Y chromosomes.

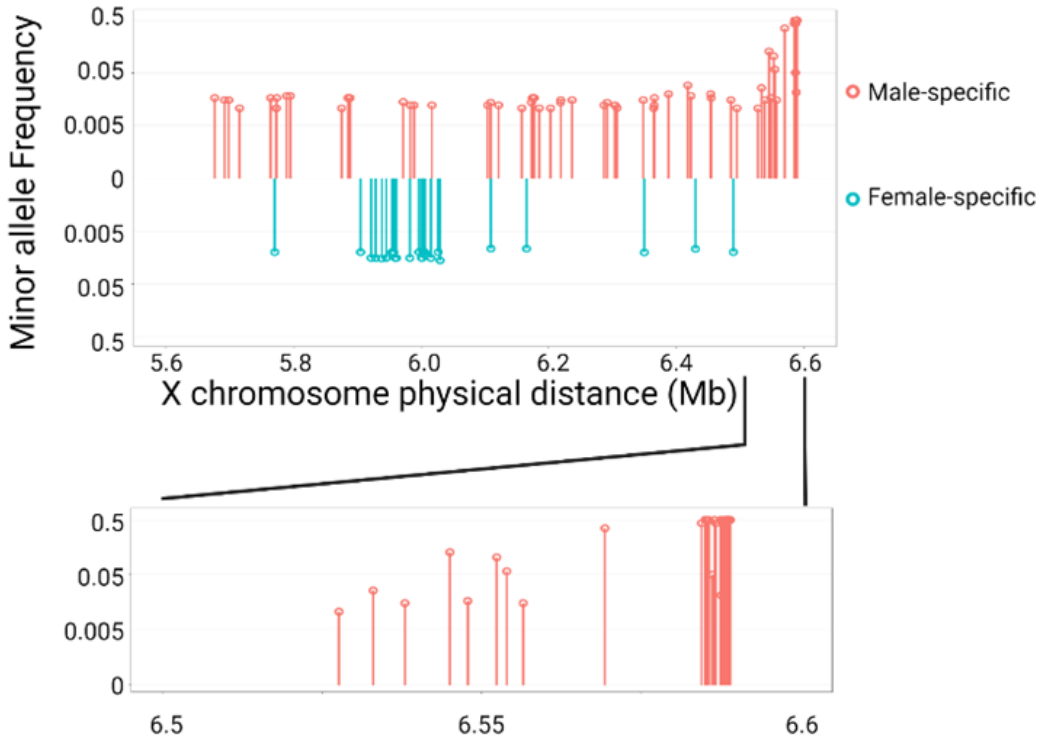


Figure 5.10. The distribution and MAF of sex-specific alleles in the PAR closing to the boundary.

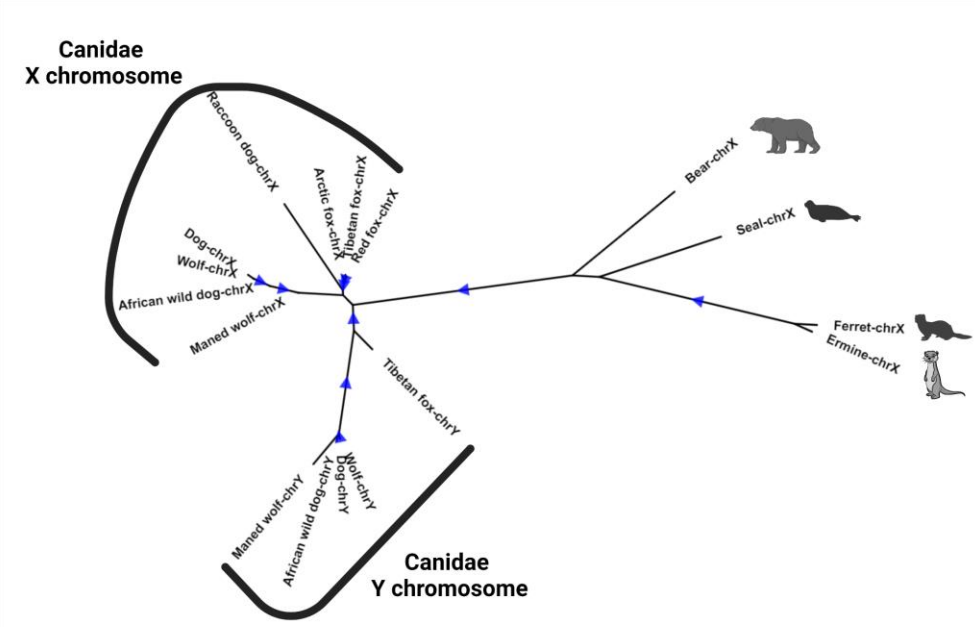


Figure 5.11. Phylogenetic tree based on PAB sequences for Caniformia. The Canidae's sequence is divided into X-linked and Y-linked PABs. The Arctoidea species retains the corresponding sequence in the PAB without sex chromosome specificity. Blue triangles refer to a bootstrapping value over 0.8.

5.3.4 SINE Activity at the PAB

Two SINE elements (called SINE1 and SINE2) were observed within the Y-linked PAB (**Figure 5.3**). Both retrotransposon sequences belonged to *Canis*-specific SINEs (329), at a similarity of 95.7% and 92.2% with SINEC2A1_CF and SINEC2A1_CF, respectively. Identical SINE1 insertions were detected in the Y-linked PAB of wild canids, maned wolves, and foxes, whereas the SINE2 was only detected among wild canids. Based on homology to their respective consensus sequences and their phylogenetic emergence among canids, it is inferred that SINE1 is younger, having derived from a common ancestor of the *Canis* species around 5 million years ago (MYA). SINE2 is relatively more diverged from its consensus and it was identified among all the Canidae species that were tested. This indicates that SINE2 insertion occurred at least 12 MYA, before the Canidae species split (**Figure 5.12**). Lending support to this interpretation, SINE1 had a relatively intact structure (**Figure 5.13, Figure 5.14**) and, by quantifying particles in plaque assays, its transposition activity was approximately 60% that of the SINEC_Cf consensus. SINE2 lacked its target site duplication (TSD) and had lost the capability of mobilizing (**Figure 5.15**). The transposition activity experiments were performed by Sarah Emery (University of Michigan Medical School) and Jeffrey Kidd (University of Michigan Medical School) based on the protocol from the Julia et. al. study (330).

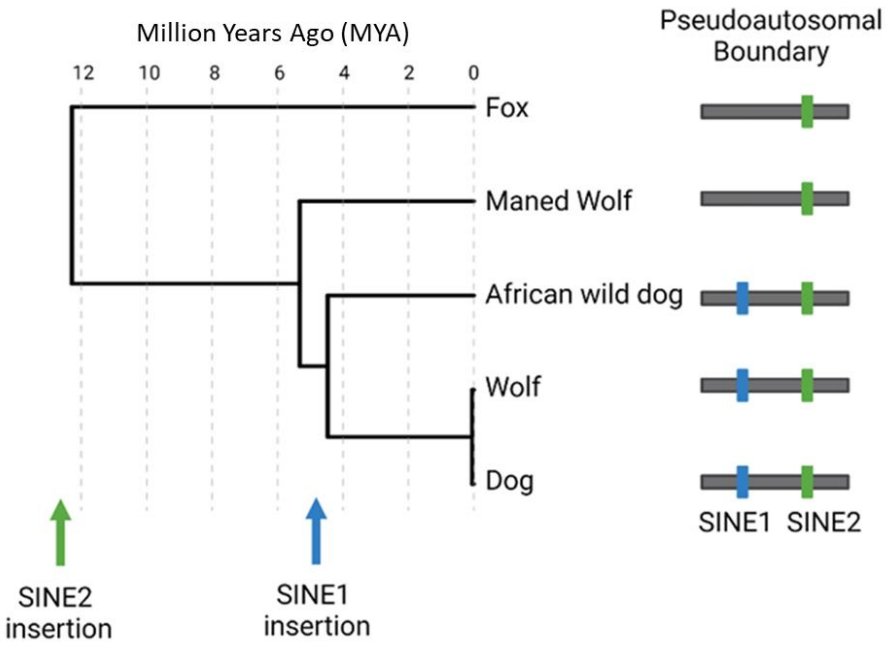


Figure 5.12. The evolution of SINE insertions in the PAB. SINE2 inserted into the Y-linked PAB before the Canidae split, while SINE1 inserted after the split that formed ancestors of Canina species (e.g. African wild dog, wolf, and domestic dog).

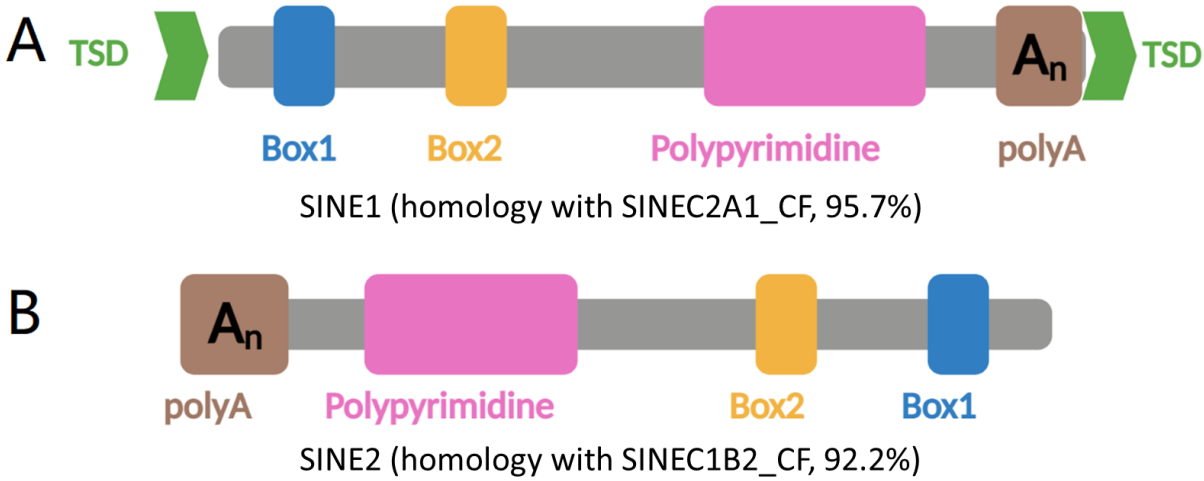


Figure 5.13. Structure of SINEs in PAB.

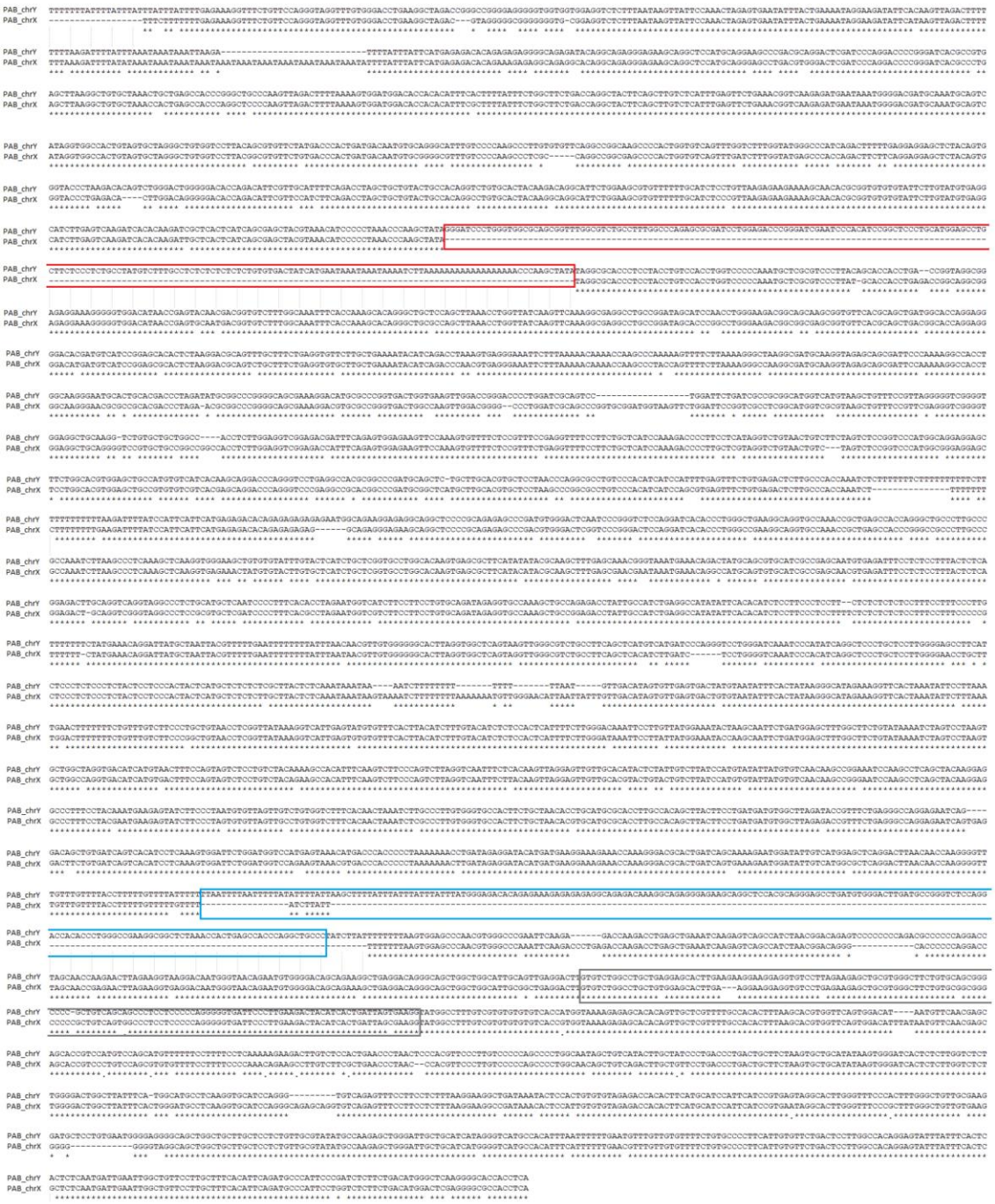


Figure 5.14. DNA sequence of the PAB for dogs. SINE1 (red box), SINE2 (blue box). The shared exon of *TETY2* and *CLDN34* is highlighted (black box). (A high resolution version is available on https://github.com/WengangXbio/script_bio/blob/main/Figure%205.14.png)

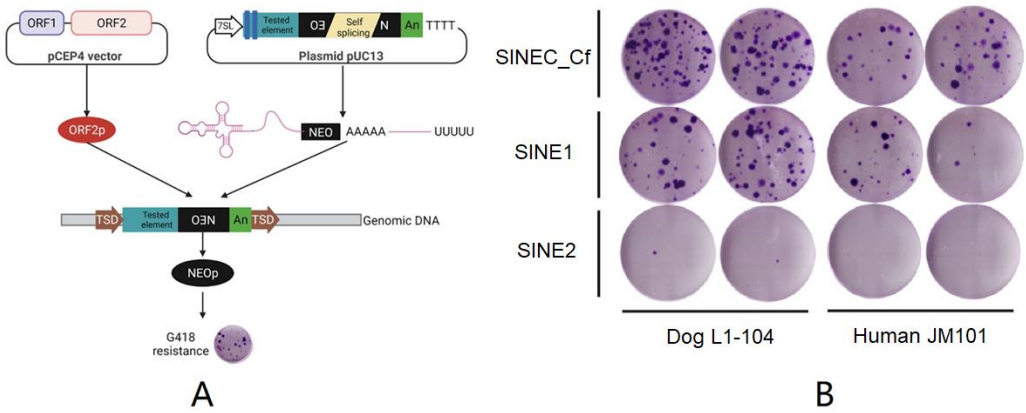


Figure 5.15. Identification of SINE elements capable of retrotransposition. (A) Schematic representation of the *in vitro* assay. Plasmids equipped with a retrotransposition indicator cassette contained the Neo resistance gene which contains an intron. If a SINE element remains retrotranposition-competent, the transcribed gene will be reconstituted and inserted into the genome. Successful transposition into the genome confers antibiotic resistance to G418. (B) Result of the retrotransposition assay using the canonical SINEC_Cf as a positive control (top row), SINE1 (middle row), and SINE2 (bottom row). Each SINE is tested through transfection into human HeLa cells. The mobility of SINE1 is roughly 60% of the consensus SINE sequence, whereas the mobility of SINE2 is nearly entirely abolished.

To test whether two SINEs were broadly distributed or fixed in Canid species, three approaches were developed (details in **Section 5.2.4**). As a result, all three methods proved the two SINE insertions in dogs and wolves, and at least one method demonstrated the existence of both SINEs in other wild canids (**Figures 5.16-5.20, Table 5.1**).

Table 5.1 Evidence of two SINE insertions in dogs, wild Canids, and close species.

Species*	SINE 1			SINE 2		
	Short-read	Long-read	PCR	Short-read	Long-read	PCR
Dog	Yes	Yes	Yes	Yes	Yes	Yes
Wolf	Yes	Yes	Yes	Yes	Yes	Yes
Dhole	Yes	n.a.	Yes	Yes	Yes	No
Coyote	Yes	n.a.	n.a.	Yes	n.a.	n.a.
Jackle	Yes	n.a.	No	Yes	n.a.	No
Ethiopian wolf	Yes	n.a.	n.a.	Yes	n.a.	n.a.
African wild dog	Yes	Yes	No	Yes	Yes	No

Maned wolf	n.a.	No	Yes**	n.a.	n.a.	Yes*
Fox	n.a.	No	n.a.	n.a.	Yes	n.a.

*Samples are describe in **5.2.4**.
**This is likely falsely discovered in the maned wolf.

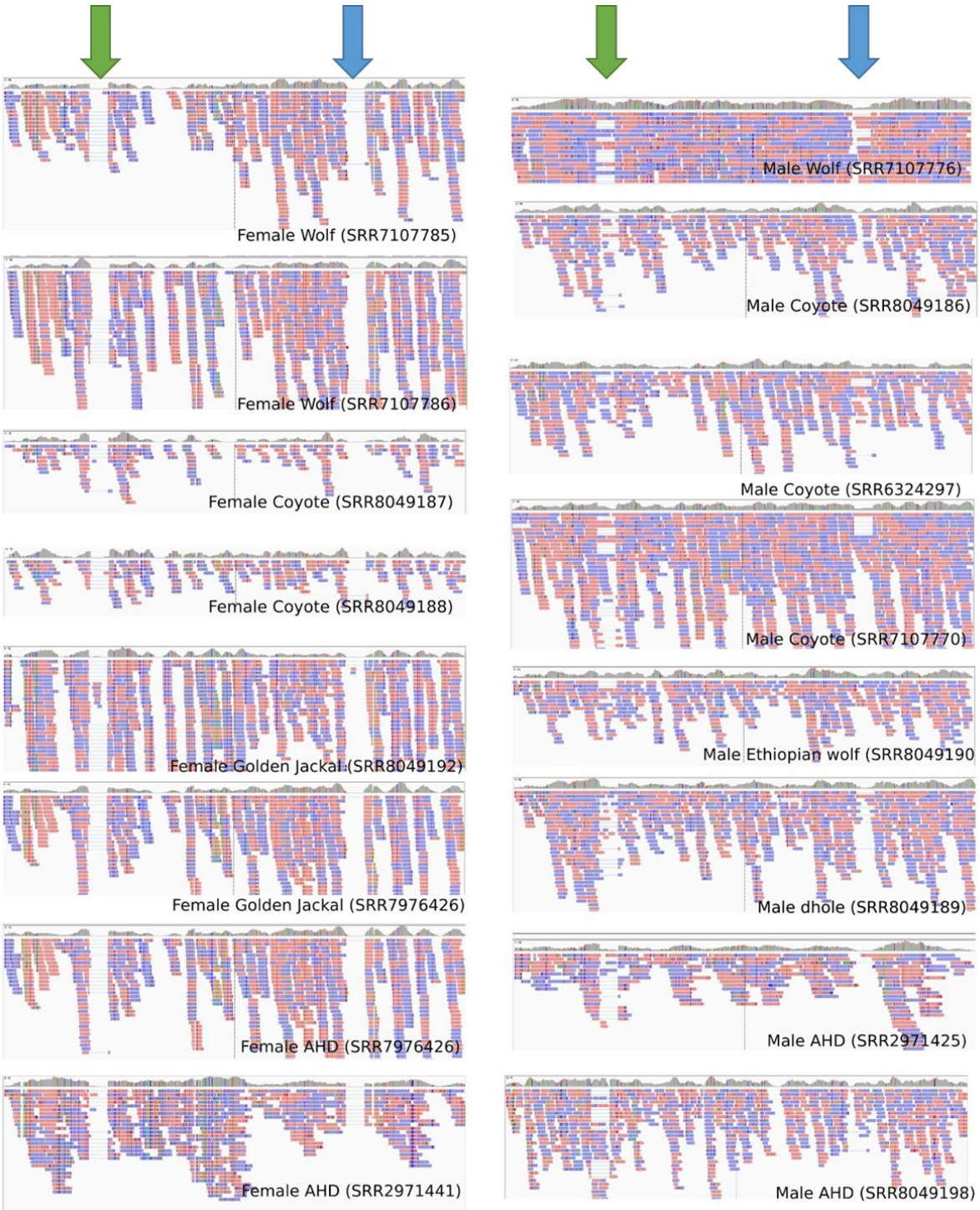


Figure 5.16. Validation of the insertion of two PAB SINEs using short read sequencing data. Visualisation of short read alignment in the PAB produced from male and female wild

canids. AHD, African hunting dog (African wild dog). Archive run accession numbers are provided. The green and blue arrows present SINE1 and SINE2, respectively.

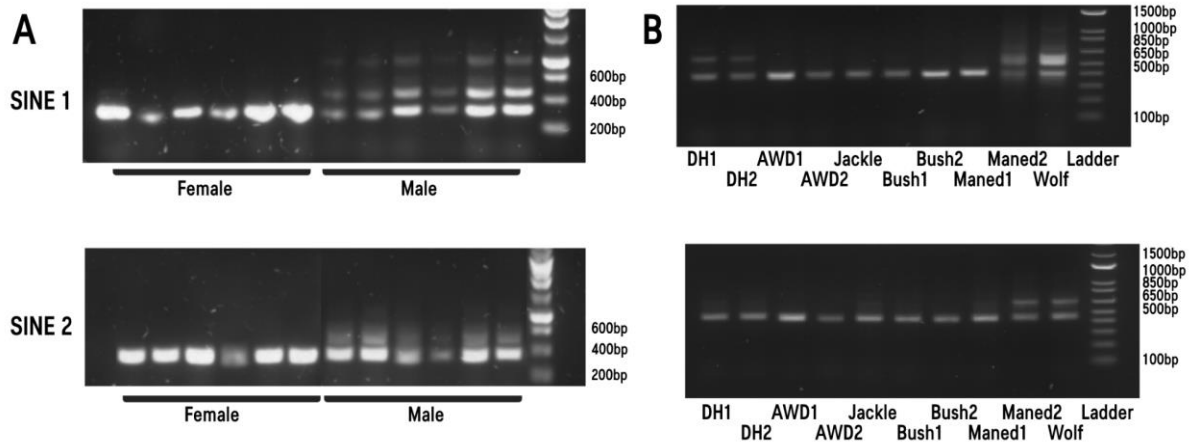


Figure 5.17. Validation of the insertion of two PAB SINEs by the PCR method. (A) Results of six male and female domestic dogs. (B) Results of wild canids. DH: dhole, AWD: African wild dog, Bush: Bush dog, Maned: Maned wolf. Samples are described in **5.2.4**.

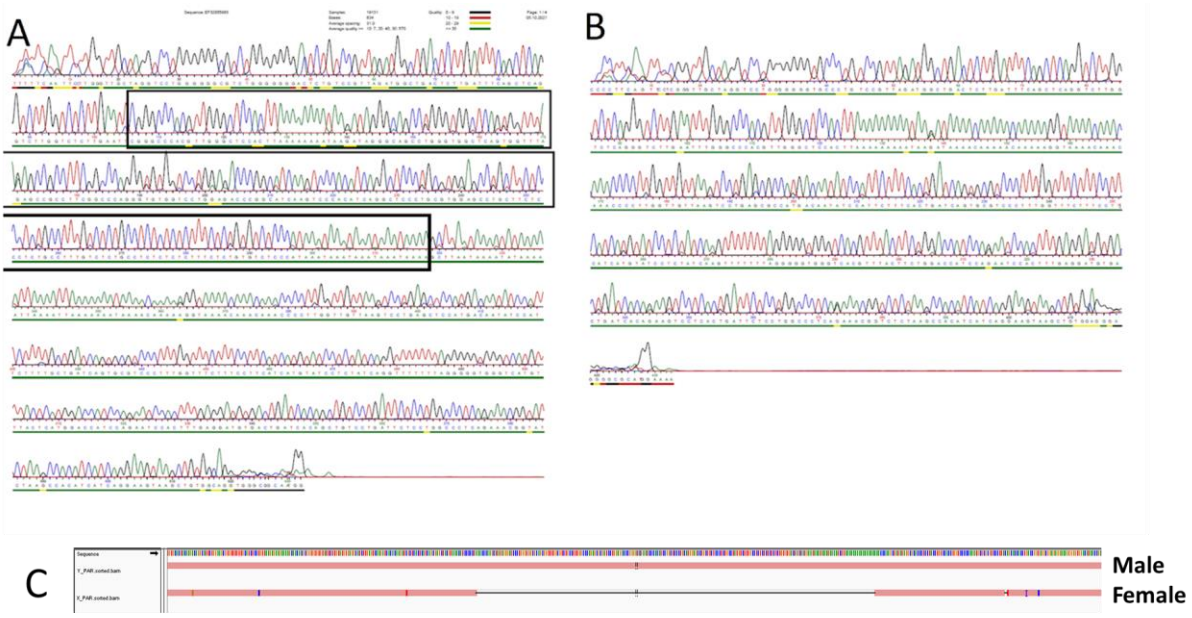


Figure 5.18. Sanger sequencing of SINE1 PCR products. (A) DNA sequences of Y-specific products from male dog samples (long segments). (B) DNA sequences of X-specific products from female dog samples. (C) Alignment of sequenced PCR products against Y-linked PAB showing the absence of SINE in females.

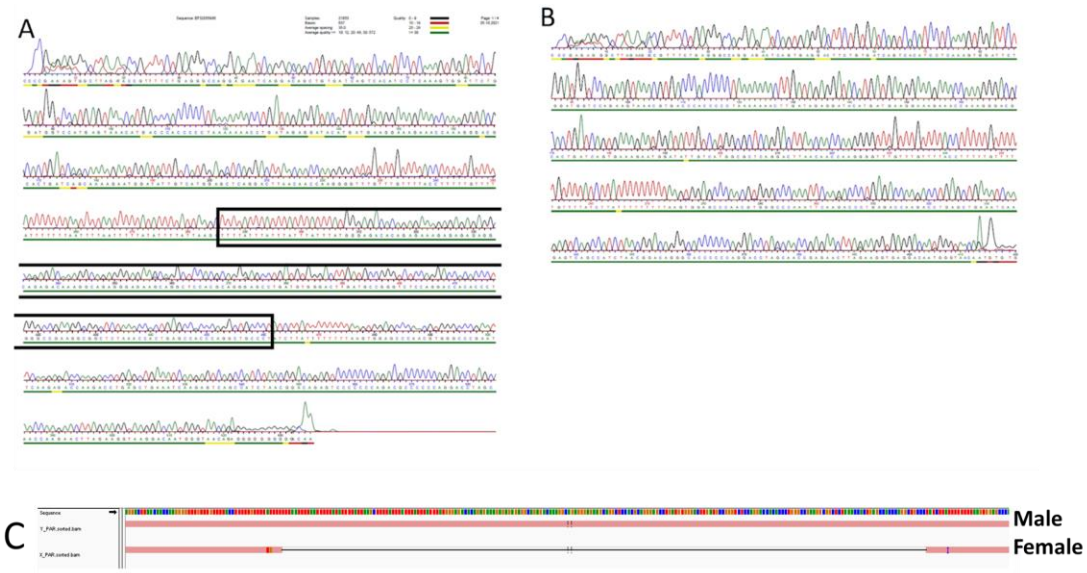


Figure 5.19. Sanger sequencing of SINE2 PCR products. (A) DNA sequences of Y-specific products from male samples (long segments). (B) DNA sequences of X-specific products from female samples. (C) Alignment of sequenced PCR products against Y-linked PAB showing the absence of SINE in females.

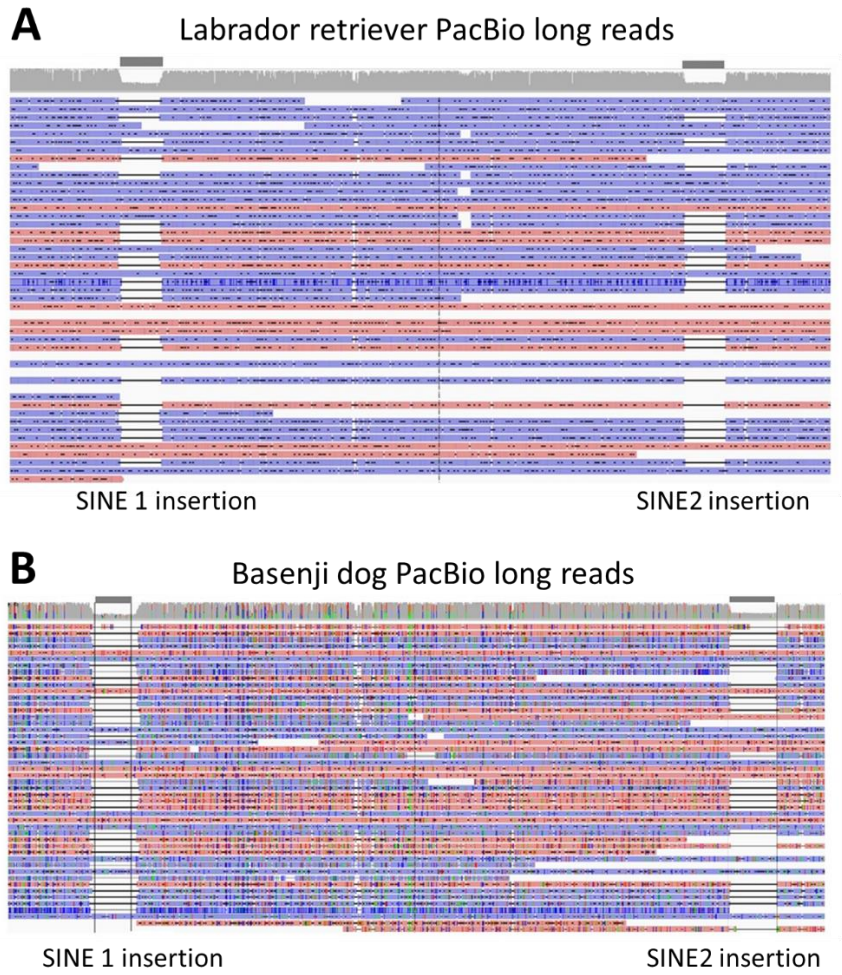


Figure 5.20. Validation of the insertion of two PAB SINEs by the long reads method. Labrador retriever (A) and basenji (B) PacBio long reads are mapped within PAB and visualised.

Intuitively, a 200 bp length SINE insertion contributed more heterozygosity than nucleotide base changes such as SNV and INDEL. Therefore, the accumulation of SINE elements near, or within the PAB, could facilitate X and Y chromosomes evolving dispersedly. SINE accumulation was significantly enriched at the interface between the PAB and sex-specific regions. This enrichment extended nearly 0.8 Mb from the PAB. SINE enrichment was also observed between the PAB and proximal PAR, albeit to a lesser extent (**Figure 5.21A**). On the dog MSY, an abundance of SINE sequences was observed within 1 Mb regions from PAB, which was significantly higher than distal regions. In contrast, SINE distributed uniformly across FSX without difference in statistics.

Also, for each SINE around the PAB, the similarity with consensus sequences of *Canis*-specific SINEs was calculated (**Figure 5.21B**). A total of 27 consensus SINEs were downloaded from Repbase (284) including SINEC1B2_CF, SINEC2_AMe, SINEC2A1_CF, SINEC2A2_CF, SINEC1D_CF, SINEC1C1_CF, SINEC1_CF, SINEC_a1, SINEC_a2, SINEC_b1, SINEC_b2, SINEC_c1, SINEC_c2, SINEC_Cf, SINEC_Cf2, SINEC_Cf3, SINEC_Fc, SINEC1_AMe, SINEC1B_AMe, SINEC1C2_CF, SINEC_Fc2, SINEC_Mv, SINEC_Pv, SINEC_old, SINEC_Fc3, SINEC1A_CF, and SINEC1B1_CF. Overall, the degree of similarity among the PAR, FSX, and MSY was the same on average, but the similarity shifted in density to around 95% for the MSY regions.

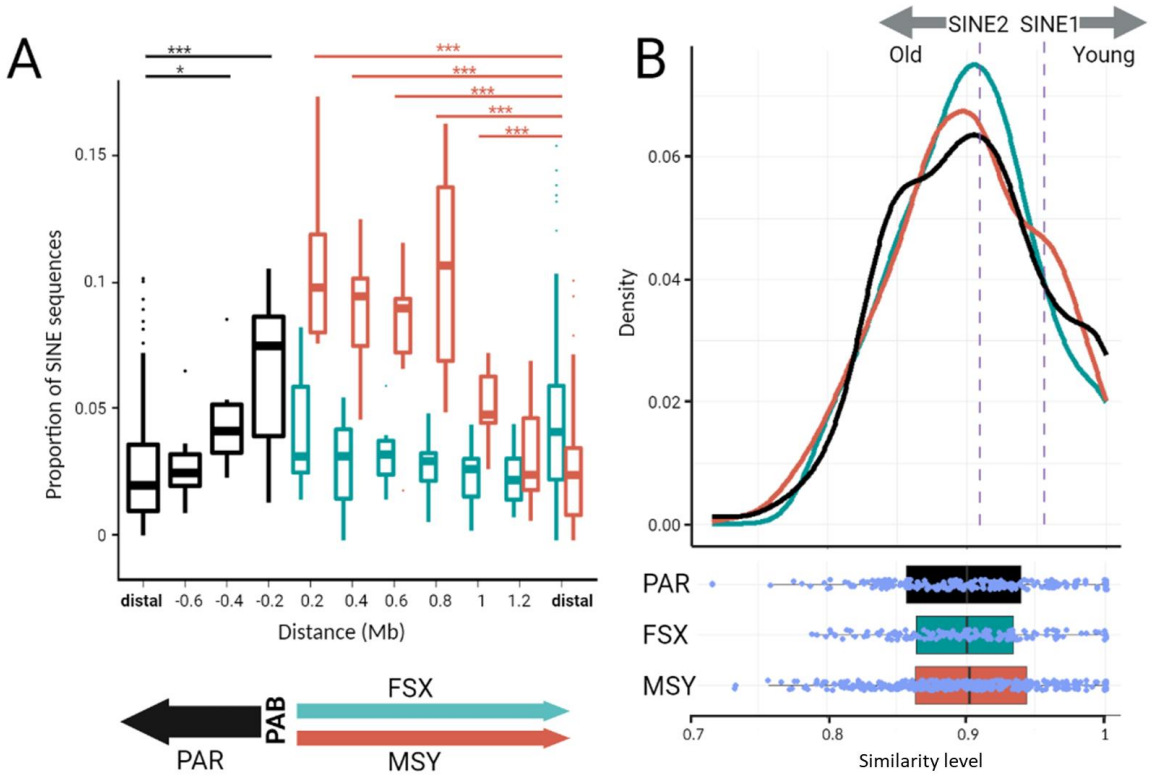


Figure 5.21. SINE distribution around the PAB. (A) The proportion of SINE sequences around PAB with a classification of PAR, FSX, and MSY. The x-axis is centred on the PAB; bins closest to the PAB are considered proximal, those furthest away are considered distal. A t-test was used to examine the significance between groups, and asterisks refer to a difference between two groups in a statistic. (B) The density plot of similarity level of SINE sequences to their consensuses in the PAR, FSX, and MSY. The similarity of SINE1 and SINE2 is highlighted with purple dash lines respectively.

5.4 Discussion

In this chapter, *CLDN34* and *TETY2* were revealed to be different genes, which evolved independently after the split of sex chromosomes. First, *CLDN34* and *TETY2* occur as distinct autosomal orthologs in the platypus, a basal mammal. Although in this species they are found on the same chromosome, the two genes are separated by a physical distance of 2.6 Mb, so they cannot be the same genes. Second, where the dog genes appear to share homology corresponds with the UTR; in other words, the two genes share no coding sequence. Thirdly, dog's Y-linked *TETY2* and X-linked *CLDN34* have corresponding gametologs which have been pseudogenised. The exonic relics of these genes were detected. *TETY2* is only expressed in the testis, indicating its function may be related to spermatogenesis. The isoform 1 and isoform 2 of *CLDN34* were expressed

ubiquitously, in agreement with its broad TSSs (300). In contrast, isoform 3 was testis-specific with a sharp peak which was expressed in high correlation with *TETY2* in testes samples. *TETY2* and isoform 3 of *CLDN34* appear to share the same promoter; perhaps during male evolution *TETY2* recruited *CLDN34* isoform 3 promoter located at exon 4. While it is also known that *CLDN34* was mainly expressed in testis for humans (332) and mice (333), we cannot reject the possibility that their coexpression is coincidental (i.e. both genes are expressed during spermatogenesis).

Next, as an enrichment of SINE elements was seen around PAB with a novel distribution, we explored how the SINE insertions worked on recombination inhibition. Recombination suppression between sex chromosomes is often established around sex-determining and sex-specific loci by structural heterozygosity caused by inversions and translocations. The effects of these multi-megabase variants occur in a stepwise fashion rather than dispersed gradually. In mammals, the non-recombining Y chromosomal segments degenerated by losing and pseudogenising ancestral genes, leaving a limited recombining region at the end of the chromosome, which is called PAR (334). As a consequence, the homology between the dog X and Y chromosomes is high in the PAR, oscillates and recedes in the PAB, and then plummets between the non-recombining regions of the two sex chromosomes. The ongoing shift of the PAR in primates and structural variation of PAB in mouse populations reveals that the non-recombining region is still expanding, and the differential length and gene content of PARs among mammals implies the evolutionary rate at PAB differs (315). Within the PAB, TE insertions were commonly described in mammals such as humans (316), cattle (321), and pigs (328). The TE insertions and recombination suppression coevolved with positive feedback (335), therefore it is unclear whether halting recombination accumulates, or is a consequence, of TE insertions. One tantalising clue comes from birds. Songbirds' W-linked PAB had an insertion of CR1-E1 elements and pieces of evidence indicate accumulation of transposable element (TE) is the cause, rather than the result, of halting recombination (336).

In this study, SINEs accumulated around the PAB, especially within the MSY proximal to the PAB. As our results show, other species within Canidae shared the same PAB as dogs, indicating their formation occurred after the split of the Canidae and Arctoidea, and before Canidae speciation (12 MYA - 45 MYA). SINE1 was estimated to have originated around 5 MYA and SINE2 was at least 12 MYA based on the phylogenetics of Canidae speciation. These two SINEs allowed us to infer ages of other SINEs, which were reflected from similarities with their respective consensus sequences. The SINE sequences around the PAB were more similar to their respective consensus sequences than SINE1 (95.7%), suggesting that they might be inserted later than the insertion of SINE1 (< 5 MYA). Also, the proportion of young SINEs (similarity > 95%) in the MSY was higher than that in the PAB and FSX, suggesting that the recombination-free MSY was more accessible to insertions by retrotransposons and under less purifying selection. Overall, the estimated insertion time of SINE1 (5 MYA), together with even younger SINEs, particularly those enriched at the MSY, suggests that reduction of the recombination rate facilitated SINE retrotransposition.

Alternatively, several patterns suggest that the accumulation of SINEs facilitated suppression of sex chromosome recombination. First, the exact order of the creation of PAB and the insertion of SINE2, both of which took place in the Canidae's common ancestor more than 12 MYA, is not known. Based on consensus homology, there are SINE sequences older than SINE2 around the PAB, whose similarity was lower than 92.2%, implying some of the SINEs inserted before recombination suppression. Second, only SINE repeats, but not LINEs, appear enriched around the PAB. Longer elements, such as LINEs, are known to have a stronger purifying selection when recombination exists (337–339). As observed, LINEs occur in abundance distal to the PAB, with a surprising decrease proximal to the PAB. Hence, these patterns are unlikely if most SINEs emerged after cessation of recombination, and by extension purifying selection. Rather, one would expect to observe an equal ratio of SINEs and LINEs around the PAB which is comparable to what is observed in the distal MSY. Finally, in accordance with PAB's movement towards PAR (340,341), the concentration of SINE elements on the proximal PAR is

elevated, as one would expect in a model where SINE insertions presage loss of recombination. Overall, we propose that SINE insertions served as a driving force for halting recombination in dogs, in turn, recombination-free regions accumulated SINE repeats, especially on the MSY, that had sex chromosomes to evolve dispersedly.

This analysis makes additional predictions concerning the loss of *CLDN34* on the Y chromosome. It is tempting to speculate that the two SINE repeats detected within the PAB are supposed to induce the dispersal of the ancestral *CLDN34* on different sex chromosomes in dogs. In the beginning, the disrupted Y-linked *CLDN34*, which was caused by SINEs inserted in the Y-linked *CLDN34*, might be favourable for male fitness, and two SINEs were fully linked with the Y chromosome due to the sexually antagonistic selection. The loss-of-function *CLDN34* had a loose selection on the gene body and its sequence diverged quickly from its X-linked gametologs. Also, accumulated SINEs on the Y chromosomes might cease recombination around the *CLDN34* loci, leading to less chance for gene conversion between sex chromosomes. All these processes promoted the generation of PAB as a consequence of distinct sequences between X and Y chromosomes.

Taken together, this chapter exclusively depicted the sequence of the dog PAB and investigated its origin and evolution across Canidae. The work has clarified the relationships of *CLDN34* and *TETY2* which were proven not to be homologs in mammals. *CLDN34* remains actively transcribed on the X chromosome where it is spanned by the PAB, whereas its Y-linked gametolog is pseudogenised. *CLDN34* is the first gene to be identified as being expressed either ubiquitously or just in the testes, in an isoform-dependent manner. *CLDN34* is in a state of transition from a pseudoautosomal gene to an X-unique gene providing evidence for a process of attrition of the pseudoautosomal region on the Y chromosome. Finally, accumulation of SINEs around the PAB is predicted to be the driving force for the reduction of recombination between sex chromosomes for dogs.

CHAPTER 6: NOVEL EVOLUTION OF MAMMALS' SEX CHROMOSOME GENE *PRSSLY*

6.1 Introduction

Mammalian sex chromosomes first arose from a pair of autosomes, following acquisition of a sex-determining gene by one of the pair. Then the sex chromosomes evolved in different trajectories and the recombination between sex chromosomes was suppressed progressively along the chromosomes in a stepwise manner (46,342). Reduced selection efficiency of the hemizygous Y-linked genes and their accumulation of deleterious mutations led to most of the genes being lost on the Y chromosome. Only a few genes were selected for sexual antagonism and survived. These became crucial for testis development and other functions (32). The rate of evolution of the X chromosome was slower than that of the Y chromosome and maintained most of its ancestral genes. To rescue the dosage for one copy of X chromosome genes in males, X chromosome genes acquired dosage compensation to balance the abundance of expression between males and females for somatic tissues (343).

Contrary to the phenomena described above, the *PRSSLY* (protease, serine-like Y) gene was the first example of a gene in mammals that existed on the ancestral autosomes, was retained on the Y chromosome, but was eliminated on the X chromosome. *PRSSLY* was first identified on the dog Y chromosome, where it was called *DYNG* (55), and was found on the mouse Y chromosomes in the same year (4), followed by the homologs annotated in pigs (69) and cattle (64). Hughes et al. (85) discovered that the *PRSSLY* was widespread across mammals' Y chromosomes, but not primates' and felines'. *PRSSLY* homologs on the X chromosome of marsupials are syntenic with an autosomal region in the monotreme and even more distant species such as lizards. Phylogenetic evidence indicated *PRSSLY* originated once in mammals. The expression pattern of *Prssly* in mice, as explored by bulk RNA-Seq and single-cell RNA-Seq suggests that this gene functions somehow in spermatogenesis, however whether the gene encoded a protein product or

not was not formally tested (85). Moreover, *Prssly* knockout mice are fertile (74,312), though one group suggested that *Prssly* may play a role in influencing sex ratios (85).

Of note, the experiments in this chapter were done before a very similar body of work was published by Hughes et al. (85). Our conclusions are discussed in detail below in **Section 6.4**.

6.2 Materials and Methods

6.2.1 *PRSSLY* Annotation in Mammals

6.2.1.1 Iso-Seq Data

Iso-Seq data of testis for mammals were downloaded from Sequence Read Archive (SRA) (**Supplementary Table 6.1**). DIAMOND (345) was used to perform protein alignments on Iso-Seq long reads with the sensitive model using *PRSSLY* sequences from dogs and mice as query sequences. Identified reads were then mapped against reference genomes, which are assembled from male samples or contain the Y chromosome. To generate annotations for *PRSSLY*, the alignment BAM files were initially converted into the BED format using bedtools (248). Subsequently, the BED files were further transformed into GTF files with UCSC utilities of "bedToGenePred" and "genePredToGtf" (346).

6.2.1.2 RNA-Seq Annotation Strategy

Due to the limited availability of Iso-Seq data for testis in mammalian species, an RNA-Seq method is widely applied to species that have access to both testis-specific RNA-Seq data and Y chromosome sequences.

First, dog *PRSSLY* protein sequences, inferred from Iso-Seq, were aligned to the potential Y chromosome in a splice-aware manner using the Spaln program (260).

Second, testis RNA-Seq for studied species was collected from SRA and aligned on the genomic sequences with HiSat2 (239) to define the boundary of exons for *PRSSLY*.

Sometimes Y chromosome sequences are assembled with errors in certain species, which can result in *PRSSLY* not being fully or correctly annotated. Hence, an additional step was developed to annotate *PRSSLY* correctly for the misassembled Y chromosomes (**Figure 6.1**). When RNA-Seq data were mapped against erroneous sequences, the alignment performed as truncated in the coding *PRSSLY*, and had long-range split junctions between coding and pseudogenised loci. The erroneous assembly was polished by RNA-Seq reads. This process is finished by using RNA-Seq reads to only align on the coding *PRSSLY*'s genome region. Then the exon sequences were corrected according to the alignment.

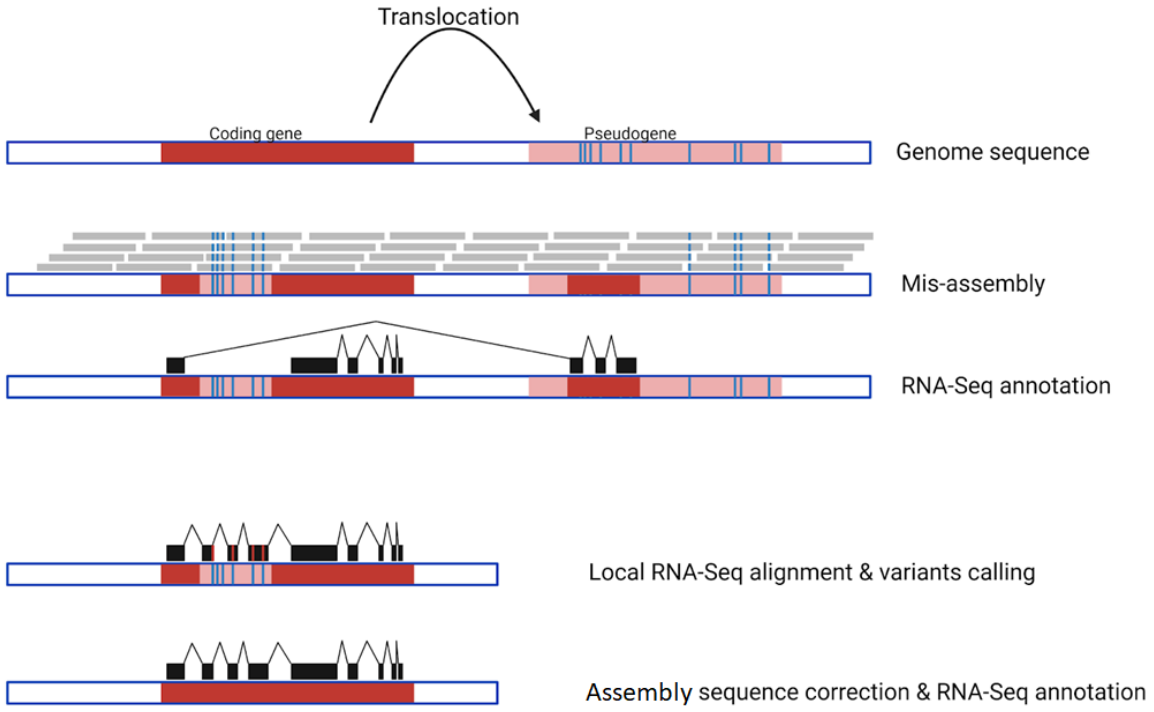


Figure 6.1. The strategy of local alignment method. Suspecting that the coding region is misassembled as a chimera of coding *PRSSLY* and pseudogenised *PRSSLY* sequences, RNA-Seq data is mapped on the potential coding region and the sequences are corrected according to the RNA-Seq alignment. Red and pink blocks present coding and pseudogenised genes respectively; grey lines refer to WGS alignment. The sequence differences between the coding gene and pseudogene are indicated by vertical blue lines, and vertical red lines mean the mismatches between RNA-Seq reads and reference sequences.

6.2.2 RNA-Seq Analysis in Mice

6.2.2.1 Bulk RNA-Seq Analysis

To investigate the expression pattern of *Prssly* in mice, bulk RNA-Seq samples with different developmental stages of testis were downloaded from SRA (**Supplementary Table 6.2**).

The mouse reference genome (GRCm39)'s Y chromosome was incomplete, where the *Prssly* locus was lost. In this study, the mouse reference genome was modified by adding the Y chromosome contigs of NW_001034423.1, which included *Prssly*'s genomic region. RNA-Seq alignment was applied by STARv2.7.8 (302) with default settings, and reads in each gene were counted with featureCounts (255) in Subread (http://subread.sourceforge.net) tools (255). Eventually, to quantify the gene abundance, reads counts were converted to transcripts per million (TPM) using a custom script.

6.2.2.2 Single Cell RNA-Seq Analysis

Single-cell RNA-Seq data of mice testis were obtained from Jung *et. al.* paper (347). The count matrix was downloaded from ZENODO (DOI: 10.5281/zenodo.3233958). Short reads were aligned on the modified mouse genome and the reads mapped on the *Prssly* locus were counted and assigned to the count matrix based on the barcode information. The R package 'Seurat' took the count matrix as input, followed by quality control (QC) and analysis steps (348). Cells with over 2500 or less than 200 feature counts and had <5% mitochondrial counts were filtered out from downstream analyses. Next, a global-scaling normalisation and log transformation were implemented to obtain the QC matrix. Principal Component Analysis (PCA) was performed on the scale data based on the determined variable features, and the dimensionality of the dataset was decided according to the PCA result. Finally, cells were clustered using the KNN graph method and visualised using a t-distributed Stochastic Neighbor Embedding (t-SNE) plot. The cell clusters were assigned to different progressive stages of spermatogonia based on the expression of selected makers (*Cd14*, *Sox9*, *Id4*, *Zpbp*, *Ddx4*, *Prm3*, *Prm2*) from previous studies (347,349,350).

6.2.3 Reverse Transcription PCR (RT-PCR) of *PRSSLY*

PRSSLY was assumed to be expressed in testes and three testes were collected for validating its expression in dogs. The starting material was prepared from snap-frozen tissues with an amount of <100mg for each extraction and was processed with the same protocol from Emily *et. al.* (351). Details are described in 2.3.1.

The starting material for synthesising complementary DNA (cDNA) was 1μg total RNA. Following the manufacturer's protocol, SuperScript III First-Strand SuperMix kit (#18080-400, Invitrogen) was used to generate first-strand cDNA with random oligonucleotide primers. Total RNA (1μg), primers (1μl), annealing buffer (1μl), and RNase/Dnase-free water (to 8μl) were mixed, and the mixture was incubated in a thermal cycler at 65°C for 5 minutes, followed by cooling down on wet ice for 1 minute. Then the tube was mixed with 2X First-Strand Reaction Mix (10μl) and SuperScript™ III/RNaseOUT™ Enzyme Mix (2μl) and was incubated in a thermal cycler at 25°C, followed by 50 minutes at 50°C. To terminate the reactions, the temperature of the thermal cycler was set to 85°C and the reactions were incubated at this temperature for 5 minutes.

Based on the transcript sequence of *PRSSLY*, two paired primers were designed with Primer3 (v4.1.0), and oligos were synthesised by ThermoFisher Scientific (**Supplementary Table 6.3**). Q5® Hot Start High-Fidelity 2X Master Mix (#M0515, NEB) was used for PCR experiments with standard reaction components and thermocycling conditions (Table 2.1, Table 2.2). The annealing temperature was empirically determined by gradient PCR and both optimal temperatures for two paired primers were both 60°C.

6.2.4 Fluorescence *In Situ* Hybridization for *PRSSLY*

Protocols for RNA fluorescence in situ hybridization (FISH) detection followed Molecular Instruments' *in situ* HCR v3.0 protocol (352). The following steps were performed by Dr. Megan Davey and Lynn McTier (Roslin Institute). In brief, testes sections from adult male dogs were immersed in 4% paraformaldehyde (PFA) for 15 minutes at 4°C, and slides were immersed four times consecutively at room temperature

for 5 minutes with 50%, 70%, and twice 100% ethanol respectively. Then slides were washed with Phosphate-buffered saline (PBS) three times and rehydrated with 2X SSC (#15557044, Thermo Fisher Scientific, UK). After adding 200µl of HCR hybridization buffer (Molecular Instruments, USA) for 10 minutes at 37°C, gene probes were loaded at a final concentration of 2nM in HCR hybridization buffer overnight in the 37°C humidified chamber. To remove excess probes, slides were washed three times at room temperature in 5X SSCT. SSCT solution was 5X SSC with 0.1% Tween 20 (#H5152, Promega, UK). HCR hairpins were added to amplify the probe sets overnight at room temperature. The hairpin solution was prepared by adding 6 pmol hairpins h1 and 6 pmol hairpins h2 to 100µl of amplification buffer. Prior to mixing the hairpin solution, hairpins h1 and h2 were snap-cooled, heated at 95°C for 90 s, and cooled down at room temperature for 30 min in the dark. After amplification, the slides were washed in 5X SSCT and stained with 100µl of Antifade Mountant (#11520686, Fisher Scientific, UK) before imaging.

6.2.5 Proteomics Analysis on Dog Testis

Liquid Chromatography-Mass Spectrometry (LC-MS) was used to validate the translation of the *PRSSLY*. Liquid chromatography (LC) is a separation process used to isolate the individual peptides of a mixture. The mixture is then separated by liquid chromatography, which uses a stationary phase and a mobile phase to separate compounds based on their interactions with these phases (353). The separated compounds are then introduced into a mass spectrometer, where they are ionized and analyzed based on their mass-to-charge ratios, providing valuable information about their molecular structures and compositions (353) (**Figure 6.2**). The following method (section **6.2.5.1** to **6.2.5.7**) has been performed by Dr. Dominic Kurian and Judit Aguilar (Proteomic facilities, the Roslin Institute).

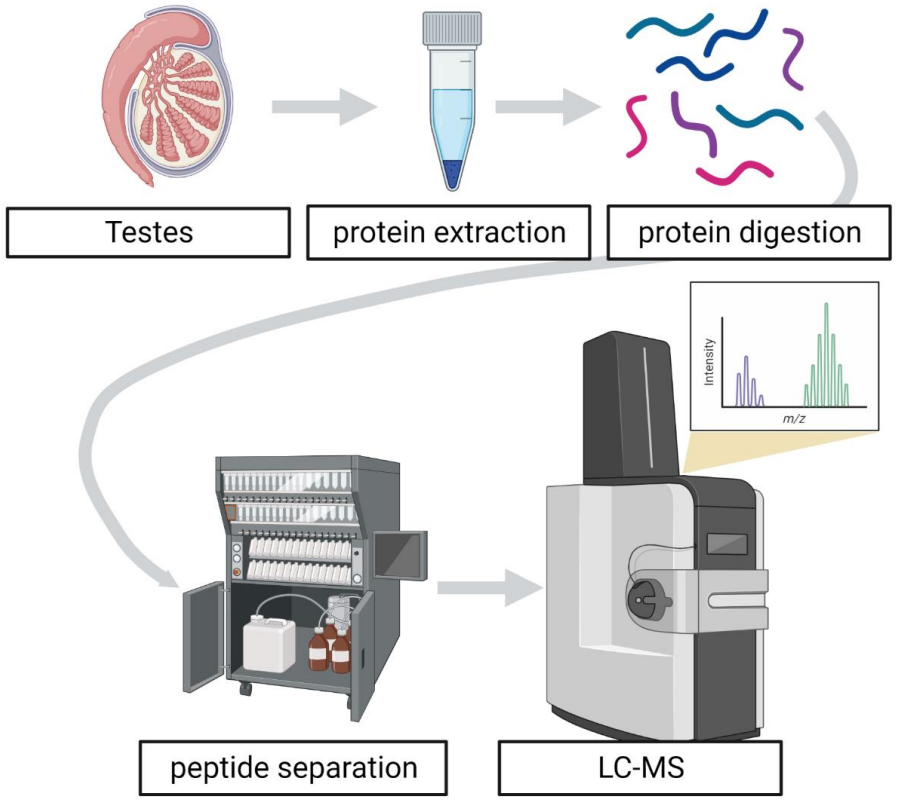


Figure 6.2. The workflow of LC-MS enables the detection of loads of novel peptides.

6.2.5.1 Protein Extraction

An aliquot of dog testes tissue was taken forward for proteomics preparation. Lysis buffer (5% Sodium dodecyl sulfate (SDS) in TEAB buffer, pH 7.5) was added to the samples. Then testes were homogenised with Precellys ceramic beads (CKmix), centrifuged at 13,000 x g for 15 minutes, before the supernatant was transferred to Protein LoBind® Tubes (Eppendorf), sonicated for 5 minutes (30 seconds on/off) in a Bioruptor® Pico sonication (Diagenode) and centrifuged once more at 10,000 x g for 10 minutes.

To remove the fat, a clean supernatant was carefully pipetted into a new vial. After that, a bicinchoninic acid assay (Pierce™ BCA Protein Assay Kit, Thermo Scientific) was used to determine the protein concentration of each sample. 200µg for each sample was taken forward for preparation, and the volume was made up to 50µL using 5% SDS in 50mM TEAB.

6.2.5.2 Reduction and Alkylation

In the reduction step, samples were sonicated for 15 minutes and incubated with 10mM Dithiothreitol (DTT) at 37°C for 1 hour at 300 rpm/min. Following reduction, samples were incubated with 18.75mM iodoacetamide at room temperature for 35 minutes. Then 12% phosphoric acid was added to lower the pH of proteins for trapping in the S-trap.

6.2.5.3 Protein Trapping

Binding buffer (100mM TEAB in 90% Methanol) was added to samples (6x volume), then transferred into S-trap mini spin columns (PROTIFI). Samples were centrifuged at 4,000 x g for 0.5 minutes.

6.2.5.4 Digestion and Peptides Elution

Trypsin was added to the samples at 1:10 (trypsin: protein) and incubated at 47°C for 2 hours. Prepared buffer (50mM TEAB, 0.1% Formic Acid in water, and 0.1% Formic Acid in 30% Acetonitrile) was added to the samples and centrifuged at 4,000 x g for 1 minute.

6.2.5.5 Liquid Chromatography Fractionation

The sample was then reconstituted and fractionated using LC Dionaex in a 96 deepwell plate, fractions with proteins were then combined to obtain a final 12 fractions for each sample. Final fractions were then dried once more using SpeedVac until incipient dryness.

6.2.5.6 Liquid Chromatography-Mass Spectrometry Analysis

The samples were reconstituted to 20µL of 0.4%FA in Acetonitrile/water (4:96, v/v), then further diluted (2µL taken from each sample) using the same solvent to obtain 0.1µg/µL in sample and 20µL placed into an HPLC vial. The purified peptide samples of 2.5µL were injected and analysed by TimsTOF Flex (Bruker, Germany) coupled with a Dionex nanoLC system (Thermo Scientific).

6.2.5.7 Data Analysis

Raw spectral data were processed by DataAnalysis (Bruker) software and the resulting peak lists were searched using Mascot 2.4 server (Matrix Science) against the modified Uniprot Dog (*Canis lupus familiaris*) sequence database containing 59,102 entries (UP000002254) and the PRSSLY peptide sequence. Trypsin was set as the proteolytic enzyme and up to one missed cleavage was allowed. Mass tolerance on peptide precursor ions was fixed at 25 ppm and on fragment ions at 0.08 Da. Carbamidomethylation of cysteine was selected as a fixed modification and oxidation of methionine and de-amidation were chosen as variable modifications. The false discovery rate was limited to <1% for peptide IDs after searching decoy databases.

6.2.6 Proteomics Data Analysis on Mice

To validate the translational potential of the *Prssly* gene in mice, a total of 65 runs from mouse proteomics data were downloaded including 31 testis samples and 34 samples from variable tissues (354). The raw data were searched against a modified protein database using MaxQuant v2.1.4.0 (355). The modified protein database included the Uniprot Mouse (UP000000589, 55,286 entries) and the Prssly peptide sequence. Oxidized methionine and acetylation (protein N terminal) were selected as variable modifications, carbamidomethyl as fixed modification, trypsin was selected as the proteolytic enzyme, and up to two missed cleavages were allowed with a minimum peptide length of seven amino acids.

6.2.7 Evolution Analyses of PRSSLY in Mammals

To construct the phylogenetic tree for the PRSSLY protein sequence for other mammals, a multiple sequence alignment was performed using the ClustalW algorithm within the MEGA7 software (243). The multiple alignments were visualised with Geneious Prime. Subsequently, the phylogenetic tree was reconstructed utilising the Jones-Taylor-Thornton model. The resulting Newick file, containing the tree structure,

was then uploaded to the iTOL (https://itol.embl.de/itol.cgi) to facilitate visualisation and exploration.

The prediction of protein domains was accomplished using InterPro (359) which aided in identifying specific functional regions within the protein sequences. Furthermore, to assess the conservation levels of these sequences, the AL2CO algorithm (262) was employed, resulting in a calculated conservation index. Visual representations, including line plots and box plots, were generated using the "ggplot2" package within the R programming.

6.3 Results

6.3.1 Transcription and Translation of *PRSSLY* in Dog

6.3.1.1 Annotation and Nomenclature of *PRSSLY* by RNA Sequencing

On the scaffold chrY2 of RosY_1.0, both RNA-Seq and Iso-Seq annotated a novel coding gene never reported in dogs (**Figure 6.3**). This gene is partially annotated by RefSeq (*LOC119868781*) without an official symbol indicating its orthologs have not yet been determined. By estimating its open reading frame (ORF) with ORFfinder (https://www.ncbi.nlm.nih.gov/orffinder/) and detecting orthologs with SmartBLAST (https://blast.ncbi.nlm.nih.gov/smartblast/), it displayed similarity with serine protease 55 (*PRSS55*). A previous study of mouse Y chromosome identified a serine-like protease, which the authors called "*PRSSLY*" (4). The molecular phylogeny tree based on protein sequence showed the detected gene and *PRSSLY* of mice were paralogous, while *PRSS55* was shared direct orthology with *PRSSLY* (**Figure 6.4**).

Using dog Iso-Seq data, two isoforms of *PRSSLY* were annotated. These differed at exon 6 where one of the isoforms included 28 amino acids (**Figure 6.5A**). RNA-Seq, which has much higher depth than Iso-Seq, supported the expression of two isoforms. Reads specific to the 28 amino acids were used to predict the relative abundance of isoform2-pb which was on average consisted of 15.5% of the total expression of *PRSSLY* (**Figure 6.5B**).

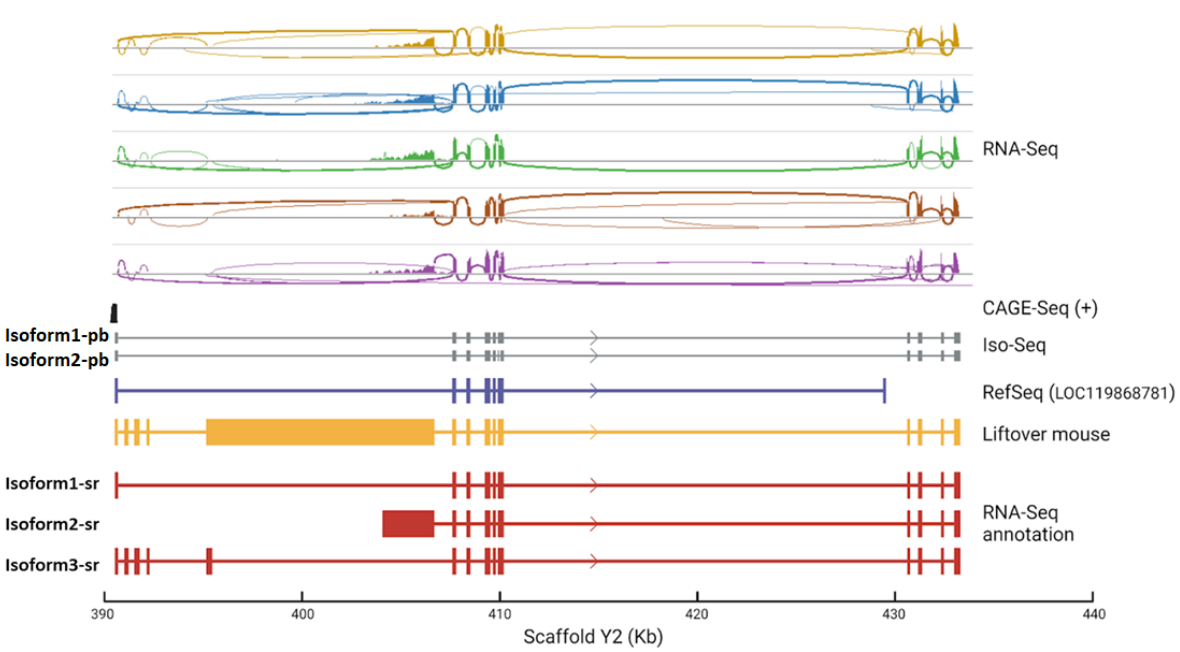


Figure 6.3. Annotation of dog *PRSSLY* by RNA-Seq and Iso-Seq.

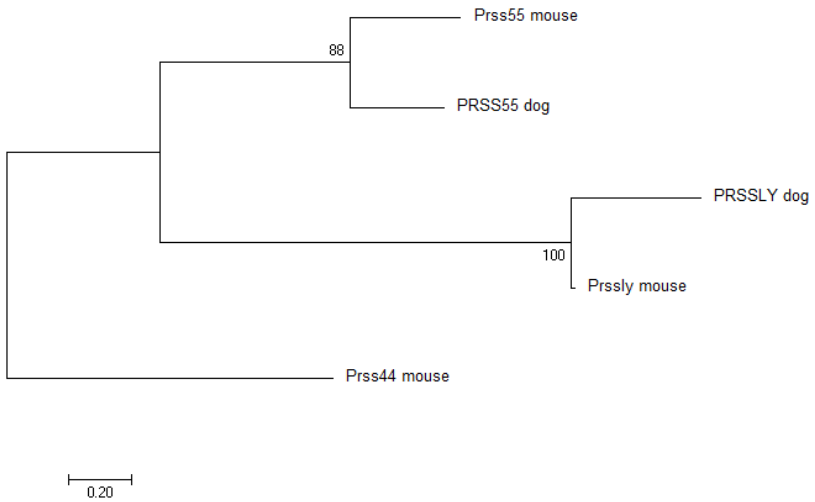


Figure 6.4. Phylogeny construction of *PRSSLY* and *PRSS55*. Molecular phylogenetic analysis of *PRSSLY* and *PRSS55* was conducted using the maximum likelihood method based on protein sequences. Mouse Prss44 is used as the root.

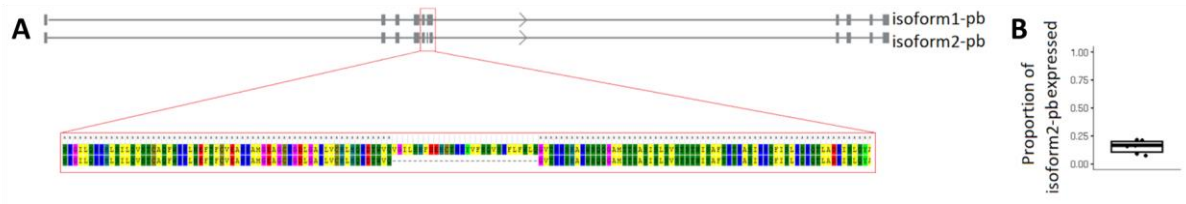


Figure 6.5. Identification and quantification of the *PRSSLY* isoforms. (A) Two isoforms of *PRSSLY* are identified in dog Iso-Seq data and their difference in amino acid. (B) The proportion of the expression of the isoform2-pb was estimated using RNA-Seq by calculating splice junctions.

Ignoring the 28 amino acids difference, RNA-Seq annotated two other isoforms that were not detected by Iso-Seq (**Figure 6.3**). The abundance of *PRSSLY*'s isoforms in 6 testis samples was estimated by read depth and splice junction methods separately (**Figure 6.6**). Isoform1-sr and Isoform2-sr were expressed 2.9 and 4.3 times higher than Isoform3-sr based on the read depth method, and their expression was higher than that of Isoform3-sr with 7.3 and 12.1 times estimated by the splice junction method. Though the methods differed, their overall interpretation was the same: isoform2-sr, and to a lesser extent isoform1-sr, are expressed in the testis.

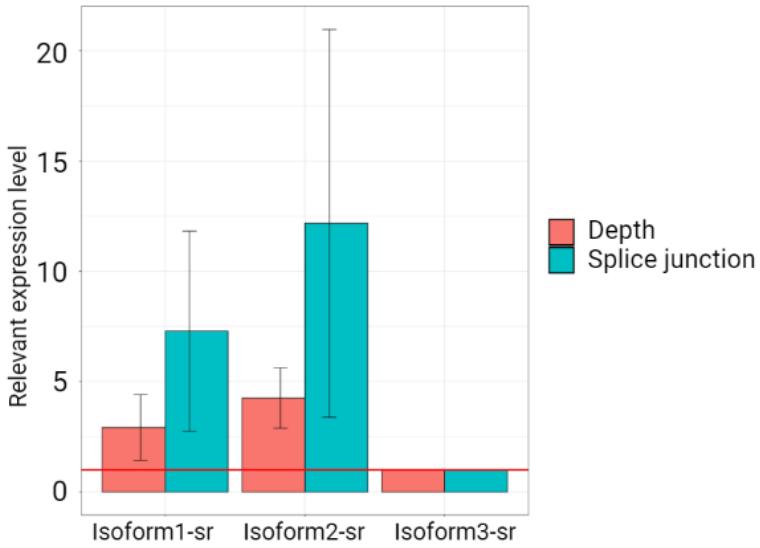


Figure 6.6. Quantification of the *PRSSLY* isoforms. The abundance of isoforms using short reads was estimated by the read depth and splice junction methods. Isoform3-sr was adjusted by 1 in each sample, which can directly reflect the relevant expression of the other two isoforms.

6.3.1.2 RNA FISH of *PRSSLY* Expression in Dog Testes

To explore spatial gene expression of *PRSSLY* in testes tissue, multiplexed HCR RNA FISH was applied using specific probes to it as well as *HOOK1* and *ZPBP*. In humans, *HOOK1* was expressed from the early stage to the late stage of spermatids (332,356), and *ZPBP* was mainly expressed in the spermatocytes and early spermatids (350,357,358). By imaging the immunolocalization of the tested genes and nuclei, *PRSSLY* was seen to co-localize with *ZPBP*, suggesting it was mainly expressed during the transition from spermatocytes to round spermatids in dog testes (**Figure 6.7**).

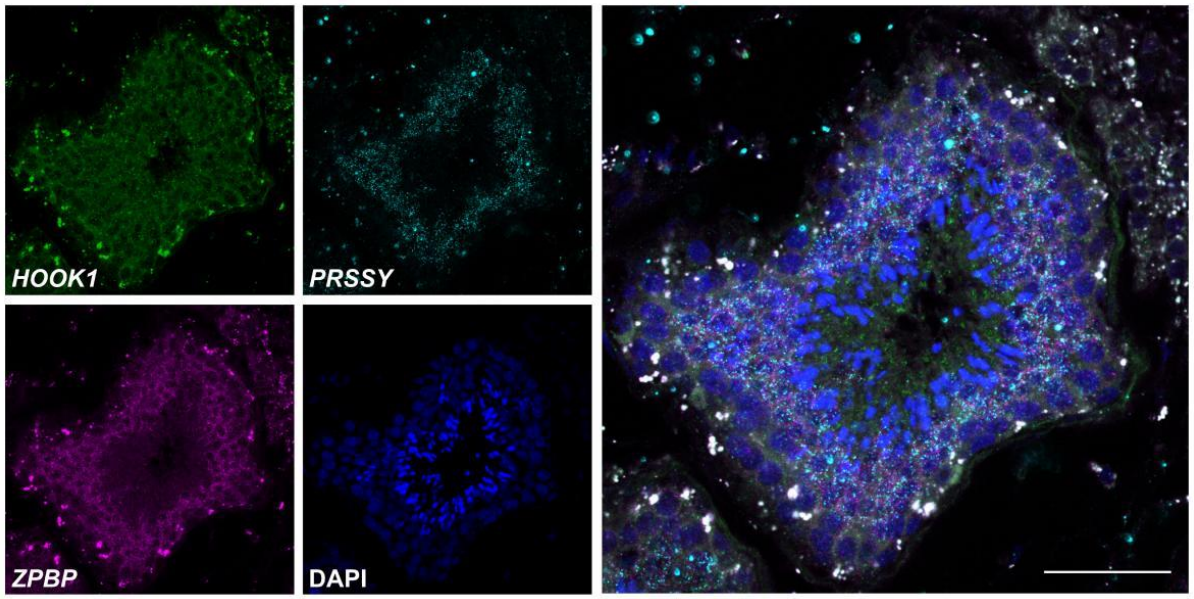


Figure 6.7. RNA FISH on testicular tubular sections for dog testes. The individual channels are as follows: *HOOK1* (green), *PRSSLY* (cyan), and *ZPBP* (pink), and nucleolus is stained with DAPI (blue). Right image is the merged view of the four channels.

6.3.1.3 *PRSSLY* Encodes a Protein

Not all genes with intact ORFs or that express mRNAs will lead to a translated protein. It was assumed that *PRSSLY* encodes a protein due to the prediction of an open reading frame. Proteomics was used to test this hypothesis. First, three candidate samples selected for proteomics were validated by RT-PCR. Two paired primers were designed on the shared exons sequences among isoforms. Two of three dog samples showed expression (**Figure 6.8**) and were used for the following proteomics analysis.

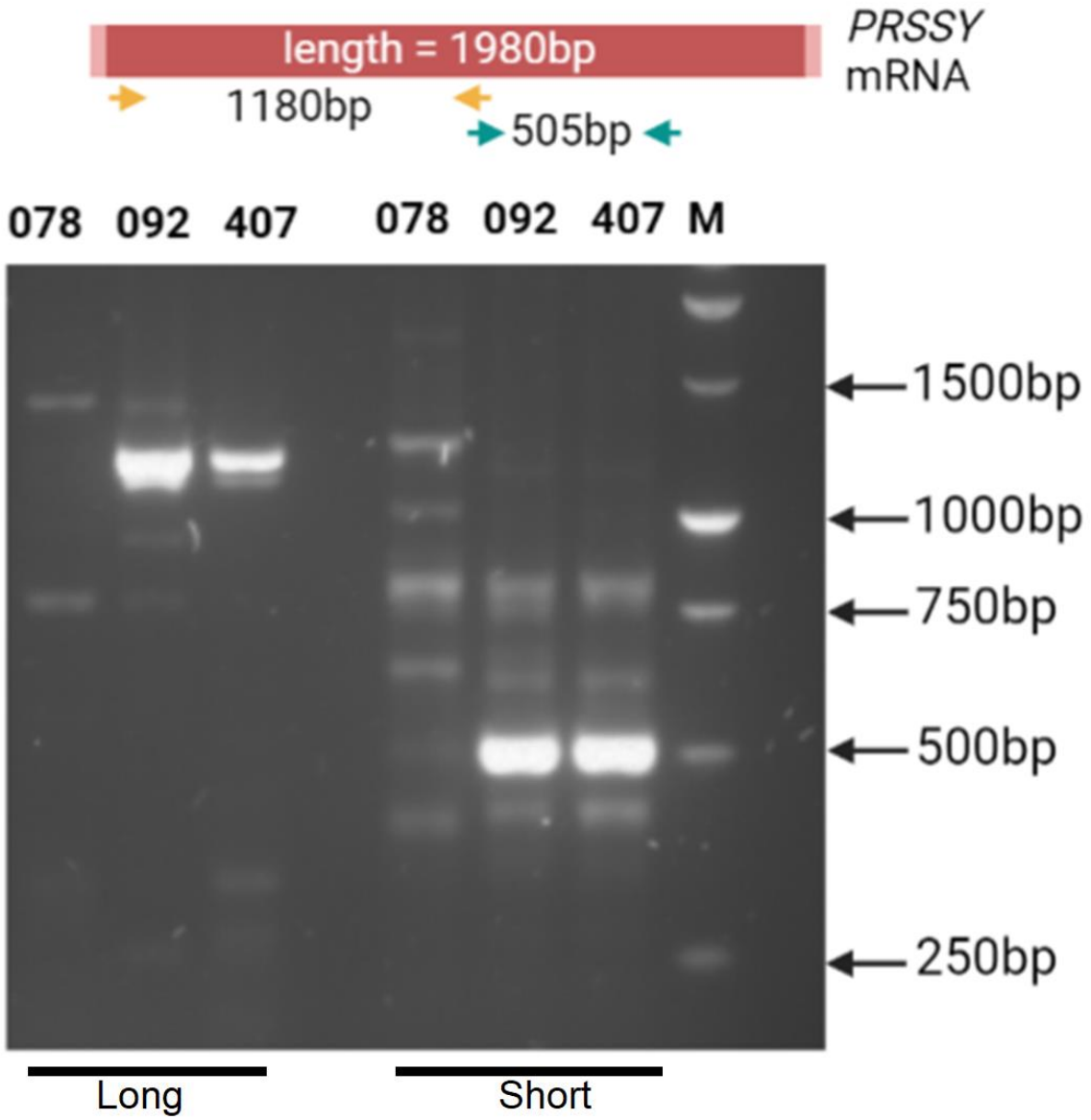


Figure 6.8. Validation of the expression of *PRSSLY* by RT-PCR. The labels "long" and "short" at the bottom present 1180 bp PCR and 505 bp PCR experiments respectively. Three different dogs' testis were conducted with RT-PCR validation. The number at the top indicates samples ID and M is the marker.

Proteomics analysis was implemented for two validated dog testes, as a result, a total of 12 peptide sequences specific to the *PRSSLY* were identified when compared to the Uniprot Dog database (**Figure 6.9**).

Figure 6.9. Validation of dog PRSSLY by proteomic data. Proteomics analysis identified unique peptides for PRSSLY in two dog testis samples.

6.3.2 Correction of the Mouse *Prssly* Gene Model

The mouse *Prssly* gene was not readily identified in the reference genome (GRCm39) but was found in an unplaced scaffold (NW_001034423.1) belonging to the Y chromosome (**Figure 6.10**). The genomic region, which showed homology with *PRSSLY* of the dog and possum, was annotated with four coding genes in the RefSeq annotation. Intuitively, RNA-Seq alignment revealed that RefSeq's four gene models correspond to one intact gene, which was thus presumably mis-split by the Gnomon pipeline. Supporting this interpretation, mouse CAGE-Seq data revealed just one TSS located at the 5' end of the RefSeq gene model *LOC102641968*.

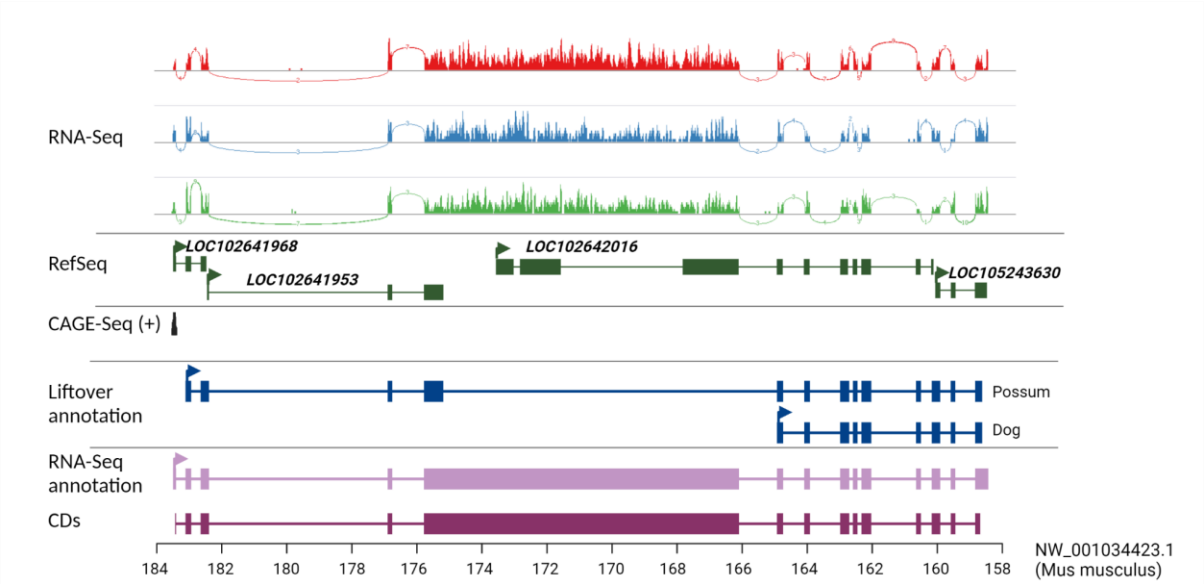


Figure 6.10. Annotation of *Prssly* by RNA-Seq in mice's Y chromosomes. RNA-Seq, CAGE-Seq, and liftover annotation by Spaln from dog and possum are shown. RefSeq truncated the *Prssly* gene due to the misassembly of the mouse's reference sequence.

In the primary step of genome assembly, coding sequences can be obfuscated and collapsed with other pseudogenic sequences. Misassembly can potentially cause ORF frameshifts and introduce stop codons. As *Prssly* was reported as multiple copies in mice (344), it is possible that the misassembly of the genomic sequences led to a truncated annotation by the Gnomon pipeline. To look at this, RNA-Seq were aligned on the modified genomic sequences of mice (GRCm39 + NW_001034423.1). The exon sequences of reference had a large amount of mismatches with the RNA-Seq reads expressed with SNVs and INDELs observed (**Figure 6.11**). For example, at the position of 80,228 bp on the scaffold NW_001034423.1, there was a 1 bp deletion in the RNA-Seq alignment and WGS data appeared heterozygous at this position and its surrounding sites (**Figure 6.12**). These results demonstrate that the *Prssly* locus exists as multiple copies in the sequenced samples and that the mouse GRCm39 was likely misassembled due to collapse of ampliconic sequences that included pseudogenised *Prssly* genes being confused with the real coding sequences. One of these variants, a deletion in the third to last exon (shown as the vertical red line in **Figure 6.12**) of the real *Prssly*, shifted the ORF, resulting in a truncated annotation in RefSeq (*LOC102642016* and *LOC105243630*).

To obtain the correct gene model, NW_001034423.1 sequences were polished by the RNA-Seq reads. Compared to the alignment on the crude genome, the realignment of RNA-Seq data on the modified sequence displayed an improvement with a few mismatches and more reads supporting splice junctions (compare **Figure 6.13** with **Figure 6.11**). The corrected annotation generated an ORF of 12,039 bp corresponding to a 4,013aa protein; this protein is orthologous to dog and possum PRSSLY protein. The predicted mouse Prssly peptide sequence was integrated with the Uniprot Mouse database, and proteomics data were exploited to search against the library for Prssly-specific peptides. As a result, a total of 31 peptides were identified in testis samples, and only one peptide was detected in non-testis samples (**Figure 6.14**)

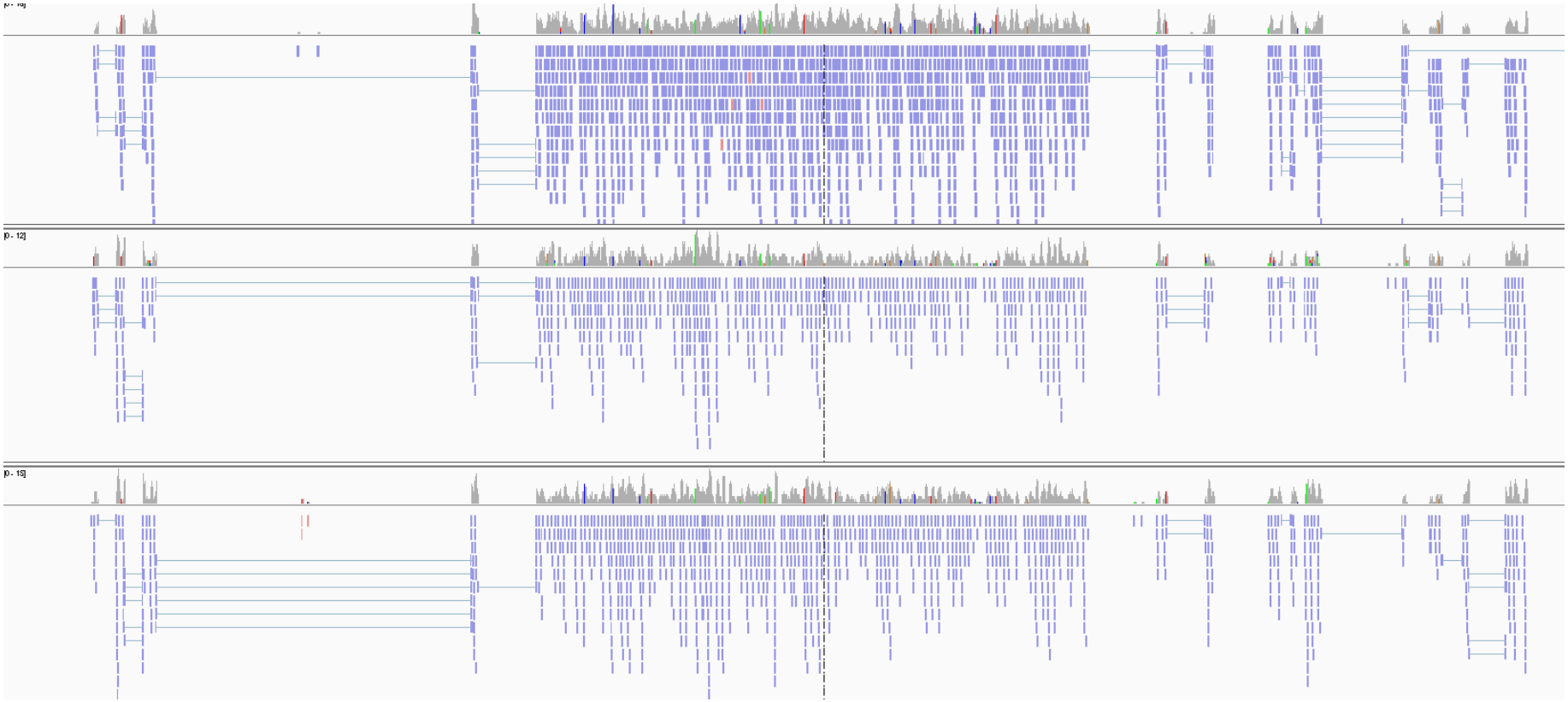


Figure 6.11. IGV screenshot at *Prssly* locus of mice. Vertical bars (green, red, brown or blue) reflect mismatches (SNVs, INDELs) between RNA-Seq reads and the mouse reference sequence.

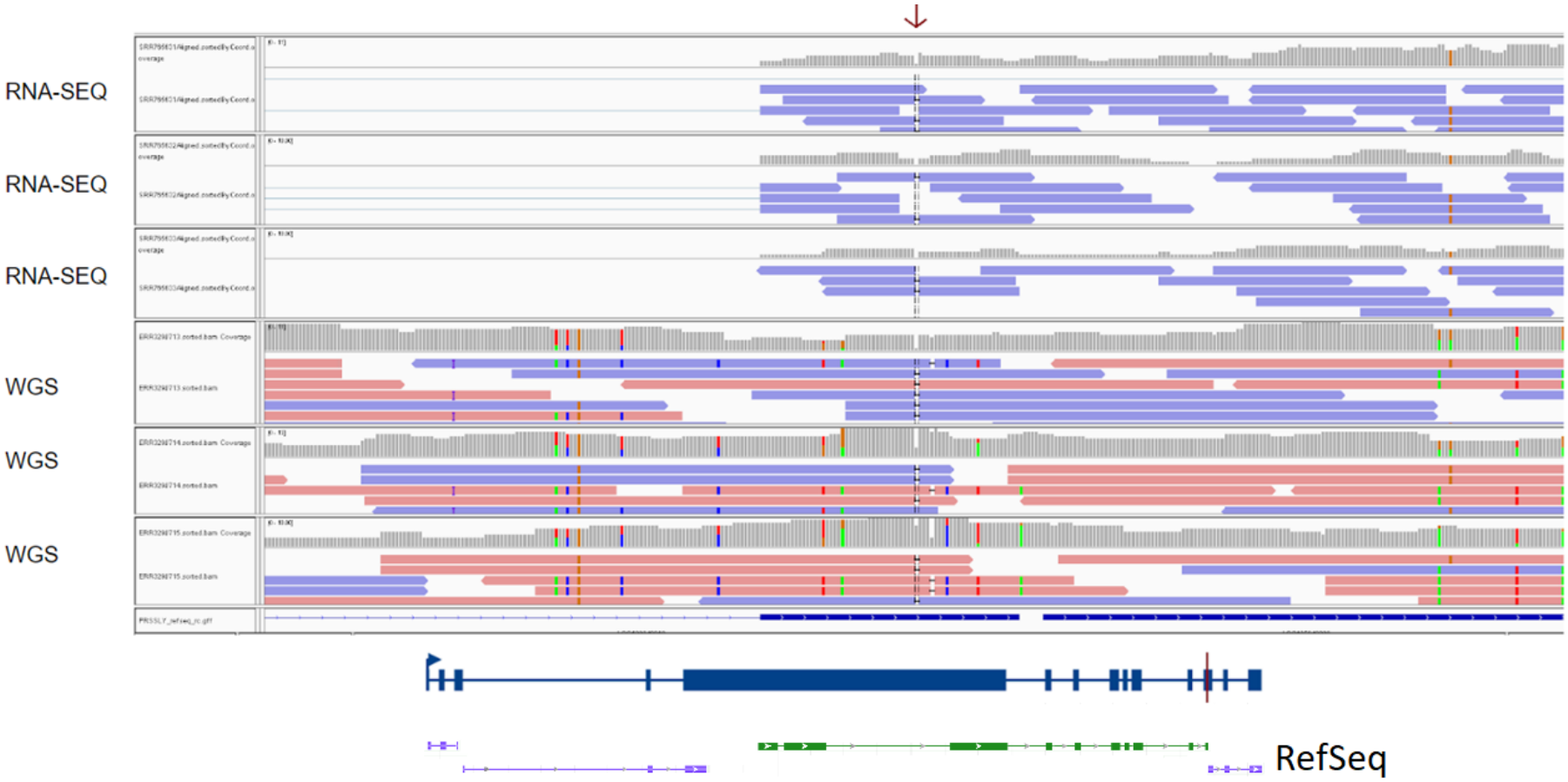


Figure 6.12. Erroneous *Prssly* annotation in misassembly. An example of the introduction of an INDEL leads to the reading frame shift. As a consequence, RefSeq annotation is truncated. The INDEL is pointed by the arrow in the IGV screenshot and by the vertical red line in the gene model below.

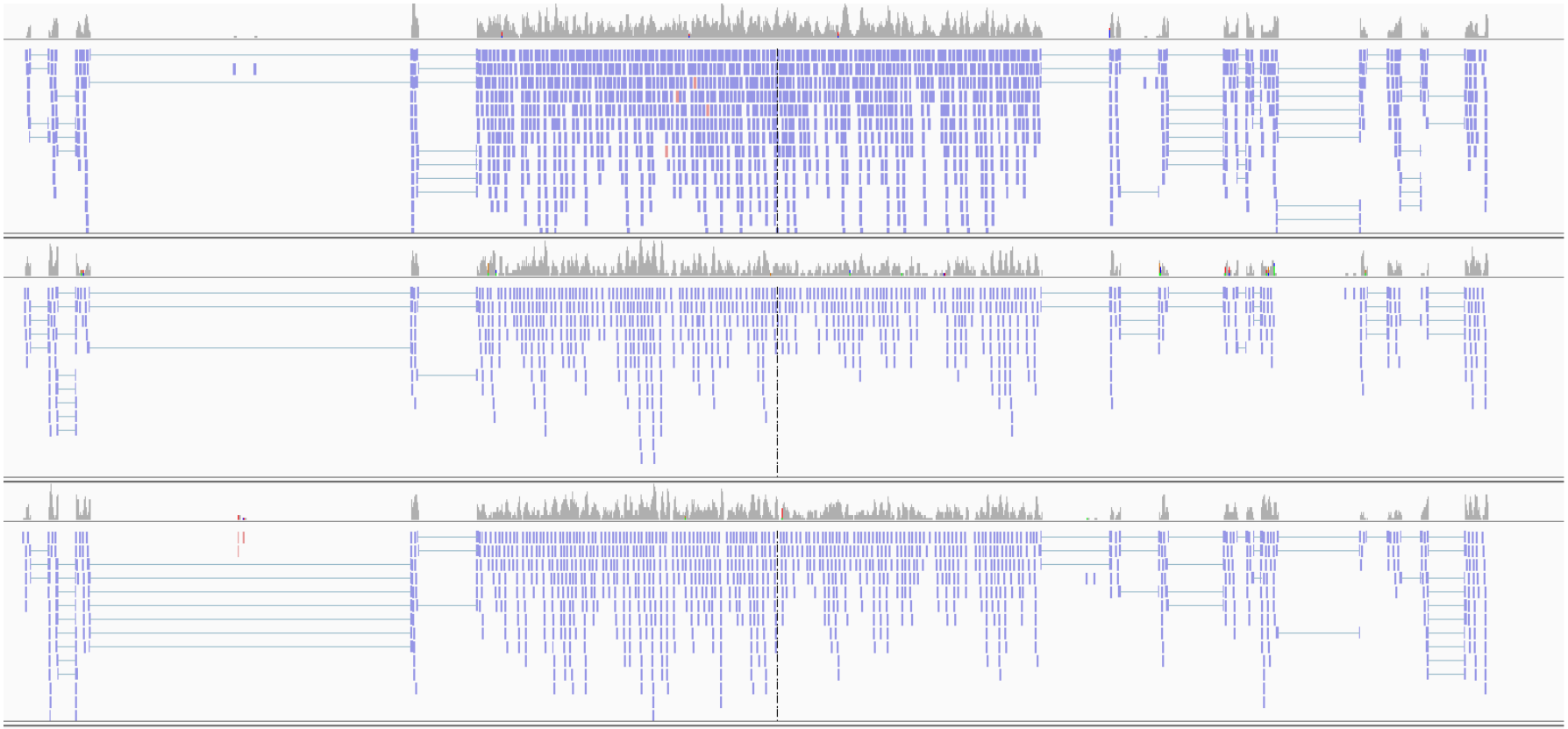


Figure 6.13. Improved alignment based on the corrected reference genome. RNA-Seq data are mapped on the polished NW_001034423.1. Compared with raw genomic sequences, an improved alignment is seen with few mismatches and more reads supporting splice junctions.

Figure 6.14. Validation of mouse Prssly by proteomic data. Proteomics analysis identifies unique peptides for Prssly protein in mice testes tissue (red). Unique peptides from other tissues are highlighted in yellow. Black arrows indicate exon boundaries. (The high resolution figure is available on https://github.com/WengangXbio/script_bio/blob/main/Figure%206.14.png)

6.3.3 Evolution of *PRSSLY*

6.3.3.1 *PRSSLY* Annotation in Mammals

The misassembly and misannotation of mouse *Prssly* raised the question whether this Y-specific gene is present in other mammals. To investigate, Iso-Seq data of testes

from mammals were downloaded and TBLASTN scanned using PRSSLY protein sequences as a query. *PRSSLY* transcripts were found in bats and common brushtail possum, but the bat's *PRSSLY* transcript appeared truncated, with 5' end exons missing compared with the RNA-Seq alignment (**Supplementary Figure 6.2**).

For the species without Iso-Seq data, *PRSSLY* was annotated by incorporation of protein sequence annotation and RNA-Seq alignment described in **6.2.1.2**. The possum *PRSSLY* protein sequence was aligned on the marsupial genome to annotate *PRSSLY*, and the *PRSSLY* sequences of dog and mouse were used in placental mammals' annotation. Species which had Y chromosome sequences (in the form of either scaffold or contigs) and RNA-Seq of the testis, were analysed, and a total of 15 mammals were annotated and presented the expression of the *PRSSLY* (**Supplementary Figure 6.2**).

For monotremes, TBLASTN searches on the genome assemblies of echidna (mTacAcu1) and platypus (mOrnAna1) was implemented using the *PRSSLY* protein sequences of dog and mouse as queries. As a result the uncharacterised genes, *LOC119930118* and *LOC100078881* were identified as paralogous genes to *PRSSLY* in echidna and platypus, respectively.

6.3.3.2 Origin of *PRSSLY* in Therian Evolution

In placental mammals, *PRSSLY* was present as Y-specific except for rats and mole rats where it translocated to autosomes (**Figure 6.15**). In rats, *PRSSLY* was located on one end of chromosome 16, distal of the short arm. The mole rats' *PRSSLY* was mapped to an unplaced scaffold that had WGS coverage in both male and female genome sequencing data. Finally, *PRSSLY* was annotated on the X chromosomes for marsupial and on chromosome 6 for monotremes (**Figure 6.16**).

It is can be inferred from its widespread representation, that *PRSSLY* was related to sex chromosome evolution in mammals. A synteny plot showed the physical co-localization of *PRSSLY* and its surrounding genes between chromosome 6 of monotremes and the X chromosome of marsupials. For the placental mammal genomes, a syntenic region can be found on X chromosome, however *PRSSLY* is missing from these regions (**Figure 6.16**).

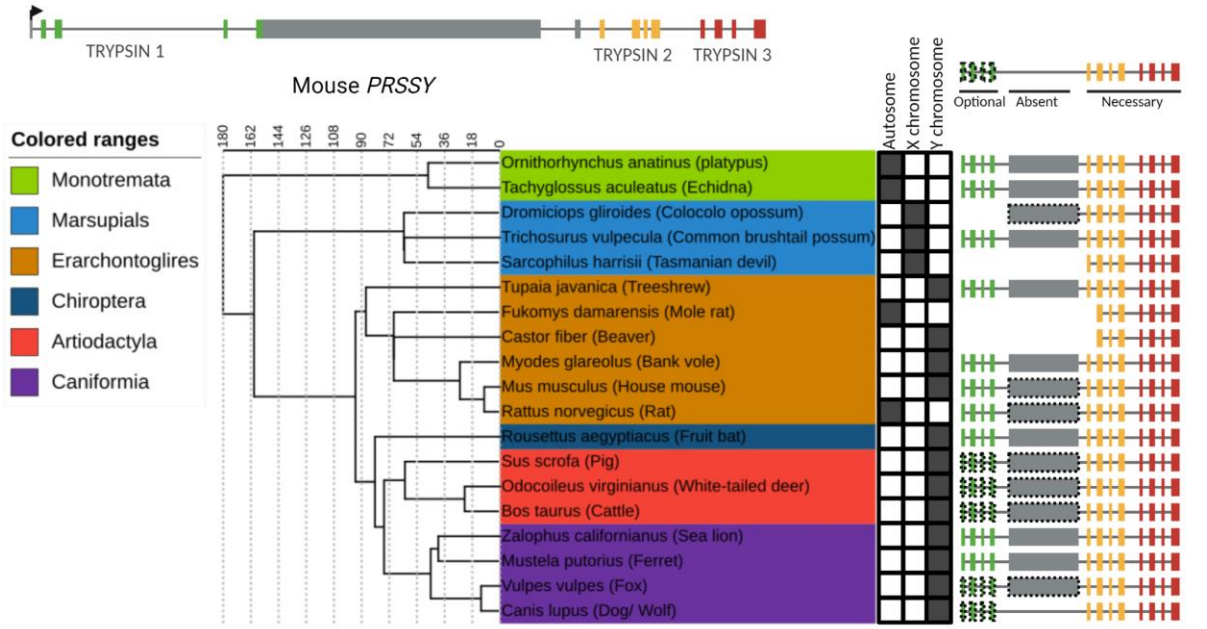


Figure 6.15. Evolution of the *PRSSLY* gene in mammals and species-specific predicted protein structures. Gene structure at the top is coloured to distinguish trypsin domains, and structure for each species is absent, optional or necessary shown on the right. The mammalian evolutionary tree is constructed based on Timetree of Life (http://www.timetree.org/).

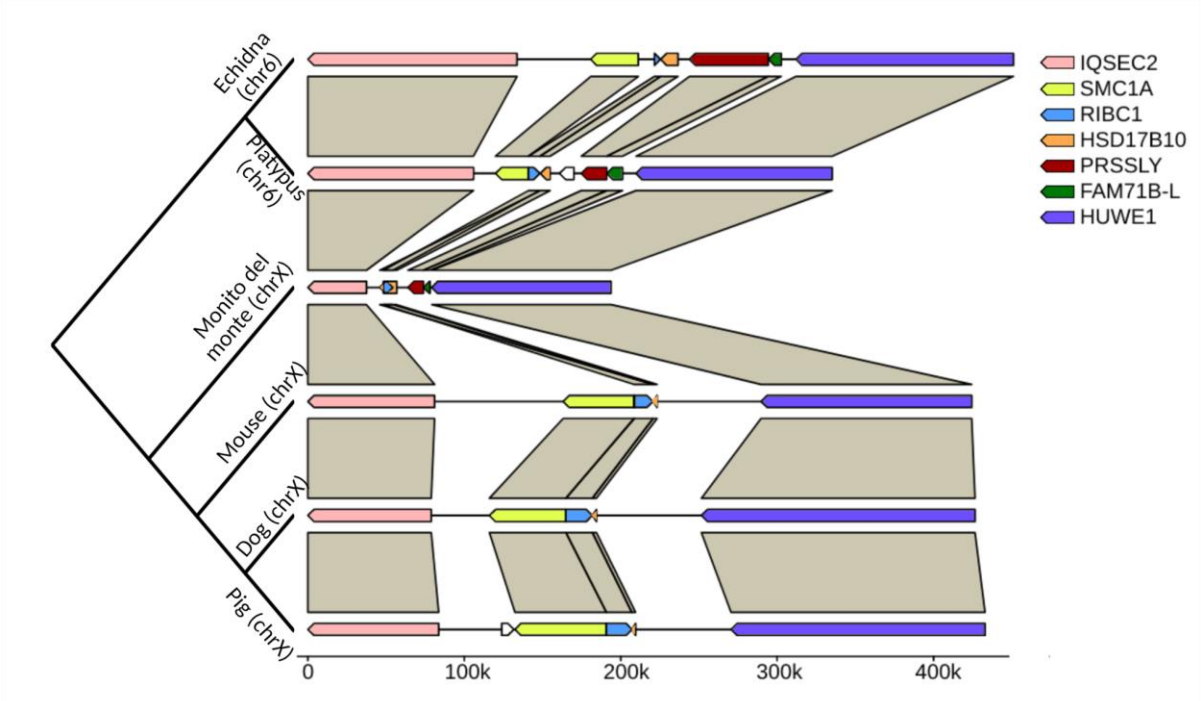


Figure 6.16. Genomic synteny around *PRSSLY* locus in mammals. *PRSSLY* and its neighbouring genes display collinear across monotremes, marsupials, and therians, but *PRSSLY* has lost from the X chromosome in Therians. This plot is made by the 'gggenomes' package of R.

PRSSLY was only retained on the Y chromosome in the placental mammal species. A phylogenetic tree based on the PRSSLY protein sequences was constructed using monotremes as a root (**Figure 6.17**). As a result, X-linked PRSSLY of *Marsupialia* and Y-linked PRSSLY of placentals generated sister groups, indicating they descended from a pair of ancestral *PRSSLY* genes from the common ancestor of the therians.

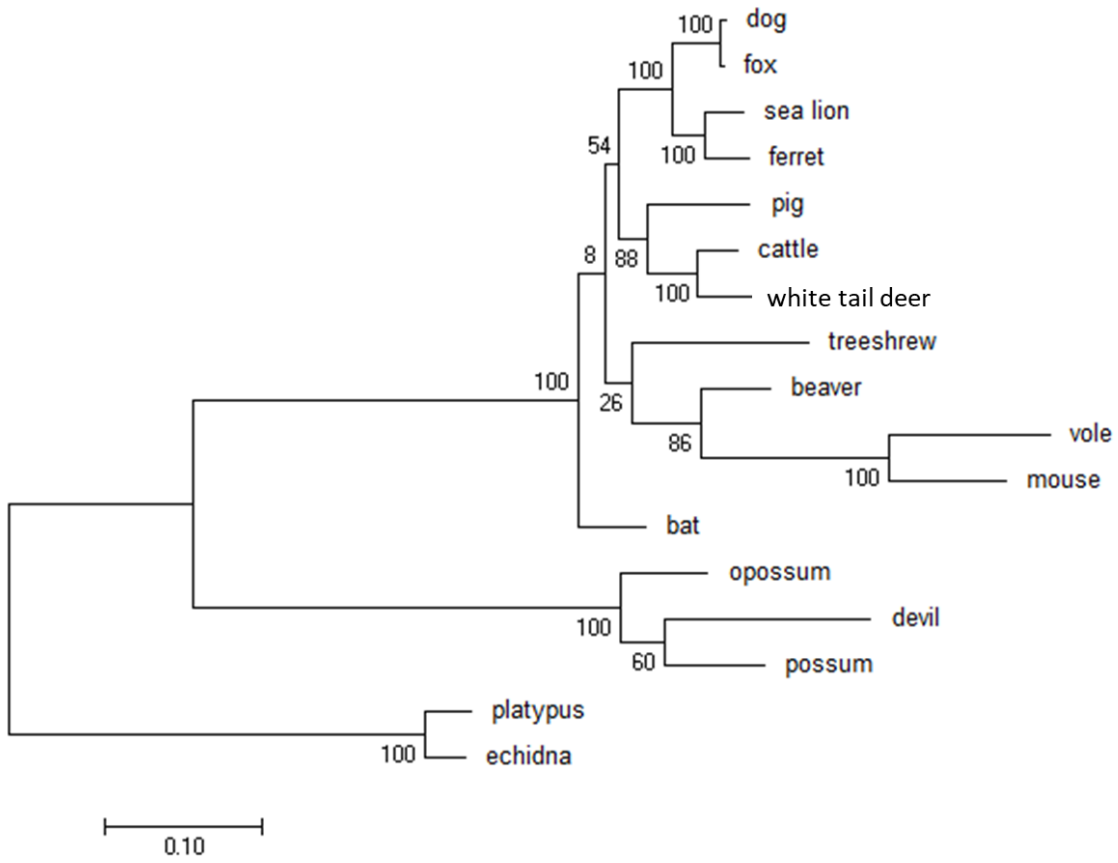


Figure 6.17. Phylogeny of PRSSLY protein across mammals. Phylogenetic tree of PRSSLY is constrctured based on protein sequences using the Maximum Likelihood method.

6.3.3.3 Characteristics and Conservation of PRSSLY Peptide Sequences

In the mouse, InterPro (359) predicted the protein domain of Prssly including three trypsin domains (**Figure 6.15**). The three domains are similar in size (224aa of trypsin1, 219aa of trypsin2, and 219aa of trypsin3) and each peptide sequence corresponds to four exons in *Prssly*. Between trypsin1 and trypsin2 domains, there was a ~3200aa

interval that had no discernible protein domain(s); this sequence was translated from an ultra-long exon (**Figure 6.15**). For clarity, we refer to this peptide interval as a "linker domain" in reference to its juxtaposition between trypsin domains, however no function is predicted.

Multiple alignment of PRSSLY protein was implemented across all peptide sequences annotated in the study. As a result, trypsin2 and trypsin3 were shown as essential domains of PRSSLY for all species analysed (**Figure 6.15**). On the contrary, the trypsin1 domain was absent in 4 of 19 species and 5 species' trypsin1 could be optionally translated from different expressed isoforms. The transcriptional sequence corresponding to the ~3200aa stretch was absent in 4 species but conserved in 7 species.

The linker domain is described as "approximate" because its length is highly variable, with a maximum of 4354aa in the fox and a minimum of 519aa in the platypus. To investigate the structure and conservation of this domain, a total of 15 PRSSLY sequences containing this linker domains were aligned using MEGA7 (243) and visualised with Geneious Prime 2021.1.1 (http://www.geneious.com/) (**Figure 6.18**). The monotreme and marsupial sequences had similar ~3200aa stretches between trypsins and were shorter than that of the therians. For the therians, their linker domains were shown to be much more complex in DNA sequence compared with the monotremes and marsupials, and most sequences were unique to monotremes and marsupials. It is suggested these novel desert sequences were generated after the Y-linked *PRSSLY* separated from ancestral *PRSSLY*. Although closely related species displayed a similarity in length and sequences, such as cattle and deer, ferret and sea lion, and mouse and rat, the linker domain was less conserved than trypsin domains indicating a quicker evolutionary rate in these loci. Furthermore, conservation index of three trypsin domains was calculated by AL2CO (262), indicating trypsin1 was lower than trypsin3 and trypsin2 (**Figure 6.19**).

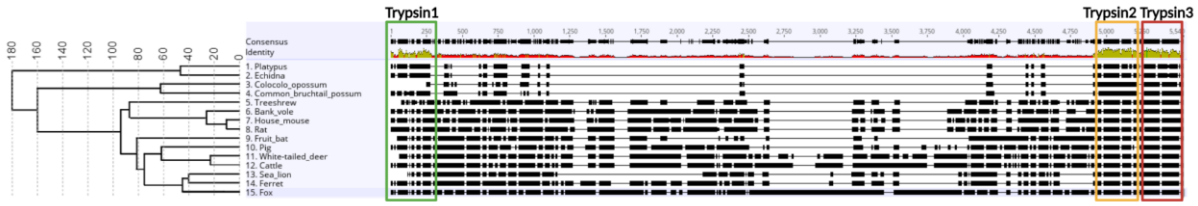


Figure 6.18. Alignments and visualisation of *PRSSLY*. Multiple alignments of *PRSSLY* DNA sequence are viewed with Geneious Prime.

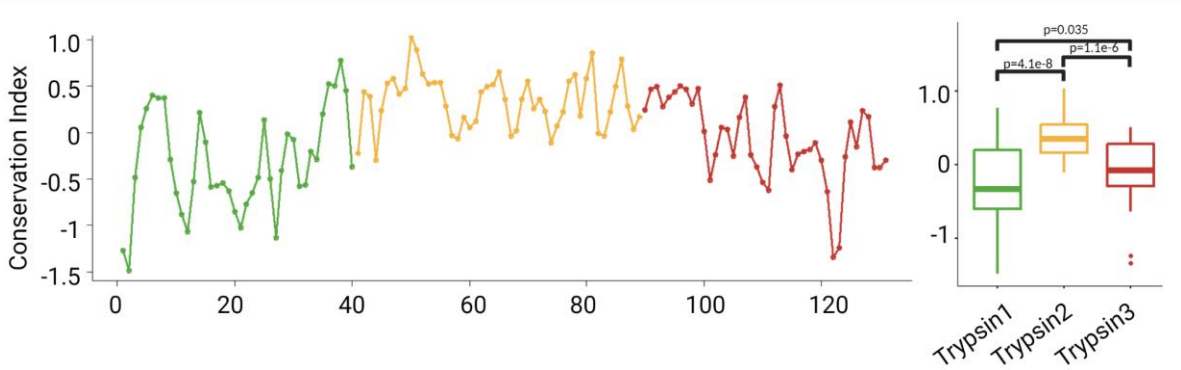


Figure 6.19. Conservation of three trypsin domains. Conservation Index across three trypsin domains is calculated using the AL2CO, and the overall conservation is compared with t-test statistics.

6.3.3.4 Pseudogenisation of *PRSSLY* in Primates

In chimpanzees, a *PRSSLY* homolog was identified on the Y chromosome by a TBLASTN search using the dog protein sequence as a query. The *PRSSLY* locus appeared transcribed as RNA-Seq reads from chimpanzee testes overlapped with the detected exons (**Figure 6.20**). However, stop codons within the potential exons were identified across all three reading frames, indicating that *PRSSLY* was pseudogenised in chimpanzees.

In humans, the *PRSSLY* and its flanking sequences were conserved with chimpanzees (**Figure 6.21**), however RNA-Seq from testes samples did not detect any evidence of expression. Also, stop codons were observed within the region suggesting the pseudogenisation of *PRSSLY* in humans. The primate order was investigated more broadly. Using dog and mouse *PRSSLY* sequences as a query, homologous sequences were detected in new-world monkeys' genomes (white-fronted capuchin and common marmoset) and old-world monkeys' genomes (rhesus macaque, snub-nosed monkey, and Francois' leaf monkey). Their pseudogenisation was predicted based on the observation of

stop codons. For Lemuriformes species, the *PRSSLY* gene of the ring-tailed lemur was believed to be pseudogenised: although there was no RNA-Seq data from testis, stop codons within a pseudo exon were observed (**Figure 6.22**).

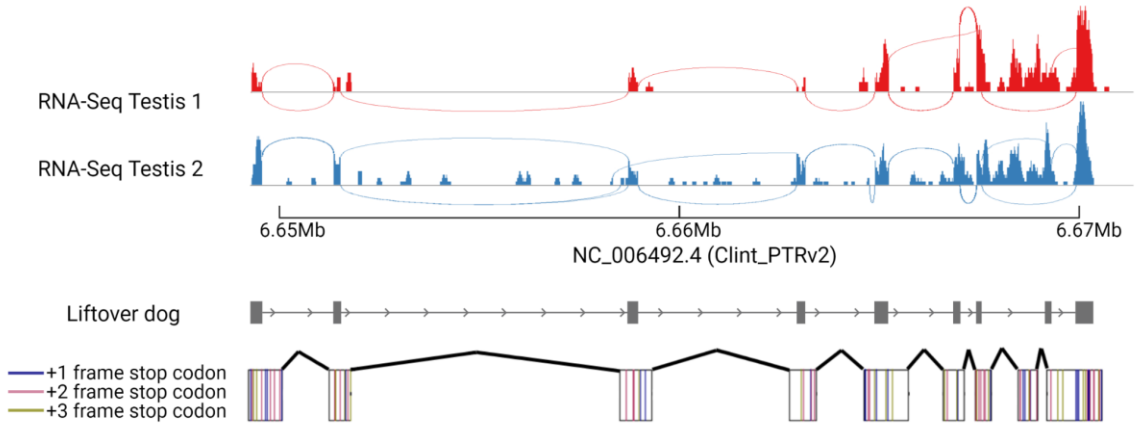


Figure 6.20. Pseudogenization of *PRSSLY* in chimpanzees. In the *PRSSLY* locus, RNA-Seq reads overlap with a liftover annotation from the *PRSSLY* of dogs. The chimpanzee *PRSSLY* gene is predicted to be a pseudogene, as stop codons are observed in all three reading frames.

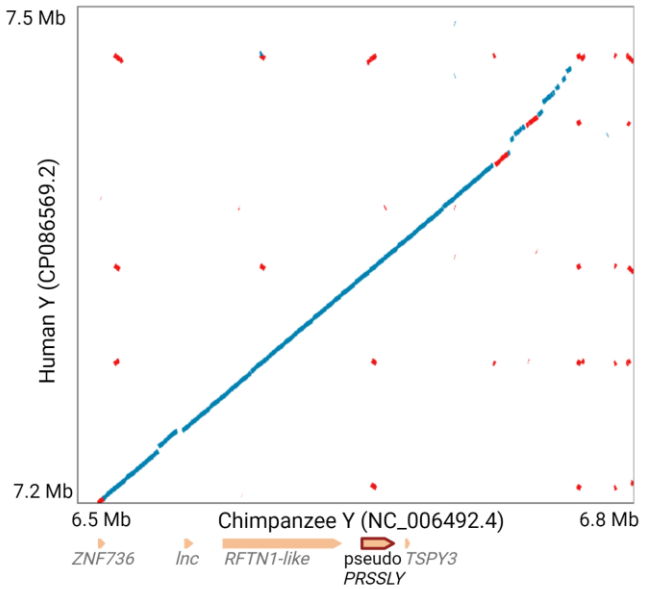


Figure 6.21. *PRSSLY* loci in the chimpanzee and human. Syntenic plot between the chimpanzee and human for the *PRSSLY* locus and its flanking regions indicates conversation in this region.

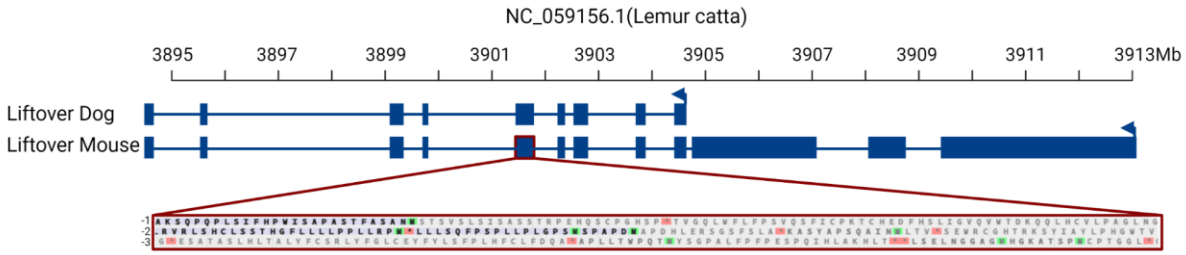


Figure 6.22. Pseudogenization of *PRSSLY* in the lemur (*Lemur catta*). The structure of the *PRSSLY* gene was defined based on annotation liftover from dogs and mice, and one of the potential exons had stop codons in all three reading frames.

For the other analysed Lemuriformes species, the gray mouse lemur, its male genome had no detectably homologous sequences. However, *PRSSLY* reads were identified in testis RNA-seq and they could reconstitute a full-length transcript. Moreover, the predicted peptide was found to be homologous with other species' PRSSLY.

Although its absence from the genome assembly could be attributed to the assembly's poor quality, the grey mouse lemur's *PRSSLY* transcript was investigated further. The similarity of other primates' dead exons shared equivalent pairwise nucleotide comparisons within the Haplorhini species (**Figure 6.23**). In contrast, the similarity between the grey mouse lemur's exons and those of Haplorhini revealed that the linker domain's exon sequence was noticeably more divergent than the exons corresponding to trypsin domains (**Figure 6.24**). Also, the linker domain's exon was seen to accumulate more deleterious variants than the trypsin domains in humans and chimpanzees (**Figure 6.25**). During evolutionary history, deleterious variants in the pseudogene were neutrally selected, while in the grey mouse lemur's coding gene, they underwent purifying selection. The trypsin domains showed less deleterious variants, indicating their ability to maintain coding functionality for a longer duration compared to the linker domain. All these results revealed a dynamic evolution pattern for *PRSSLY*: the linker domain's exon was pseudogenised prior to the trypsin domains, and all domains were dead within the common ancestor of the *Haplorhini* (**Figure 6.26**). In this case, the linker domain's exon was non-functional for a longer time than the trypsin domain exons, resulting in higher sequence divergence in the linker domain exon than in the

trypsin domain exons and enrichment of detrimental variants in the desert exon. All the exons' similarity at the same level when compared within the *Haplorhini* species excluded the possibility that the *PRSSL Y* was pseudogenised after the *Haplorhini* species separated.

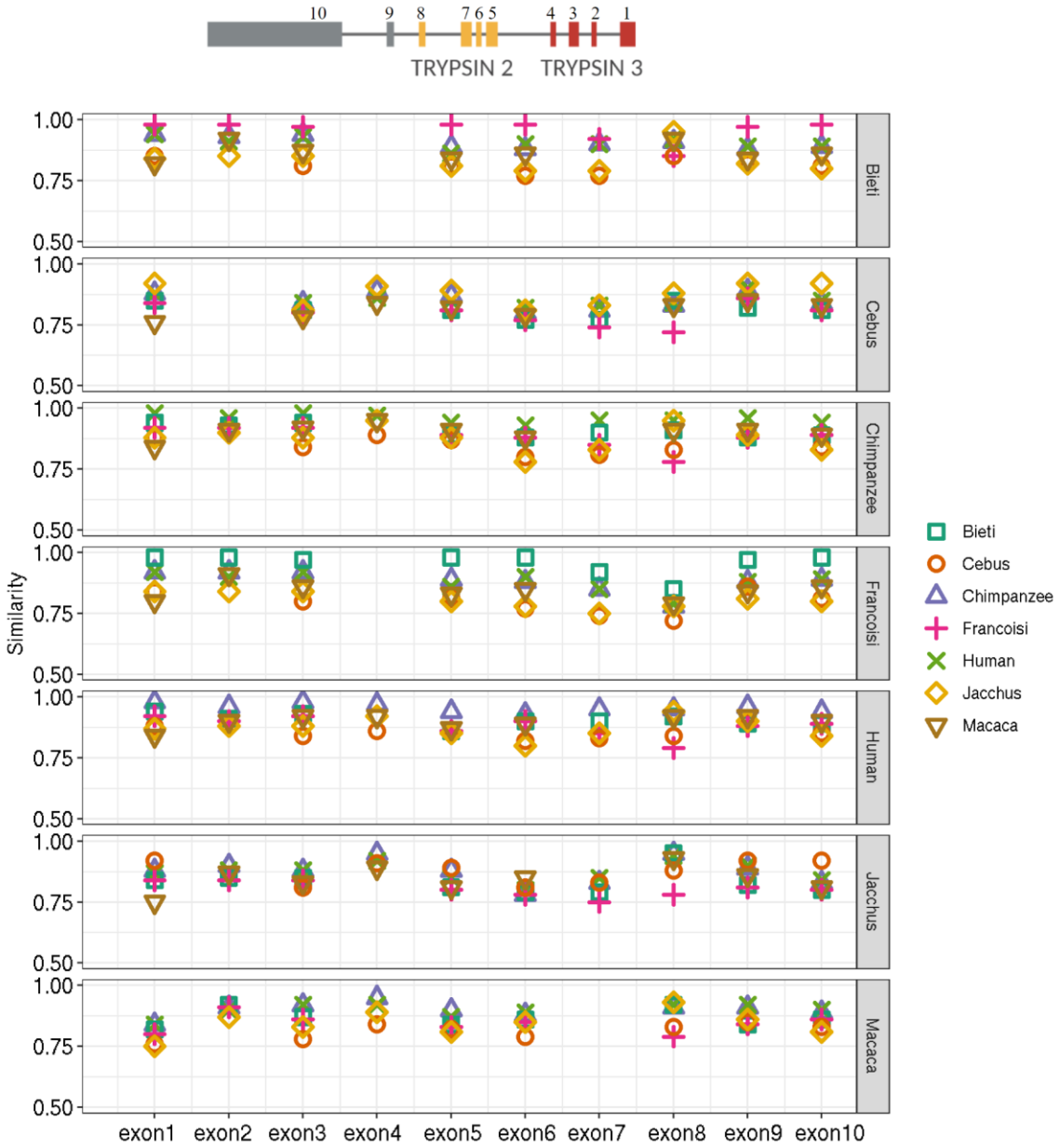


Figure 6.23. Paired comparison of nucleotide similarity among Haplorhini species at the exon level. The exons are numbered 1 to 10 from the last exon to the linker exon accordingly.

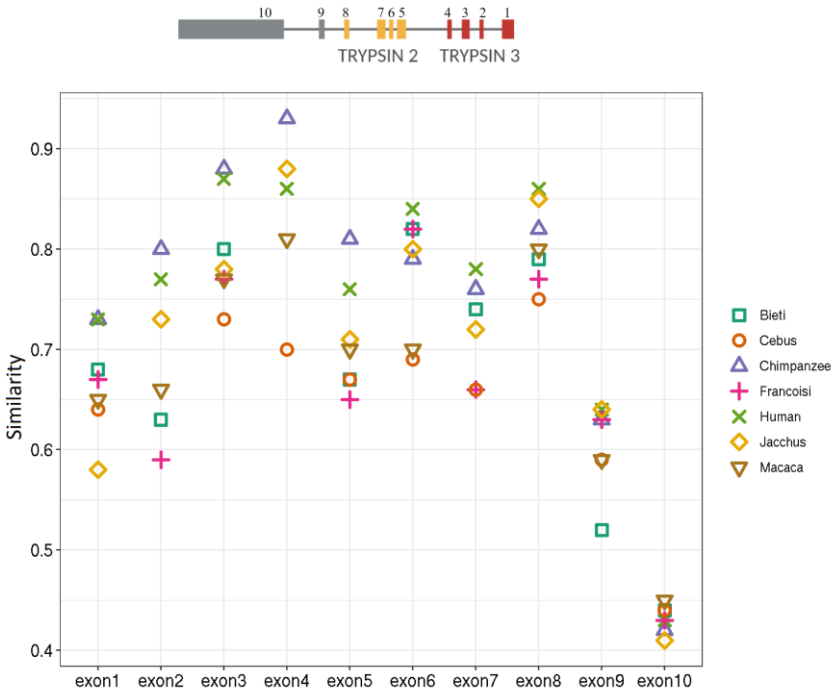


Figure 6.24. The similarity of Haplorhini species compared with mouse lemur for each exon.

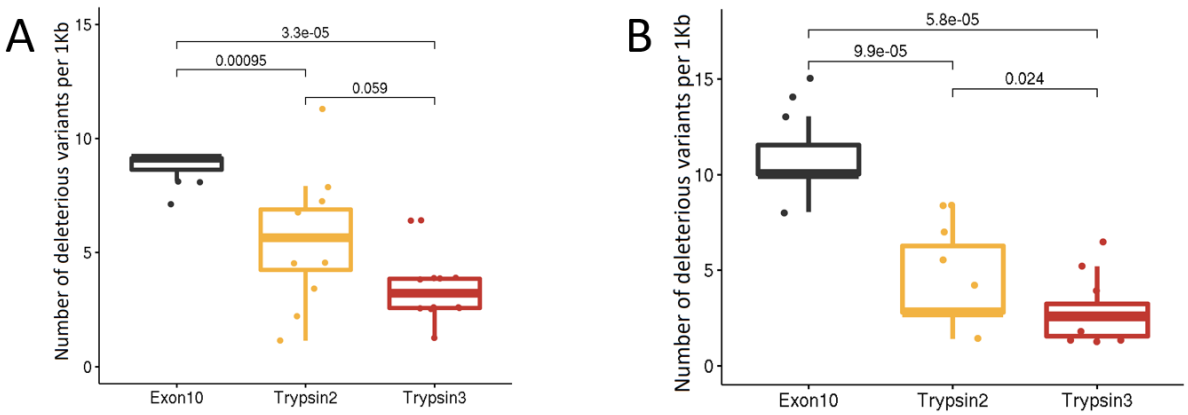


Figure 6.25. Enrichment of deleterious variants in the linker domain. Number of deleterious variants per 1 kb is calculated for human (A) and chimpanzee (B) in exons corresponding to the linker domain (exon 10), trypsin 2, and trypsin 3. Deleterious variants are characterized by insertions or deletions that induce frameshifts or nucleotide changes that introduce a stop codon.

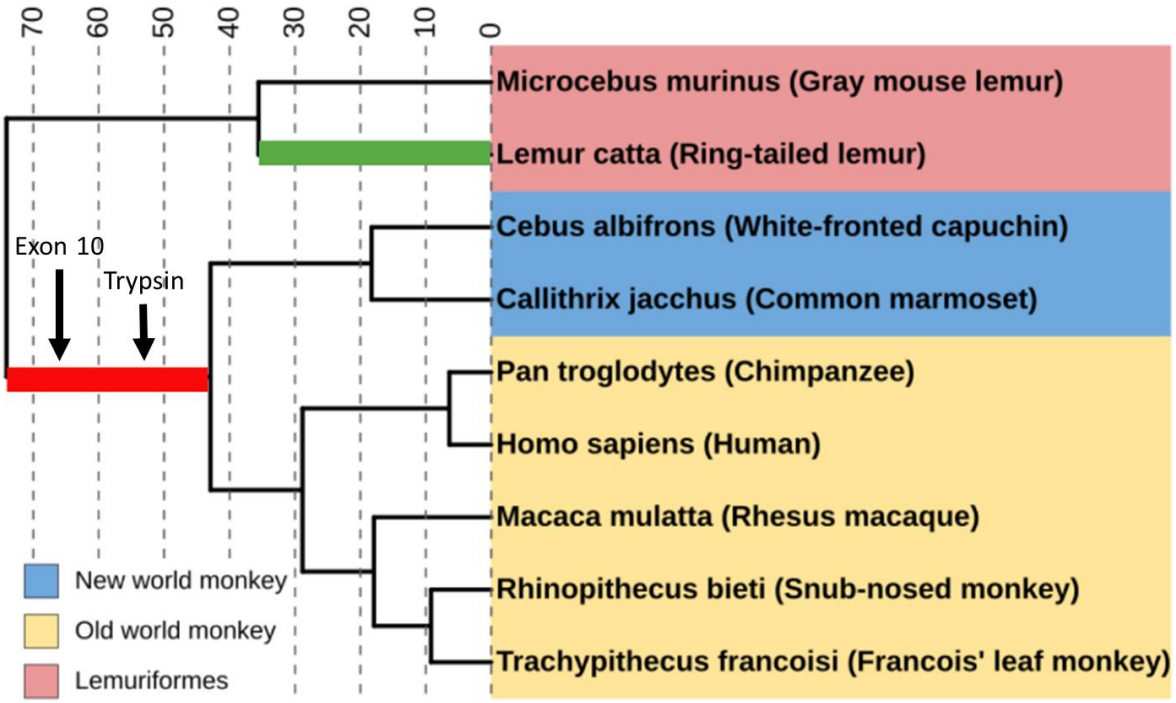


Figure 6.26. Schematic representation of stepwise pseudogenization for *PRSSLY*. Evolutionary tree in primates demonstrates a dynamic process of pseudogenization for *PRSSLY*. For the common ancestor of Haplorhini (the highlighted red branch), the linker domain and trypsin domains lost their coding capability in turn (shown by arrows). Ring-tailed lemur has a pseudogenization event independent of the Haplorhini species (the highlighted green branch).

6.3.4 Spatiotemporal Expression of Mouse *Prssly*

The expression of *Prssly* in different tissues was quantified with the bulk RNA-Seq samples. *Prssly* displayed tissue-specific with its expression only being seen in the testis (**Figure 6.27**). Therefore, bulk RNA-Seq of testis from different developmental stages was analysed, and the timing of *Prssly* expression was four weeks post-birth (**Figure 6.28**).

To examine the cell types that expressed *Prssly* in testis, published single-cell data were reanalysed. Seven marker genes (*Cd14*, *Sox9*, *Id4*, *Zpbp*, *Ddx4*, *Prm3*, *Prm2*) were selected as they are expressed.

The k-means clustering divided all the cells into 12 clusters and the full set of 12 clusters was simplified into 9 major classes according to the expression pattern of molecular markers. As seen, the *Prssly* was mainly expressed in the round spermatids (**Figure 6.29, Figure 6.30**), consistent with the result of RNA FISH in dogs (**Figure 6.8**).

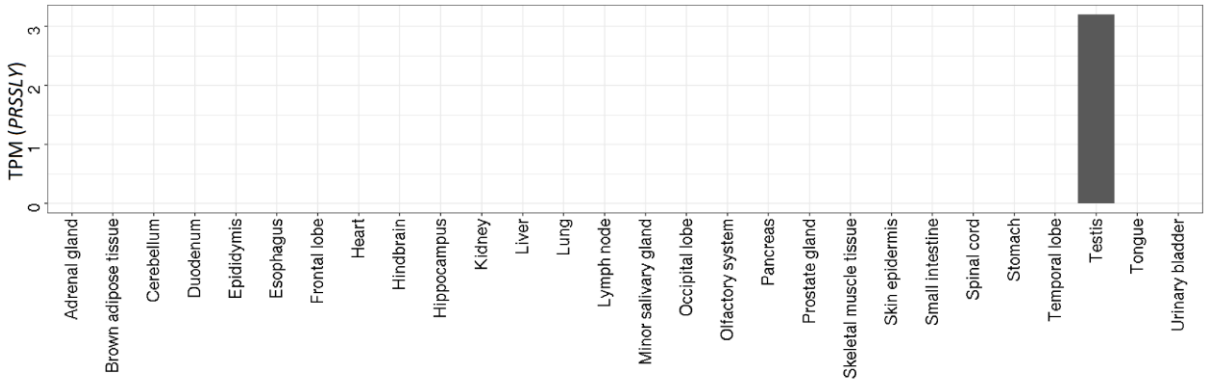


Figure 6.27. Expression level of the *PRSSLY* across various tissues in dogs. *PRSSLY* is testes-specific.

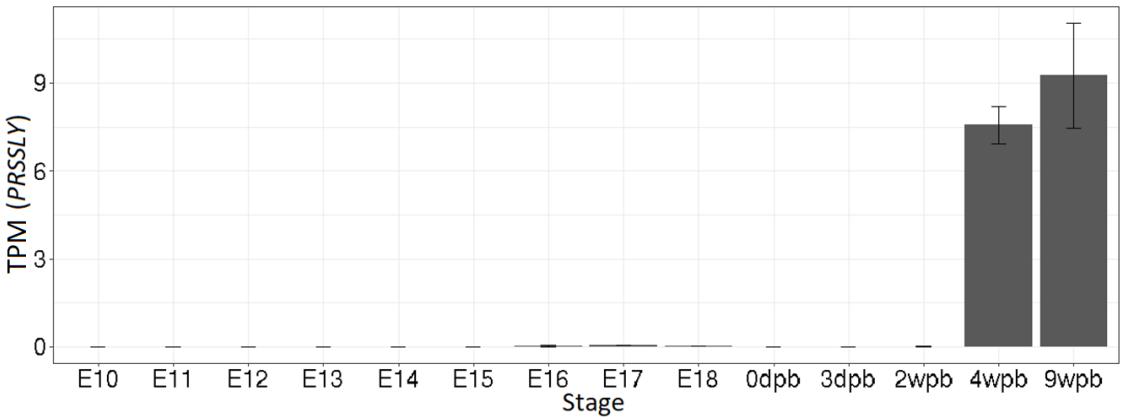


Figure 6.28. RNA-seq analysis of PRSSLY across development in mice. E: Embryo, dpb: days post birth, wpb: weeks post birth.

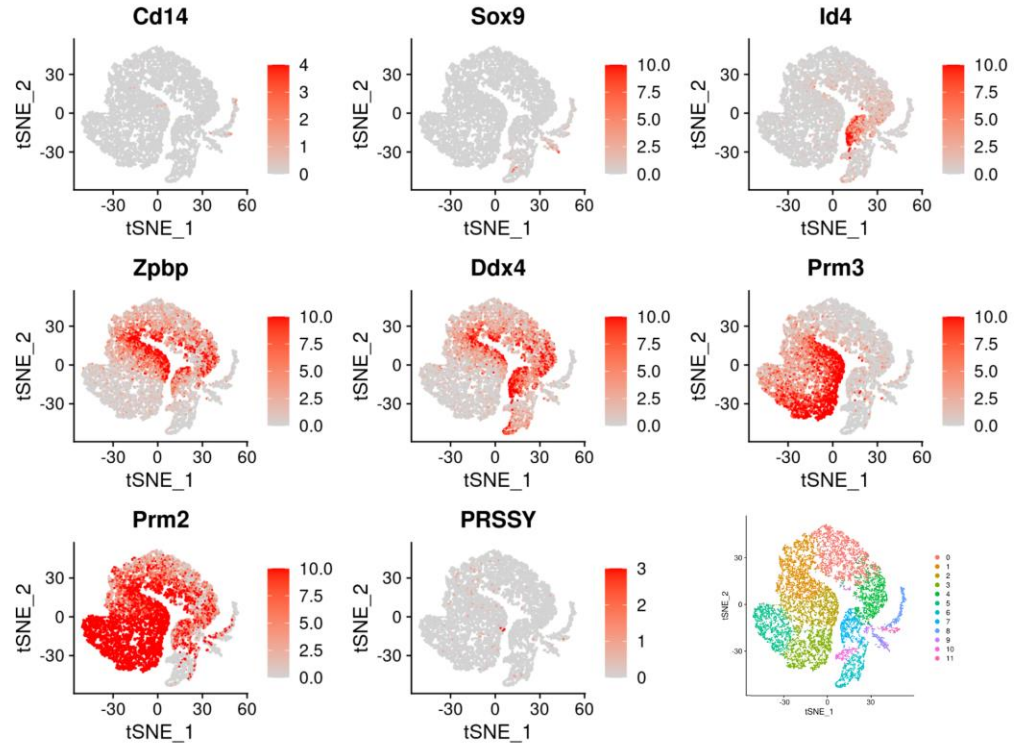


Figure 6.29. t-SNE projection of single cells and clustering in Seurat. Single-cell RNA sequencing analysis reveals the expression pattern of *PRSSLY* in mice, and tSNE and clustering analysis of single-cell data. Expression patterns of selected genes projected on the tSNE plot. Red and gray indicate a high and low abundance expressed individually. A high-resolution figure is available at https://github.com/WengangXbio/script_bio/blob/main/Figure%206.29.png

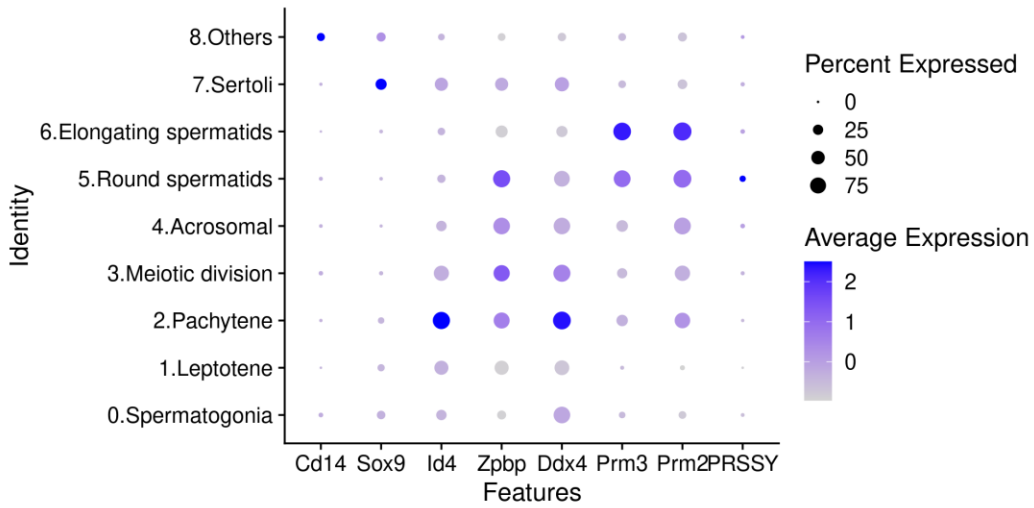


Figure 6.30. Quantification of selected makers and *Prssly* across all cell types. The expression level is reflected by colour and the percent of expressed cells for each type is indicated by the size of dots.

6.4 Discussion

This chapter identified a unique gene, *PRSSLY*, on mammalian sex chromosomes whose evolutionary pattern has never been fully investigated previously. *PRSSLY* survived on the Y chromosomes in most of the placental mammal lineages and was lost on the X chromosomes. For marsupials, *PRSSLY* was retained on the X chromosomes and disappeared on the Y chromosomes. The *PRSSLY* gene was derived from a single origin of an ancestral autosome gene, which is still maintained as an autosomal gene in extant monotremes.

The analysis of this chapter are in agreement with many of the findings described by Hughes et al, which were published concurrent to these analyses (74). Nevertheless, the analyses presented here reveal some differences and they extend our understanding of *PRSSLY* gene evolution and suspected function.

First, the annotation used by Hughes et. al. and Hayden et. al. of the *PRSSLY* for mice was erroneous, and a corrected gene model was determined *in silico* in this chapter. In theory, the existence of multiple-copy of *PRSSLY* on the mouse Y chromosome genome could lead to an introduction of stop codons or a shift of reading frames by INDELs once the pseudogenised *PRSSLY* sequences were mis-assembled into functional *PRSSLY* loci. The alignment of RNA-Seq data on the crude sequence of NW_001034423.1 generated a 'plausible' gene model, but its transcribed region was predicted as four ORFs with head-to-tail tiling from the beginning to the end of the real *PRSSLY*. It looks like the four ORFs predicted by RefSeq were truncated from an entire ORF as many mismatches between the crude reference sequences and RNA-Seq data were seen (**Figure 6.11**). After polishing NW_001034423.1 using RNA-Seq reads, a complete gene model was annotated with a long ORF that linked the four shorter ORFs together. With only one peak being observed immediately upstream of the corrected gene model, CAGE-Seq also supported a long transcript existing in this locus rather than four short ORFs corresponding to four independent transcripts, in which case every transcript should have a CAGE-Seq peak at TSS (**Figure 6.10**).

The corrected mouse *Prssly* gene model contained 14 exons with the longest one of 9585 bp encoding a 3195 aa peptide. Proteomics analysis proved that this extremely long exon was translated. Based on tissue expression data, *PRSSLY* is only expressed in testes, so mouse proteomics data were grouped into the testis and the non-testis tissues followed by searching for unique peptides of *PRSSLY* in proteomics data. Testis samples identified 41 unique peptides belonging to *PRSSLY*, some of which supported the translation of the longest exon. One unique peptide was detected in other tissues, perhaps the result of sporadic expression or the false discovery. Hayden et al. (344) proved that *PRSSLY* had an extension at the 3' end of Hughes's annotation through 3' RACE-Seq, and our works predicted the full length of *PRSSLY in silico*, which also had an extension at the 5' end of the previous one.

Two knockout studies independently concluded that mouse *Prssly* is dispensable for male fertility (85,344). They disrupted *Prssly* by introducing frameshifts at exons 6, exon 8 and exon 5 of Hughes's gene model. While, when considering the full-length gene model, it is unknown whether the modified *PRSSLY* really lost its functions. This is because their DNA editing only disrupted the translation of the peptides after the mutated sites, which corresponded to the functional domains of trypsin 2 and trypsin 3; it is possible that trypsin 1 domain was translated successfully and retained sufficient function. Therefore, to explore its relationship with male fertility, reliable *PRSSLY* knock-out mice are warranted in the future. Moreover, Hughes et al. found that *Prssly* mutant mice produced offspring in favour of females. This observation is believed to be a random phenomenon or caused by other factors when the published data was reanalysed (**Table 6.1**). On the one hand, there was no significant difference between mutated and wild mice for sex ratio in Chi-Square tests; on the other hand, Holmlund's mutated mice did not show a consistent observation on sex ratio, which had a skew towards males.

In summary, *PRSSLY* is a gene with male-specific functions retained on the Y chromosomes for most mammalian lineages. In this chapter, the gene model for the *PRSSLY* was corrected and fully annotated, and evidence of the transcription and translation of *PRSSLY* in dogs and mice was provided. Primates showed a dynamic

degeneration of *PRSSLY* inferred by comparative analyses. The analysis presented an accurate annotation of the *PRSSLY* gene, which is important to future studies such as gene editing, transcriptomic analysis, and proteomics analysis, and provided a fundamental understanding of this gene.

CHAPTER 7: Conclusion and Discussion

Mammalian sex chromosomes descended from paired ancestral autosomes 180 million years ago (MYA), and the Y chromosome has evolved with substantial distinction across species in terms of size, structure and organisation. Previously, our general understanding of mammalian Y chromosomes mainly came from well-assembled genomes from animals such as primates, mice, and bulls. The dog is one of the most popular companion animals in the world, and the investigation of its Y chromosome is paramount to addressing welfare concerns regarding male dogs fertility and development. Additionally, the dog Y chromosome can contribute to phylogenetic completeness in investigations of mammalian Y chromosome evolution. To this end, this thesis endeavoured to construct dog Y chromosome sequences that enable its gene contents, structure, and evolution to be processed.

7.1 Is it Necessary to Improve the Assembly of the Dog Y Chromosome Further?

In this thesis, a novel assembly approach was used to generate dog Y chromosome sequences. This method used PacBio long reads to generate contigs, followed by implementing Bionano and Hi-C technologies to orient and connect contigs into the scaffold level. RosY_1.0 consisted of 3 scaffolds in a total length of 6.78 Mb with gaps of 0.26 Mb. Although RosY_1.0 is the most complete assembly of the dog Y chromosome to date, including single-copy regions, multiple-copy regions, and PAB sequences, this assembly still lacks sequences that correspond to the centromere, NOR, and parts of highly repetitive regions. In this study, the PacBio sequencing was conducted by the Sequel I system, where the error rate ranges from 10% to 15%. This high error rate of long reads was prone to collapse multiple repetitive units into single copy during assembly. As seen in RosY_1.0, the 0.26 Mb gap showed high repeats and palindrome patterns with its flanking regions, and around 300 copies of LINE1_CF tandem repeat only presented as 9 copies.

To improve the assembly of the dog Y chromosome, the newest PacBio platform (Sequel II and Revio) and Nanopore ultra-long DNA sequencing can be used in the future. These technologies were developed recently, and have been applied in genome assembly studies (360–363). The accuracy of sequencing in PacBio HiFi reads by Sequel II is as high as 99.5% and Revio HiFi reads is even 99.95%, which can distinguish variants among repetitive copies. Such highly accurate long reads may enable assembly across long stretches of repetitive DNA through the identification of distinguishing variants as markers. For the Nanopore ultra-long DNA sequencing, sequenced reads can reach 300 Kb in length with an accuracy of approximately 95% (173,364). Although currently Nanopore reads cannot differentiate repeat units with the equivalent resolution as PacBio sequencers, complicated regions can be resolved by single reads that span repetitive sequences (364).

Also, flow-sorting and BAC can be used to improve the quality of assembly. Flow-sorting method is to isolate and sequence Y chromosome sequences individually (**Section 1.2.3**). But if enriched DNA is sequenced with error-prone technologies such as Nanopore and PacBio Sequel I, the complicated regions are still challenging to generate completely. BAC methods (**Section 1.2.3**) can assemble genome sequences by sequencing DNA in a block manner. However, it costs money and time that is exponentially higher than PacBio and Nanopore sequencing.

A recent emerging technology by Oxford Nanopore called 'adaptive sequencing', could help with the Y chromosome's assembly. By switching the voltage across individual nanopores to reject sequences, Nanopore sequencers can be applied to selectively examine DNA molecules in a pool, enabling targeting user-defined sequences (365). Combining with the newest R10.4.1 flow cell, whose accuracy averages 99% (366), this performed with ultra-long reads may be able to improve the assembly of the Y chromosome beyond what is reported here.

7.2 Evolution and Functions of MSY Genes

MSY genes were annotated with multipronged methods using Iso-Seq, RNA-Seq, and protein sequence data. Four novel coding genes (*WWC3Y, AP1S2Y*, *TMSB4Y*, and *PRSSY*) were detected. The MSY genes were categorised into three groups: ubiquitous, low-expression, and testis-specific. Inferred from the sequence divergence analyses and expression pattern, ubiquitous genes were implicated in housekeeping functions for somatic cells, and testis-specific genes may be functional in spermatogenesis and male development.

The functions of MSY genes were mainly studied in humans and mice, and not investigated in dogs before. Examples like *UBE1Y*, which was testis-specific in mice and broadly expressed in dogs, indicates the functions of dog MSY genes may differ to MSY genes of mice. Therefore, more molecular studies to understand the roles of MSY genes are needed for dogs. Knockout is one of the ways to explore the function of a gene. But, in most cases, using genetic modification on dogs is impractical due to ethical issues and welfare concerns. As an alternative, pluripotent stem cells, which are either reprogrammed from somatic cells or directly derived from embryos, have the ability to develop any type of cell or tissue in the body (367,368). So, this makes it plausible to study gene functions by editing genes in cell lines.

With the growing number of genome assemblies from diverse mammalian species, orthologs and paralogs of dog MSY genes were discovered across mammals, which enables us to construct phylogenies. The outcomes reflected a dynamic evolution of dog Y chromosomes, where historical conversion events between X and Y-linked genes took place at different time points. The limited number of caniformia species, especially canid species, is a deficit in our analysis. The closest analysed species available currently is the dhole, so more recent gene conversion after the split of dholes and dogs cannot be found in this study. To date, there is no genome assembly or Y chromosome assembly for coyotes, golden jackals, Ethiopian wolves, black-backed jackals, or side-striped jackals. Also problematic, the African wild dogs and dholes' Y chromosome was assembled with Illumina short reads (369). Therefore, to fully understand the evolution of dog MSY

genes, Y chromosome assemblies from canids species will be needed in the future. These will need to be produced using cutting-edge sequencing and scaffolding technologies.

Dogs and wolves have 2-4 copies of *SRY* on their Y chromosomes. In humans and mice, the normal Y chromosome carries one copy of *SRY*, while some rodent species possess multiple-copies of *SRY* genes in variable amounts (370–372). Even-toed ungulates such as pigs consistently present with two copies of *SRY* in their populations (69,373), and donkeys, horses, bulls and buffaloes displayed a variable number of *SRY* copies among individuals (374–377). For the carnivorans, cats presented single or multiple-copy of *SRY* and other *Carnivora* species have not been studied (55,78,378). Notably, the dog is the second reported case where *SRY* is located within the spacer of palindrome sequences after mice (100) (**Figure 7.1A**). In mammals, another organisation is that *SRY* resides within palindrome arms such as in pigs (69) and rabbits (379) (**Figure 7.1A**). The palindrome structure enables intrachromosomal recombination coupled with gene conversion or crossing over between arms or palindromes (71,380,381) (**Figure 7.1B**). Gene conversion potentially inhibits stochastic Y chromosome erosion, leading to the maintenance of gene functions (380,382). In dogs, which contain two *SRY* genes in the spacer of palindromes, gene conversion can be achieved by recombination between the two palindromes' loops (**Figure 7.1C**) or recombination between homologous palindromes (**Figure 7.1D**). As cats do not exhibit this palindrome structure of *SRY*, it is clear that dogs developed it after the separation of dogs and cats. More finished dog and wild canid Y chromosome assemblies are required to fully comprehend the evolution of *SRY* genes in dogs, including when they were embedded with palindromes, how many duplication events of *SRY* and its flanking palindromes have taken place within the dog population, and the divergence rate of *SRY* genes. Meanwhile, we will get more insight into *SRY* copy number variation, structures, and sequence divergence in relation to other canids.

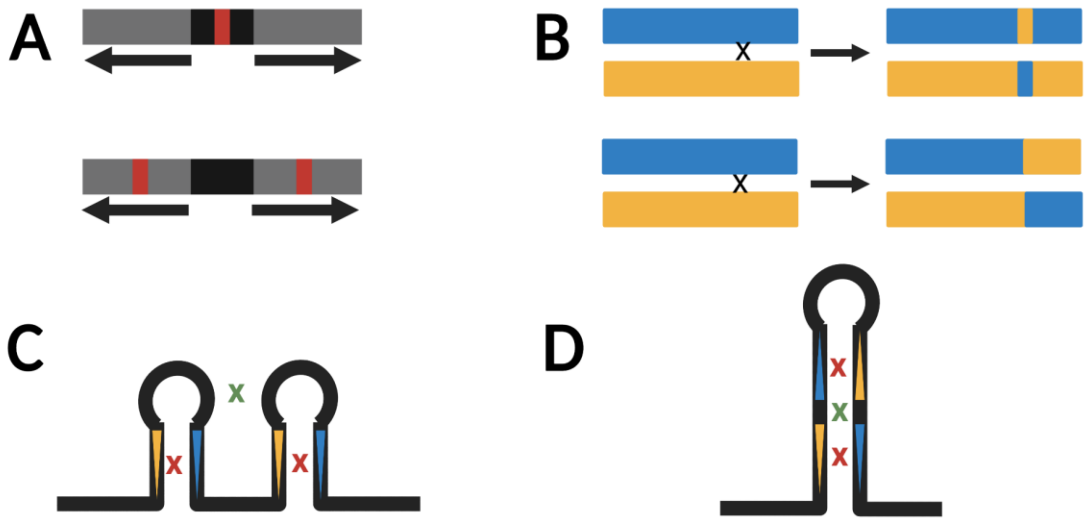


Figure 7.1. Hypothesized mechanism of gene conversion by homologous recombination in MSY palindromes. (A) Schematic representation of *SRY* located within palindromes. The top figure presents that *SRY* is located in the spacer of palindrome sequences, such as dogs and mice; and the bottom figure shows that *SRY* resides within the arms of palindromes, such as bulls and rabbits. The two arms of palindromes are shown as grey boxes, whose direction is indicated by arrows. Black boxes refer to the spacer of palindromes and red boxes indicate the *SRY* genes. (B) Recombination between homologous sequences. DNA double-strand break and resolution can yield gene conversion (top) or crossing over (bottom). (C) Intrachromosomal recombination between two arms of palindromes. (D) Intrachromosomal recombination between two homologous palindromes. For *SRY* that is located in the spacer of palindromes, *SRY* copies can be converted between two loops (green 'X' in C) or between spacers' homologues (green 'X' in D). For *SRY* that resides within the arms, gene conversion can occur between two arms of the same palindromes (red 'X' in C) or between arms of different palindromes (red 'X' in D).

7.3 PAB Variation within dogs and among Canids

In this thesis, PAB sequences were defined for the sequenced Labrador retriever, which enabled the examination of its structure, gene contents, and characteristics. We already knew that its PAB was 5 Kb in length, and that the *TETY2* and *CLDN34* were embedded within the Y-linked and X-linked PABs, respectively. Two SINEs were inserted on the Y-linked PAB at various times, suggesting a dynamic evolution of the

PAB, which raised the question of whether the PAB existed stably within the dog population. For instance, mice showed such high variability in PAB that distinct subspecies developed with independent structures. Dog Y chromosome haplotypes were seen to be separated from another as early as 50,000 years ago, and their PABs may have evolved in various ways.

Another question is whether the canid species had the same PAB intervals as dogs. Despite the discovery of syntenic X-linked and Y-linked PAB sequences in the maned wolf, African wild dog, and Tibetan fox, the precise point at which the recombination between X and Y chromosomes starts to be inhibited is uncertain. These breakpoints may be expanded further or even encapsulate the closest PAB gene, *SHROOM2*, as is known to occur in pigs. Exploring diverse movement rates of PAB can help us understand the mechanisms underpinning PAB movement, and studying gene degradation on the Y-linked PAB can help us better understand how dosage compensation for mammalian sex chromosomes works.

7.4 Future Research Direction for the *PRSSLY* gene

The first instance of *PRSSLY*, which retained its Y chromosome, but lost its X chromosome, was discovered. According to bulk and single-cell RNA-Seq, and FISH analyses, one can conclude its functions are related to spermatogenesis. Two independent knockout (KO) experiments concluded that *PRSSLY* is dispensable for male fertility (85,344), but their annotation was erroneous, potentially leading to an ineffective KO and false conclusions.

For future exploration of *PRSSLY*, its gene model in mice generated in this study has to be fully validated. Although our revised gene model was confirmed by CAGE-Seq, RNA-Seq, and proteomics data, full length RNA sequencing data is still lacking, which would provide the direct evidence for the model of *PRSSLY*. There are a few Iso-Seq datasets from mice's testes, but *PRSSLY* was not expressed in these data. Therefore, deeper Iso-Seq sequencing for adult mice's testes is needed.

Secondly, the conclusion drawn from two KO experiments is still uncertain, although not entirely dismissed. This is exemplified by the fact that human genes *CDY1*, *BPY2*, and *DAZ*, which are exclusively expressed in the testis, have been found not to be essential for male fertility. A future objective should be to determine whether *PRSSLY* is necessary for male fertility or not. To achieve this, mutation sites should be designed in the first few exons to disrupt all the trypsin domains, and a successful knockout should be confirmed by techniques such as Iso-Seq and proteomics.

If *PRSSLY* is proved necessary for male fertility, further studies will focus on the molecular mechanism of infertility. Considering *PRSSLY* is mainly expressed in the round spermatids, the absence of PRSSLY protein may disrupt pathways or down-regulate downstream genes in the process of spermatogenesis.

If mice KO for *PRSSLY* are still fertile, here, I am proposing one hypothesis that *PRSSLY* may be involved in the process of testicular thermoregulation or testicular heat stress. In mammals, testicular thermoregulation is a critical process that maintains the optimal temperature for spermatogenesis in the testes. The testes require a temperature that is approximately 2-4°C lower than the core body temperature for proper functioning. To maintain an optimal temperature, the testes rely on a number of thermoregulatory mechanisms, including the countercurrent heat exchange system, the regulation of blood flow to the testes and the descended scrotum. Testicular heat stress is a condition where testicular temperature is dysregulated. When the testicles are exposed to high temperatures for a prolonged period of time, deleterious outcomes are expected such as decreased sperm production and lower sperm quality. There are three reasons why we think *PRSSLY* is a candidate gene for testicular thermoregulation or heat stress. First, the camel, which adapts to desert conditions, is the only species that contained four copies of *PRSSLY* based on our analyses (data not shown). It is possible that multiple-copies of *PRSSLY* are positively selected within the camel population under an extremely hot environment in favour of male fertility. Second, *PRSSLY* is missing or pseudogenised in Simiiformes and Feliformia. We observed their loss of functional *PRSSLY* occurred earlier than 50 million years ago, when Earth was in a warm period (**Figure 7.2A**). We also

observed the primates and Feliformia are geographically distributed around tropics, where the temperature is relevant higher than other areas (**Figure 7.2B-C**). Compared to Feliformia, Caniformia are more uniformly distributed across the world (**Figure 7.2D**). We hypothesed that the common ancestors of primates and Feliformia were living under a hot environment for a long time leading to the creation of different mechanisms responding to heat stress. As a result, *PRSSLY* was unnecessary in these two linages but still took functions in other mammals. Third, *PRSSLY* was shown to be exclusively expressed in the testis in eight investigated eutherian mammals (**Figure 7.3A**) and abundantly expressed across somatic tissues in marsupials, monotremes, and lizards (**Figure 7.3B**). In terms of the structure of *PRSSLY*, the linked domain's exon of marsupials and monotremes were short and conserved, and the eutherians' were long and had less conservation (**Figure 6.18**). As known, the marsupials (383) and monotremes (384,385) escaped from testicular thermoregulation displaying a similar temperature between body and testis, and lizards had different sex-determination systems. It is suggested that *PRSSLY* was retained in eutherians with a particular function in temperature-sensitive testis, while it has broad functions for marsupials, monotremes, and lizards. Taken together, *PRSSLY* is implicated to be functional in testicular thermoregulation or testicular heat stress, and further experiments are warranted to dissect its biological roles.

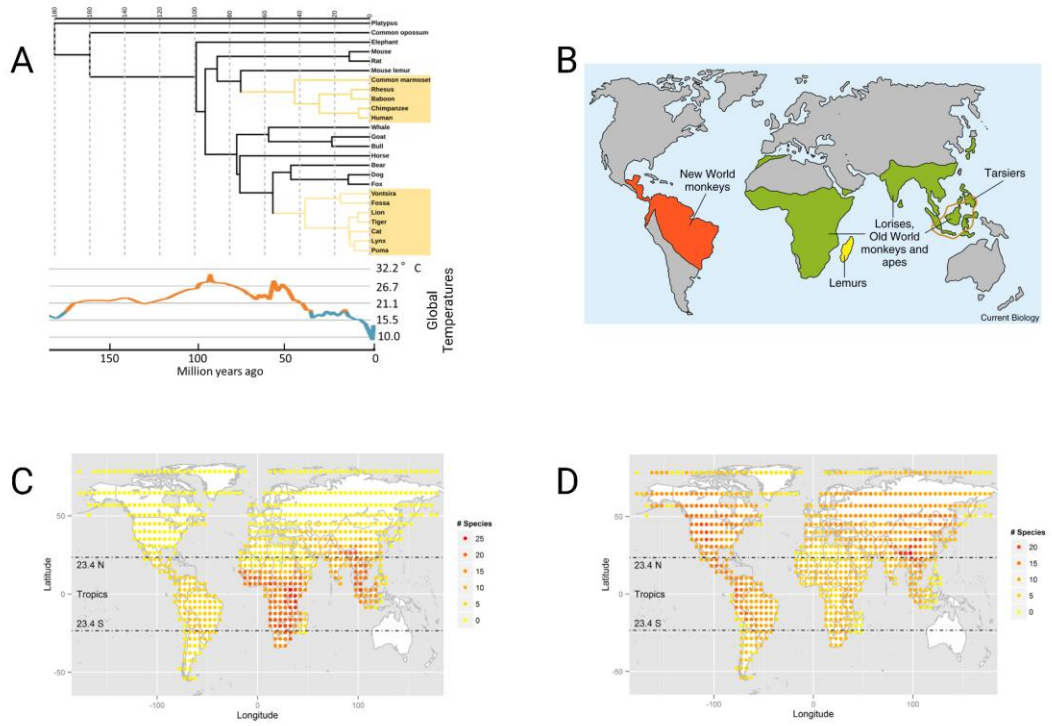


Figure 7.2. *PRSSLY* is a candidate gene for testicular thermoregulation. (A) Mammalian phylogeny scaled by divergence time and global temperature over 180 million years ago. (B) Geographical distribution of extant nonhuman primates. (C) Geographical distribution of Feliformia with the tropics marked. (D) Geographical distribution of Caniformia species with the tropics marked. Figures are adapted from previous studies (386–388).

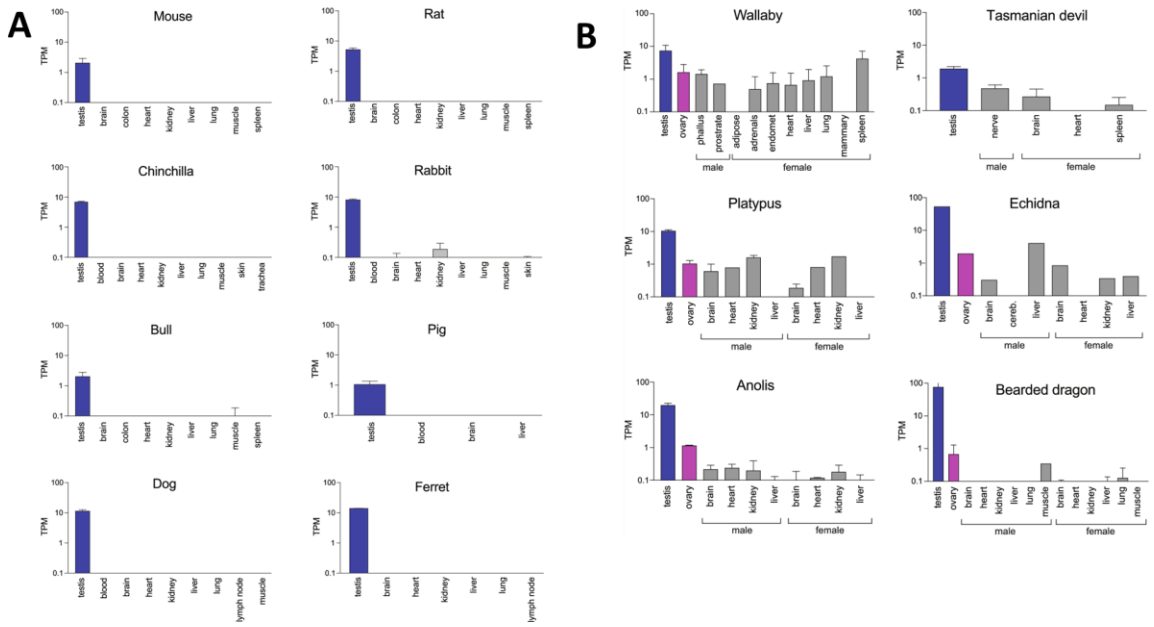


Figure 7.3. Expression level of *PRSSLY* in mammals and reptiles. (A) *PRSSLY* shows Testis-specific expression in *Placentalia* species. (B) *PRSSLY* is expressed broadly in marsupials, monotremes, and reptiles. Figures are modified based on previous research (85).

Reference List

1. Bellott DW, Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Cho TJ, et al. Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. Nature. 2014 Apr;508(7497):494–9.
2. Lahn BT, Pearson NM, Jegalian K. The human Y chromosome, in the light of evolution. Nat Rev Genet. 2001;2(3):207–16.
3. Jobling MA, Tyler-Smith C. Human Y-chromosome variation in the genome-sequencing era. Nat Rev Genet. 2017 Aug;18(8):485–97.
4. Soh YS, Alföldi J, Pyntikova T, Brown LG, Graves T, Minx PJ, et al. Sequencing the mouse Y chromosome reveals convergent gene acquisition and amplification on both sex chromosomes. Cell. 2014;159(4):800–13.
5. Hughes JF, Page DC. The Biology and Evolution of Mammalian Y Chromosomes. Annu Rev Genet. 2015;49(1):507–27.
6. Cannon-Albright LA, Farnham JM, Bailey M, Albright FS, Teerlink CC, Agarwal N, et al. Identification of specific Y chromosomes associated with increased prostate cancer risk. The Prostate. 2014;74(9):991–8.
7. Dumanski JP, Rasi C, Lönn M, Davies H, Ingelsson M, Giedraitis V, et al. Smoking is associated with mosaic loss of chromosome Y. Science. 2015 Jan 2;347(6217):81–3.
8. Charchar FJ, Bloomer LD, Barnes TA, Cowley MJ, Nelson CP, Wang Y, et al. Inheritance of coronary artery disease in men: an analysis of the role of the Y chromosome. Lancet. 2012 Mar 10;379(9819):915–22.
9. Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature. 2015 Jun;522(7555):207–11.
10. Eriksson J, Siedel H, Lukas D, Kayser M, Erler A, Hashimoto C, et al. Y-chromosome analysis confirms highly sex-biased dispersal and suggests a low male effective population size in bonobos (Pan paniscus). Mol Ecol. 2006;15(4):939–49.
11. Freedman AH, Wayne RK. Deciphering the Origin of Dogs: From Fossils to Genomes. Annu Rev Anim Biosci. 2017;5(1):281–307.
12. Germonpré M, Lázničková-Galetová M, Sablin MV. Palaeolithic dog skulls at the Gravettian Předmostí site, the Czech Republic. J Archaeol Sci. 2012 Jan 1;39(1):184–202.
13. Speijer D, Lukeš J, Eliáš M. Sex is a ubiquitous, ancient, and inherent attribute of eukaryotic life. Proc Natl Acad Sci. 2015 Jul 21;112(29):8827–34.
14. Maynard Smith JM, Szathmary E. The major transitions in evolution. OUP Oxford; 1997.
15. Williams GC. Sex and Evolution.(MPB-8), Volume 8. Vol. 8. Princeton University Press; 2020.
16. Lehtonen J, Kokko H, Parker GA. What do isogamous organisms teach us about sex and the two sexes? Philos Trans R Soc B Biol Sci. 2016 Oct 19;371(1706):20150532.
17. Cox PA. Introduction: the evolutionary mystery of gamete dimorphism. Evol Anisogamy. 2011;1–16.
18. Ezaz T, Stiglec R, Veyrunes F, Marshall Graves JA. Relationships between Vertebrate ZW and XY Sex Chromosome Systems. Curr Biol. 2006 Sep 5;16(17):R736–43.
19. Turnover of mammal sex chromosomes in the Sry-deficient Amami spiny rat is due to male-specific upregulation of Sox9 [Internet]. [cited 2023 Jul 26]. Available from: https://www.pnas.org/doi/10.1073/pnas.2211574119
20. Grutzner F. Evolution and Organization of Monotreme Sex Chromosomes. Oxford University Press; 2008.
21. Deakin JE, Graves J a. M, Rens W. The Evolution of Marsupial and Monotreme Chromosomes. Cytogenet Genome Res. 2012;137(2–4):113–29.
22. Rens W, O'Brien PC, Grützner F, Clarke O, Graphodatskaya D, Tsend-Ayush E, et al. The multiple sex chromosomes of platypus and echidna are not completely identical and several share homology with the avian Z. Genome Biol. 2007 Nov 16;8(11):R243.
23. Schnedl W. Analysis of the human karyotype using a reassociation technique.

Chromosoma. 1971;34(4):448–54.
24. Richard F, Lombard M, Dutrillaux B. Reconstruction of the ancestral karyotype of eutherian mammals. Chromosome Res. 2003;11:605–18.
25. Akeson EC, Davisson MT. Mitotic chromosome preparations from mouse cells for karyotyping. Curr Protoc Hum Genet. 2000;25(1):4.10. 1-4.10. 19.
26. Langford CF, Fischer PE, Binns MM, Holmes NG, Carter NP. Chromosome-specific paints from a high-resolution flow karyotype of the dog. Chromosome Res. 1996 Mar;4(2):115–23.
27. Carvalho B de A, Oliveira LFB, Nunes AP, Mattevi MS. Karyotypes of nineteen marsupial species from Brazil. J Mammal. 2002;83(1):58–70.
28. Grützner F, Rens W, Tsend-Ayush E, El-Mogharbel N, O'Brien PCM, Jones RC, et al. In the platypus a meiotic chain of ten sex chromosomes shares genes with the bird Z and mammal X chromosomes. Nature. 2004 Dec;432(7019):913–7.
29. Watson JM, Spencer JA, Riggs AD, Graves JA. The X chromosome of monotremes shares a highly conserved region with the eutherian and marsupial X chromosomes despite the absence of X chromosome inactivation. Proc Natl Acad Sci U S A. 1990 Sep;87(18):7125–9.
30. Veyrunes F, Waters PD, Miethke P, Rens W, McMillan D, Alsop AE, et al. Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. Genome Res. 2008 Jun 1;18(6):965–73.
31. Cortez D, Marin R, Toledo-Flores D, Froidevaux L, Liechti A, Waters PD, et al. Origins and functional evolution of Y chromosomes across mammals. Nature. 2014 Apr 24;508(7497):488–93.
32. Bachtrog D. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. Nat Rev Genet. 2013 Feb;14(2):113–24.
33. Berta P, Hawkins JB, Sinclair AH, Taylor A, Griffiths BL, Goodfellow PN, et al. Genetic evidence equating SRY and the testis-determining factor. Nature. 1990 Nov;348(6300):448–50.
34. Brashear WA, Bredemeyer KR, Murphy WJ. Genomic architecture constrained placental mammal X Chromosome evolution. Genome Res. 2021 Aug 1;31(8):1353–65.
35. Graves J a. M, Gécz J, Hameister H. Evolution of the human X – a smart and sexy chromosome that controls speciation and development. Cytogenet Genome Res. 2002;99(1–4):141–5.
36. Murphy WJ, Sun S, Chen ZQ, Pecon-Slattery J, O'Brien SJ. Extensive Conservation of Sex Chromosome Organization Between Cat and Human Revealed by Parallel Radiation Hybrid Mapping. Genome Res. 1999 Dec 1;9(12):1223–30.
37. Raudsepp T, Kata SR, Piumi F, Swinburne J, Womack JE, Skow LC, et al. Conservation of Gene Order between Horse and Human X Chromosomes as Evidenced through Radiation Hybrid Mapping. Genomics. 2002 Mar 1;79(3):451–7.
38. Lyon MF. Some Milestones in the History of X-Chromosome Inactivation. Annu Rev Genet. 1992;26(1):17–29.
39. Ohno S. Sex chromosomes and sex-linked genes. 2013.
40. Lucotte EA, Skov L, Jensen JM, Macià MC, Munch K, Schierup MH. Dynamic Copy Number Evolution of X- and Y-Linked Ampliconic Genes in Human Populations. Genetics. 2018 Jul 1;209(3):907–20.
41. Li G, Figueiró HV, Eizirik E, Murphy WJ. Recombination-Aware Phylogenomics Reveals the Structured Genomic Landscape of Hybridizing Cat Species. Yoder A, editor. Mol Biol Evol. 2019 Oct 1;36(10):2111–26.
42. Livernois AM, Graves J a. M, Waters PD. The origin and evolution of vertebrate sex chromosomes and dosage compensation. Heredity. 2012 Jan;108(1):50–8.
43. Janečka JE, Davis BW, Ghosh S, Paria N, Das PJ, Orlando L, et al. Horse Y chromosome assembly displays unique evolutionary features and putative stallion fertility genes. Nat Commun. 2018 Dec;9(1):2945.
44. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence

classes. Nature. 2003 Jun;423(6942):825–37.
45. Furman BLS, Metzger DCH, Darolti I, Wright AE, Sandkam BA, Almeida P, et al. Sex Chromosome Evolution: So Many Exceptions to the Rules. Genome Biol Evol. 2020 Jun 1;12(6):750–63.
46. Wright AE, Dean R, Zimmer F, Mank JE. How to make a sex chromosome. Nat Commun. 2016 Jul 4;7(1):12087.
47. Haigh J. The accumulation of deleterious genes in a population—Muller's Ratchet. Theor Popul Biol. 1978 Oct 1;14(2):251–67.
48. Charlesworth B. Model for evolution of Y chromosomes and dosage compensation. Proc Natl Acad Sci. 1978;75(11):5618–22.
49. Rice WR. Genetic Hitchhiking and the Evolution of Reduced Genetic Activity of the Y Sex Chromosome. Genetics. 1987 May 1;116(1):161–7.
50. Felsenstein J. The evolutionary advantage of recombination. Genetics. 1974 Oct;78(2):737–56.
51. Charlesworth B, Sniegowski P, Stephan W. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature. 1994 Sep;371(6494):215–20.
52. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993 Aug 1;134(4):1289–303.
53. Charlesworth B. The evolution of chromosomal sex determination and dosage compensation. Curr Biol. 1996 Feb 1;6(2):149–62.
54. Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Graves T, Fulton RS, et al. Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. Nature. 2012 Mar;483(7387):82–6.
55. Li G, Davis B, Raudsepp T, Wilkerson AP, Mason V, Ferguson-Smith M, et al. Comparative analysis of mammalian Y chromosomes illuminates ancestral structure and lineage-specific evolution. Genome Res. 2013 Jun 20;gr.154286.112.
56. Houck ML, Kumamoto AT, Jr DSG, Benirschke K. Comparative cytogenetics of the African elephant (Loxodonta africana) and Asiatic elephant (Elephas maximus). Cytogenet Genome Res. 2001;93(3–4):249–52.
57. Skov L, Consortium TDPG, Schierup MH. Analysis of 62 hybrid assembled human Y chromosomes exposes rapid structural changes and high rates of gene conversion. PLOS Genet. 2017 Aug 28;13(8):e1006834.
58. Cechova M, Harris RS, Tomaszkiewicz M, Arbeithuber B, Chiaromonte F, Makova KD. High Satellite Repeat Turnover in Great Apes Studied with Short- and Long-Read Technologies. Mol Biol Evol. 2019 Nov 1;36(11):2415–31.
59. Repping S, van Daalen SKM, Brown LG, Korver CM, Lange J, Marszalek JD, et al. High mutation rates have driven extensive structural polymorphism among human Y chromosomes. Nat Genet. 2006 Apr;38(4):463–7.
60. Oetjens MT, Shen F, Emery SB, Zou Z, Kidd JM. Y-Chromosome Structural Diversity in the Bonobo and Chimpanzee Lineages. Genome Biol Evol. 2016 Jul 1;8(7):2231–40.
61. Murphy WJ, Wilkerson AJP, Raudsepp T, Agarwala R, Schäffer AA, Stanyon R, et al. Novel gene acquisition on carnivore Y chromosomes. PLoS Genet. 2006;2(3):e43.
62. Brashear WA, Raudsepp T, Murphy W. Evolutionary conservation of Y chromosome ampliconic gene families despite extensive structural variation. Genome Res. 2018 Oct 31;gr.237586.118.
63. Giachini C, Nuti F, Turner DJ, Laface I, Xue Y, Daguin F, et al. TSPY1 copy number variation influences spermatogenesis and shows differences among Y lineages. J Clin Endocrinol Metab. 2009;94(10):4016–22.
64. Hughes JF, Skaletsky H, Pyntikova T, Koutseva N, Raudsepp T, Brown LG, et al. Sequence analysis in *Bos taurus* reveals pervasiveness of X–Y arms races in mammalian lineages. Genome Res. 2020 Dec;30(12):1716–26.
65. Cocquet J, Ellis PJI, Mahadevaiah SK, Affara NA, Vaiman D, Burgoyne PS. A Genetic Basis for a Postmeiotic X Versus Y Chromosome Intragenomic Conflict in the Mouse. PLOS Genet. 2012 Sep 13;8(9):e1002900.
66. Hughes JF, Skaletsky H, Pyntikova T, Graves TA, van Daalen SKM, Minx PJ, et al.

Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. Nature. 2010 Jan;463(7280):536–9.
67. Tomaszkiewicz M, Rangavittal S, Cechova M, Sanchez RC, Fescemyer HW, Harris R, et al. A time- and cost-effective strategy to sequence mammalian Y Chromosomes: an application to the de novo assembly of gorilla Y. Genome Res. 2016 Apr;26(4):530–40.
68. Yang Y, Chang TC, Yasue H, Bharti AK, Retzel EF, Liu WS. ZNF280BY and ZNF280AY: autosome derived Y-chromosome gene families in Bovidae. BMC Genomics. 2011 Jan 7;12(1):13.
69. Skinner BM, Sargent CA, Churcher C, Hunt T, Herrero J, Loveland JE, et al. The pig X and Y Chromosomes: structure, sequence, and evolution. Genome Res. 2016 Jan 1;26(1):130–9.
70. Bidon T, Schreck N, Hailer F, Nilsson MA, Janke A. Genome-Wide Search Identifies 1.9 Mb from the Polar Bear Y Chromosome for Evolutionary Analyses. Genome Biol Evol. 2015 Jul 1;7(7):2010–22.
71. Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, et al. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. Nature. 2003 Jun 19;423(6942):873–6.
72. Hallast P, Balaresque P, Bowden GR, Ballereau S, Jobling MA. Recombination Dynamics of a Human Y-Chromosomal Palindrome: Rapid GC-Biased Gene Conversion, Multi-kilobase Conversion Tracts, and Rare Inversions. PLOS Genet. 2013 Jul 25;9(7):e1003666.
73. Jobling MA. Copy number variation on the human Y chromosome. Cytogenet Genome Res. 2008;123(1–4):253–62.
74. Lahn BT, Page DC. Four evolutionary strata on the human X chromosome. Science. 1999;286(5441):964–7.
75. Mueller JL, Skaletsky H, Brown LG, Zaghlul S, Rock S, Graves T, et al. Independent specialization of the human and mouse X chromosomes for the male germ line. Nat Genet. 2013 Sep;45(9):1083–7.
76. Mácha J, Teichmanová R, Sater AK, Wells DE, Tlapáková T, Zimmerman LB, et al. Deep ancestry of mammalian X chromosome revealed by comparison with the basal tetrapod Xenopus tropicalis. BMC Genomics. 2012 Jul 16;13(1):315.
77. Lemaitre C, Braga MDV, Gautier C, Sagot MF, Tannier E, Marais GAB. Footprints of Inversions at Present and Past Pseudoautosomal Boundaries in Human Sex Chromosomes. Genome Biol Evol. 2009 Jan 1;1:56–66.
78. Wilkerson AJP, Raudsepp T, Graves T, Albracht D, Warren W, Chowdhary BP, et al. Gene discovery and comparative analysis of X-degenerate genes from the domestic cat Y chromosome. Genomics. 2008;92(5):329–38.
79. Wilson MA, Makova KD. Evolution and Survival on Eutherian Sex Chromosomes. Gojobori T, editor. PLoS Genet. 2009 Jul 17;5(7):e1000568.
80. Pandey RS, Wilson Sayres MA, Azad RK. Detecting Evolutionary Strata on the Human X Chromosome in the Absence of Gametologous Y-Linked Sequences. Genome Biol Evol. 2013;5(10):1863–71.
81. Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, Muzny D, et al. The DNA sequence of the human X chromosome. Nature. 2005 Mar;434(7031):325–37.
82. Cortez D, Marin R, Toledo-Flores D, Froidevaux L, Liechti A, Waters PD, et al. Origins and functional evolution of Y chromosomes across mammals. Nature. 2014 Apr;508(7497):488–93.
83. Kuroda-Kawaguchi T, Skaletsky H, Brown LG, Minx PJ, Cordum HS, Waterston RH, et al. The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. Nat Genet. 2001 Nov;29(3):279–86.
84. Yu YH, Lin YW, Yu JF, Schempp W, Yen PH. Evolution of the DAZ gene and the AZFc region on primate Y chromosomes. BMC Evol Biol. 2008 Mar 26;8(1):96.
85. Hughes JF, Skaletsky H, Nicholls PK, Drake A, Pyntikova T, Cho TJ, et al. A gene deriving from the ancestral sex chromosomes was lost from the X and retained on the Y chromosome in eutherian mammals. BMC Biol. 2022 Jun 9;20(1):133.

86. Lahn BT, Page DC. Functional coherence of the human Y chromosome. Science. 1997;278(5338):675–80.
87. Godfrey AK, Naqvi S, Chmátal L, Chick JM, Mitchell RN, Gygi SP, et al. Quantitative analysis of Y-Chromosome gene expression across 36 human tissues. Genome Res. 2020 Jun 1;30(6):860–73.
88. Pessia E, Engelstädter J, Marais GAB. The evolution of X chromosome inactivation in mammals: the demise of Ohno's hypothesis? Cell Mol Life Sci. 2014 Apr 1;71(8):1383–94.
89. Tukiainen T, Villani AC, Yen A, Rivas MA, Marshall JL, Satija R, et al. Landscape of X chromosome inactivation across human tissues. Nature. 2017 Oct;550(7675):244–8.
90. Jegalian K, Page DC. A proposed path by which genes common to mammalian X and Y chromosomes evolve to become X inactivated. Nature. 1998 Aug;394(6695):776–80.
91. Naqvi S, Bellott DW, Lin KS, Page DC. Conserved microRNA targeting reveals preexisting gene dosage sensitivities that shaped amniote sex chromosome evolution. Genome Res. 2018 Apr;28(4):474–83.
92. Cotton AM, Ge B, Light N, Adoue V, Pastinen T, Brown CJ. Analysis of expressed SNPs identifies variable extents of expression from the human inactive X chromosome. Genome Biol. 2013 Nov 1;14(11):R122.
93. Carrel L, Willard HF. X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature. 2005 Mar;434(7031):400–4.
94. Jacobs PA, Strong JA. A case of human intersexuality having a possible XXY sex-determining mechanism. Nature. 1959 Jan 31;183(4657):302–3.
95. Ford CE, Jones KW, Polani PE, De Almeida JC, Briggs JH. A sex-chromosome anomaly in a case of gonadal dysgenesis (Turner's syndrome). Lancet Lond Engl. 1959 Apr 4;1(7075):711–3.
96. Sinclair AH, Berta P, Palmer MS, Hawkins JR, Griffiths BL, Smith MJ, et al. A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. Nature. 1990 Jul 19;346(6281):240–4.
97. Koopman P, Münsterberg A, Capel B, Vivian N, Lovell-Badge R. Expression of a candidate sex-determining gene during mouse testis differentiation. Nature. 1990 Nov;348(6300):450–2.
98. Gubbay J, Collignon J, Koopman P, Capel B, Economou A, Münsterberg A, et al. A gene mapping to the sex-determining region of the mouse Y chromosome is a member of a novel family of embryonically expressed genes. Nature. 1990 Jul;346(6281):245–50.
99. Tucker PK, Lundrigan BL. The nature of gene evolution on the mammalian Y chromosome: lessons from Sry. Philos Trans R Soc Lond B Biol Sci. 1995 Nov 29;350(1333):221–7.
100. Miyawaki S, Kuroki S, Maeda R, Okashita N, Koopman P, Tachibana M. The mouse Sry locus harbors a cryptic exon that is essential for male sex determination. Science. 2020;370(6512):121–4.
101. Jeske YW, Mishina Y, Cohen DR, Behringer RR, Koopman P. Analysis of the role of Amh and Fra1 in the Sry regulatory pathway. Mol Reprod Dev. 1996 Jun;44(2):153–8.
102. Bullejos M, Koopman P. Spatially dynamic expression of Sry in mouse genital ridges. Dev Dyn Off Publ Am Assoc Anat. 2001 Jun;221(2):201–5.
103. Sekido R, Bar I, Narváez V, Penny G, Lovell-Badge R. SOX9 is up-regulated by the transient expression of SRY specifically in Sertoli cell precursors. Dev Biol. 2004 Oct 15;274(2):271–9.
104. Eggers S, Ohnesorg T, Sinclair A. Genetic regulation of mammalian gonad development. Nat Rev Endocrinol. 2014 Nov;10(11):673–83.
105. Li Y, Zheng M, Lau YFC. The sex-determining factors SRY and SOX9 regulate similar target genes and promote testis cord formation during testicular differentiation. Cell Rep. 2014 Aug 7;8(3):723–33.
106. Arango NA, Lovell-Badge R, Behringer RR. Targeted mutagenesis of the endogenous mouse Mis gene promoter: in vivo definition of genetic pathways of vertebrate sexual development. Cell. 1999 Nov 12;99(4):409–19.

107. Wilson MJ, Jeyasuria P, Parker KL, Koopman P. The Transcription Factors Steroidogenic Factor-1 and SOX9 Regulate Expression of Vanin-1 during Mouse Testis Development*. J Biol Chem. 2005 Feb 18;280(7):5917–23.
108. Bradford ST, Wilhelm D, Bandiera R, Vidal V, Schedl A, Koopman P. A cell-autonomous role for WT1 in regulating Sry in vivo. Hum Mol Genet. 2009 Sep 15;18(18):3429–38.
109. Kinoshita K, Shinka T, Sato Y, Kurahashi H, Kowa H, Chen G, et al. Expression analysis of a mouse orthologue of HSFY, a candidate for the azoospermic factor on the human Y chromosome. J Med Invest. 2006;53(1,2):117–22.
110. Shinka T, Sato Y, Chen G, Naroda T, Kinoshita K, Unemi Y, et al. Molecular characterization of heat shock-like factor encoded on the human Y chromosome, and implications for male infertility. Biol Reprod. 2004 Jul;71(1):297–306.
111. Sato Y, Yoshida K, Shinka T, Nozawa S, Nakahori Y, Iwamoto T. Altered expression pattern of heat shock transcription factor, Y chromosome (HSFY) may be related to altered differentiation of spermatogenic cells in testes with deteriorated spermatogenesis. Fertil Steril. 2006 Sep;86(3):612–8.
112. Vinci G, Raicu F, Popa L, Popa O, Cocos R, McElreavey K. A deletion of a novel heat shock gene on the Y chromosome associated with azoospermia. Mol Hum Reprod. 2005 Apr;11(4):295–8.
113. Hamilton CK, Verduzco-Gómez AR, Favetta LA, Blondin P, King WA. Testis-specific protein Y-encoded copy number is correlated to its expression and the field fertility of Canadian Holstein bulls. Sex Dev Genet Mol Biol Evol Endocrinol Embryol Pathol Sex Determ Differ. 2012;6(5):231–9.
114. Yue XP, Dechow C, Chang TC, DeJarnette JM, Marshall CE, Lei CZ, et al. Copy number variations of the extensively amplified Y-linked genes, HSFY and ZNF280BY, in cattle and their association with male reproductive traits in Holstein bulls. BMC Genomics. 2014 Feb 8;15(1):113.
115. Elliott DJ, Bourgeois CF, Klink A, Stévenin J, Cooke HJ. A mammalian germ cell-specific RNA-binding protein interacts with ubiquitously expressed proteins involved in splice site selection. Proc Natl Acad Sci U S A. 2000 May 23;97(11):5717–22.
116. Liu Y, Bourgeois CF, Pang S, Kudla M, Dreumont N, Kister L, et al. The Germ Cell Nuclear Proteins hnRNP G-T and RBMY Activate a Testis-Specific Exon. PLoS Genet. 2009 Nov 6;5(11):e1000707.
117. Dreumont N, Bourgeois CF, Lejeune F, Liu Y, Ehrmann IE, Elliott DJ, et al. Human RBMY regulates germline-specific splicing events by modulating the function of the serine/arginine-rich proteins 9G8 and Tra2-β. J Cell Sci. 2010 Jan 1;123(1):40–50.
118. Elliott DJ, Millar MR, Oghene K, Ross A, Kiesewetter F, Pryor J, et al. Expression of RBM in the nuclei of human germ cells is dependent on a critical region of the Y chromosome long arm. Proc Natl Acad Sci U S A. 1997 Apr 15;94(8):3848–53.
119. Elliott DJ. The role of potential splicing factors including RBMY, RBMX, hnRNPG-T and STAR proteins in spermatogenesis. Int J Androl. 2004 Dec;27(6):328–34.
120. Mahadevaiah SK, Odorisio T, Elliott DJ, Rattigan A, Szot M, Laval SH, et al. Mouse homologues of the human AZF candidate gene RBM are expressed in spermatogonia and spermatids, and map to a Y chromosome deletion interval associated with a high incidence of sperm abnormalities. Hum Mol Genet. 1998 Apr;7(4):715–27.
121. Tsuei DJ, Hsu HC, Lee PH, Jeng YM, Pu YS, Chen CN, et al. RBMY, a male germ cell-specific RNA-binding protein, activated in human liver cancers and transforms rodent fibroblasts. Oncogene. 2004;23(34):5815–22.
122. Kido T, Tabatabai ZL, Chen X, Lau YFC. Potential dual functional roles of the Y-linked RBMY in hepatocarcinogenesis. Cancer Sci. 2020 Aug;111(8):2987–99.
123. Gahadzikwa T. Investigating the potential role for RBMY in cancer. University of Kent (United Kingdom); 2021.
124. Tsuei DJ, Lee PH, Peng HY, Lu SL, Su DS, Jeng YM, et al. Male Germ Cell-Specific RNA Binding Protein RBMY: A New Oncogene Explaining Male Predominance in Liver Cancer. PLOS ONE. 2011 Nov 4;6(11):e26948.

125. Nickkholgh B, Noordam MJ, Hovingh SE, van Pelt AMM, van der Veen F, Repping S. Y chromosome TSPY copy numbers and semen quality. Fertil Steril. 2010 Oct 1;94(5):1744–7.
126. Murata C, Kuroki Y, Imoto I, Kuroiwa A. Ancestral Y-linked genes were maintained by translocation to the X and Y chromosomes fused to an autosomal pair in the Okinawa spiny rat Tokudaia muenninki. Chromosome Res. 2016 Sep 1;24(3):407–19.
127. Ozbun LL, You L, Kiang S, Angdisen J, Martinez A, Jakowlew SB. Identification of differentially expressed nucleolar TGF-beta1 target (DENTT) in human lung cancer cells that is a new member of the TSPY/SET/NAP-1 superfamily. Genomics. 2001 Apr 15;73(2):179–93.
128. Ayyanathan K, Lechner MS, Bell P, Maul GG, Schultz DC, Yamada Y, et al. Regulated recruitment of HP1 to a euchromatic gene induces mitotically heritable, epigenetic gene silencing: a mammalian cell culture model of gene variegation. Genes Dev. 2003 Aug 1;17(15):1855–69.
129. Vera J, Jaumot M, Estanyol JM, Brun S, Agell N, Bachs O. Heterogeneous nuclear ribonucleoprotein A2 is a SET-binding protein and a PP2A inhibitor. Oncogene. 2006;25(2):260–70.
130. Oram SW, Liu XX, Lee TL, Chan WY, Lau YFC. TSPY potentiates cell proliferation and tumorigenesis by promoting cell cycle progression in HeLa and NIH3T3 cells. BMC Cancer. 2006 Jun 9;6:154.
131. Muto S, Senda M, Akai Y, Sato L, Suzuki T, Nagai R, et al. Relationship between the structure of SET/TAF-Ibeta/INHAT and its histone chaperone activity. Proc Natl Acad Sci U S A. 2007 Mar 13;104(11):4285–90.
132. Lau YFC, Li Y, Kido T. Gonadoblastoma locus and the TSPY gene on the human Y chromosome. Birth Defects Res Part C Embryo Today Rev. 2009;87(1):114–22.
133. Gallagher WM, Bergin OE, Rafferty M, Kelly ZD, Nolan IM, Fox EJP, et al. Multiple markers for melanoma progression regulated by DNA methylation: insights from transcriptomic studies. Carcinogenesis. 2005 Nov;26(11):1856–67.
134. Yin YH, Li YY, Qiao H, Wang HC, Yang XA, Zhang HG, et al. TSPY is a cancer testis antigen expressed in human hepatocellular carcinoma. Br J Cancer. 2005 Aug;93(4):458–63.
135. Mitchell MJ, Woods DR, Wilcox SA, Marshall Graves JA, Bishop CE. Marsupial Y chromosome encodes a homologue of the mouse Y-linked candidate spermatogenesis gene Ube1y. Nature. 1992 Oct;359(6395):528–31.
136. Mitchell MJ, Woods DR, Tucker PK, Opp JS, Bishop CE. Homology of a candidate spermatogenic gene from the mouse Y chromosome to the ubiquitin-activating enzyme E1. Nature. 1991 Dec 12;354(6353):483–6.
137. Shpargel KB, Sengoku T, Yokoyama S, Magnuson T. UTX and UTY demonstrate histone demethylase-independent function in mouse embryonic development. PLOS Genet. 2012 Sep 27;8(9):e1002964.
138. Luddi A, Margollicci M, Gambera L, Serafini F, Cioni M, De Leo V, et al. Spermatogenesis in a Man with Complete Deletion of USP9Y. N Engl J Med. 2009 Feb 26;360(9):881–5.
139. Venkataramanan S, Gadek M, Calviello L, Wilkins K, Floor SN. DDX3X and DDX3Y are redundant in protein synthesis. RNA. 2021 Dec;27(12):1577–88.
140. Yamauchi Y, Riel JM, Ruthig VA, Ortega EA, Mitchell MJ, Ward MA. Two genes substitute for the mouse Y chromosome for spermatogenesis and reproduction. Science. 2016 Jan 29;351(6272):514–6.
141. Chen CY, Chan CH, Chen CM, Tsai YS, Tsai TY, Wu Lee YH, et al. Targeted inactivation of murine Ddx3x: essential roles of Ddx3x in placentation and embryogenesis. Hum Mol Genet. 2016 Jul 15;25(14):2905–22.
142. Ramathal C, Angulo B, Sukhwani M, Cui J, Durruthy-Durruthy J, Fang F, et al. DDX3Y gene rescue of a Y chromosome AZFa deletion restores germ cell formation and transcriptional programs. Sci Rep. 2015 Oct 12;5(1):15041.
143. Matsubara Y, Kato T, Kashimada K, Tanaka H, Zhi Z, Ichinose S, et al. TALEN-

Mediated Gene Disruption on Y Chromosome Reveals Critical Role of EIF2S3Y in Mouse Spermatogenesis. Stem Cells Dev. 2015 May 15;24(10):1164–70.
144. Link JC, Chen X, Arnold AP, Reue K. Metabolic impact of sex chromosomes. Adipocyte. 2013 Apr;2(2):74–9.
145. Skuse DH. Sexual dimorphism in cognition and behaviour: the role of X-linked genes. Eur J Endocrinol. 2006 Nov 1;155(suppl_1):S99–106.
146. Qi S, Al Mamun A, Ngwa C, Romana S, Ritzel R, Arnold AP, et al. X chromosome escapee genes are involved in ischemic sexual dimorphism through epigenetic modification of inflammatory signals. J Neuroinflammation. 2021 Mar 12;18(1):70.
147. Skuse DH. X-linked genes and mental functioning. Hum Mol Genet. 2005 Apr 15;14 Spec No 1:R27-32.
148. Dean R, Mank JE. The role of sex chromosomes in sexual dimorphism: discordance between molecular and phenotypic data. J Evol Biol. 2014;27(7):1443–53.
149. Sekido R. The potential role of SRY in epigenetic gene regulation during brain sexual differentiation in mammals. Adv Genet. 2014;86:135–65.
150. Heidecker B, Lamirault G, Kasper EK, Wittstein IS, Champion HC, Breton E, et al. The gene expression profile of patients with new-onset heart failure reveals important gender-specific differences†. Eur Heart J. 2010 May 1;31(10):1188–96.
151. Si H, Banga RS, Kapitsinou P, Ramaiah M, Lawrence J, Kambhampati G, et al. Human and Murine Kidneys Show Gender- and Species-Specific Gene Expression Differences in Response to Injury. PLOS ONE. 2009 Mar 11;4(3):e4802.
152. Lee J, Pinares-Garcia P, Loke H, Ham S, Vilain E, Harley VR. Sex-specific neuroprotection by inhibition of the Y-chromosome gene, SRY, in experimental Parkinson's disease. Proc Natl Acad Sci U S A. 2019 Aug 13;116(33):16577–82.
153. Dumanski JP, Lambert JC, Rasi C, Giedraitis V, Davies H, Grenier-Boley B, et al. Mosaic Loss of Chromosome Y in Blood Is Associated with Alzheimer Disease. Am J Hum Genet. 2016 Jun 2;98(6):1208–19.
154. Forsberg LA, Rasi C, Malmqvist N, Davies H, Pasupulati S, Pakalapati G, et al. Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. Nat Genet. 2014 Jun;46(6):624–8.
155. Qin N, Li N, Wang C, Pu Z, Ma Z, Jin G, et al. Association of Mosaic Loss of Chromosome Y with Lung Cancer Risk and Prognosis in a Chinese Population. J Thorac Oncol Off Publ Int Assoc Study Lung Cancer. 2019 Jan;14(1):37–44.
156. Kido T, Lau YFC. Roles of the Y chromosome genes in human cancers. Asian J Androl. 2015 Jun;17(3):373–80.
157. Wright DJ, Day FR, Kerrison ND, Zink F, Cardona A, Sulem P, et al. Genetic variants associated with mosaic Y chromosome loss highlight cell cycle genes and overlap with cancer susceptibility. Nat Genet. 2017 May;49(5):674–9.
158. Sano S, Horitani K, Ogawa H, Halvardson J, Chavkin NW, Wang Y, et al. Hematopoietic loss of Y chromosome leads to cardiac fibrosis and heart failure mortality. Science. 2022 Jul 15;377(6603):292–7.
159. Vakilian H, Mirzaei M, Sharifi Tabar M, Pooyan P, Habibi Rezaee L, Parker L, et al. DDX3Y, a Male-Specific Region of Y Chromosome Gene, May Modulate Neuronal Differentiation. J Proteome Res. 2015 Sep 4;14(9):3474–83.
160. Molina E, Chew GS, Myers SA, Clarence EM, Eales JM, Tomaszewski M, et al. A Novel Y-Specific Long Non-Coding RNA Associated with Cellular Lipid Accumulation in HepG2 cells and Atherosclerosis-related Genes. Sci Rep. 2017 Dec 1;7(1):16710.
161. Oliver SG, van der Aart QJM, Agostoni-Carbone ML, Aigle M, Alberghina L, Alexandraki D, et al. The complete DNA sequence of yeast chromosome III. Nature. 1992 May;357(6373):38–46.
162. Wilson R, Ainscough R, Anderson K, Baynes C, Berks M, Bonfield J, et al. 2.2 Mb of contiguous nucleotide sequence from chromosome III of C. elegans. Nature. 1994 Mar;368(6466):32–8.
163. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001 Feb;409(6822):860–921.

164. Chinwalla AT, Cook LL, Delehaunty KD, Fewell GA, Fulton LA, Fulton RS, et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002 Dec;420(6915):520–62.
165. Metzker ML. Emerging technologies in DNA sequencing. Genome Res. 2005 Dec 1;15(12):1767–76.
166. Mardis ER. A decade's perspective on DNA sequencing technology. Nature. 2011 Feb;470(7333):198–203.
167. Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. Nat Methods. 2011 Jan;8(1):61–5.
168. Zimin AV, Salzberg SL. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. PLOS Comput Biol. 2020 Jun 26;16(6):e1007981.
169. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-Time DNA Sequencing from Single Polymerase Molecules. Science. 2009 Jan 2;323(5910):133–8.
170. Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol. 2009 Apr;4(4):265–70.
171. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 2015 Oct;13(5):278–89.
172. Ardui S, Ameur A, Vermeesch JR, Hestand MS. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Res. 2018 Mar 16;46(5):2159–68.
173. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol. 2021 Nov;39(11):1348–65.
174. Schwartz DC, Li X, Hernandez LI, Ramnarain SP, Huff EJ, Wang YK. Ordered Restriction Maps of Saccharomyces cerevisiae Chromosomes Constructed by Optical Mapping. Science. 1993 Oct;262(5130):110–4.
175. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol. 2013 Dec;31(12):1119–25.
176. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017 May;27(5):722–36.
177. Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods. 2016 Dec;13(12):1050–4.
178. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021 Feb;18(2):170–5.
179. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019 May;37(5):540–6.
180. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Nat Methods. 2020 Feb;17(2):155–8.
181. Stein L. Genome annotation: from sequence to biology. Nat Rev Genet. 2001 Jul;2(7):493–503.
182. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016 Jan 4;44(D1):D733–45.
183. Frankish A, Uszczynska B, Ritchie GRS, Gonzalez JM, Pervouchine D, Petryszak R, et al. Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. BMC Genomics. 2015;16 Suppl 8:S2.
184. Guigó R, Knudsen S, Drake N, Smith T. Prediction of gene structure. J Mol Biol. 1992 Jul 5;226(1):141–57.
185. Borodovsky M, McIninch J. GENMARK: Parallel gene recognition for both DNA strands. Comput Chem. 1993 Jun 1;17(2):123–33.
186. Solovyev VV, Salamov AA, Lawrence CB. Predicting internal exons by

oligonucleotide composition and discriminant analysis of spliceable open reading frames. Nucleic Acids Res. 1994 Dec 11;22(24):5156–63.
187. Birney E, Durbin R. Using GeneWise in the Drosophila annotation experiment. Genome Res. 2000 Apr;10(4):547–8.
188. Gelfand MS, Mironov AA, Pevzner PA. Gene recognition via spliced sequence alignment. Proc Natl Acad Sci U S A. 1996 Aug 20;93(17):9061–6.
189. Yeh RF, Lim LP, Burge CB. Computational inference of homologous gene structures in the human genome. Genome Res. 2001 May;11(5):803–16.
190. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. Bioinforma Oxf Engl. 2003 Oct;19 Suppl 2:ii215-225.
191. Korf I, Flicek P, Duan D, Brent MR. Integrating genomic homology into gene structure prediction. Bioinforma Oxf Engl. 2001;17 Suppl 1:S140-148.
192. Parra G, Agarwal P, Abril JF, Wiehe T, Fickett JW, Guigó R. Comparative gene prediction in human and mouse. Genome Res. 2003 Jan;13(1):108–17.
193. Meyer IM, Durbin R. Comparative ab initio prediction of gene structures using pair HMMs. Bioinforma Oxf Engl. 2002 Oct;18(10):1309–18.
194. Korf I. Genomics: the state of the art in RNA-seq analysis. Nat Methods. 2013 Dec;10(12):1165–6.
195. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008 Jan;18(1):188–96.
196. Thierry-Mieg D, Thierry-Mieg J. AceView: a comprehensive cDNA-supported gene and transcripts annotation. Genome Biol. 2006;7 Suppl 1:S12.1-14.
197. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 2008;9(1):R7.
198. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011 Dec 22;12:491.
199. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinforma Oxf Engl. 2016 Mar 1;32(5):767–9.
200. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010 May;28(5):511–5.
201. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015 Mar;33(3):290–5.
202. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat Biotechnol. 2011;29(7):644–52.
203. Martin J, Bruno VM, Fang Z, Meng X, Blow M, Zhang T, et al. Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. BMC Genomics. 2010 Nov 24;11:663.
204. Wang L, Park HJ, Dasari S, Wang S, Kocher JP, Li W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res. 2013 Apr 1;41(6):e74.
205. Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W345-349.
206. Wucher V, Legeai F, Hédan B, Rizk G, Lagoutte L, Leeb T, et al. FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. Nucleic Acids Res. 2017 May 5;45(8):e57.
207. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for

reference generation and analysis. Nat Protoc. 2013 Aug;8(8):1494–512.
208. Dunham I, Shimizu N, Roe BA, Chissoe S, Hunt AR, Collins JE, et al. The DNA sequence of human chromosome 22. Nature. 1999 Dec 2;402(6761):489–95.
209. Kuo RI, Tseng E, Eory L, Paton IR, Archibald AL, Burt DW. Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. BMC Genomics. 2017 Apr 24;18(1):323.
210. Mudge JM, Harrow J. The state of play in higher eukaryote gene annotation. Nat Rev Genet. 2016 Dec;17(12):758–72.
211. Morillon A, Gautheret D. Bridging the gap between reference and real transcriptomes. Genome Biol. 2019 Jun 3;20(1):112.
212. Beiki H, Liu H, Huang J, Manchanda N, Nonneman D, Smith TPL, et al. Improved annotation of the domestic pig genome through integration of Iso-Seq and RNA-seq data. BMC Genomics. 2019 May 7;20(1):344.
213. Abdel-Ghany SE, Hamilton M, Jacobi JL, Ngam P, Devitt N, Schilkey F, et al. A survey of the sorghum transcriptome using single-molecule long reads. Nat Commun. 2016 Jun 24;7:11706.
214. Minoche AE, Dohm JC, Schneider J, Holtgräwe D, Viehöver P, Montfort M, et al. Exploiting single-molecule transcript sequencing for eukaryotic gene prediction. Genome Biol. 2015 Sep 2;16:184.
215. Nouhaud P. Long-read based assembly and annotation of a Drosophila simulans genome. bioRxiv. 2018;425710.
216. Fiddes IT, Armstrong J, Diekhans M, Nachtweide S, Kronenberg ZN, Underwood JG, et al. Comparative Annotation Toolkit (CAT)-simultaneous clade and personal genome annotation. Genome Res. 2018 Jul;28(7):1029–38.
217. Cook DE, Valle-Inclan JE, Pajoro A, Rovenich H, Thomma BPHJ, Faino L. Long-Read Annotation: Automated Eukaryotic Genome Annotation Based on Long-Read cDNA Sequencing. Plant Physiol. 2019 Jan;179(1):38–54.
218. Nachtweide S, Stanke M. Multi-Genome Annotation with AUGUSTUS. Methods Mol Biol Clifton NJ. 2019;1962:139–60.
219. Hoff KJ, Stanke M. Predicting Genes in Single Genomes with AUGUSTUS. Curr Protoc Bioinforma. 2019 Mar;65(1):e57.
220. Kuo RI, Cheng Y, Zhang R, Brown JWS, Smith J, Archibald AL, et al. Illuminating the dark side of the human transcriptome with long read transcript sequencing. BMC Genomics. 2020 Oct 30;21(1):751.
221. Hughes JF, Rozen S. Genomics and genetics of human and primate y chromosomes. Annu Rev Genomics Hum Genet. 2012;13:83–108.
222. Kuderna LFK, Lizano E, Julià E, Gomez-Garrido J, Serres-Armero A, Kuhlwilm M, et al. Selective single molecule sequencing and assembly of a human Y chromosome of African origin. Nat Commun. 2019 Jan 2;10(1):4.
223. Yano Y, Chiba T, Asahara H. Analysis of the mouse Y chromosome by single-molecule sequencing with Y chromosome enrichment. Front Genet. 2020;11:406.
224. Jevit MJ, Davis BW, Castaneda C, Hillhouse A, Juras R, Trifonov VA, et al. An 8.22 Mb Assembly and Annotation of the Alpaca (Vicugna pacos) Y Chromosome. Genes. 2021 Jan;12(1):105.
225. Cechova M, Vegesna R, Tomaszkiewicz M, Harris RS, Chen D, Rangavittal S, et al. Dynamic evolution of great ape Y chromosomes. Proc Natl Acad Sci U S A. 2020 Oct 20;117(42):26273–80.
226. Xiao C, Li J, Xie T, Chen J, Zhang S, Elaksher SH, et al. The assembly of caprine Y chromosome sequence reveals a unique paternal phylogenetic pattern and improves our understanding of the origin of domestic goat. Ecol Evol. 2021;11(12):7779–95.
227. Li R, Yang P, Li M, Fang W, Yue X, Nanaei HA, et al. A Hu sheep genome with the first ovine Y chromosome reveal introgression history after sheep domestication. Sci China Life Sci. 2021 Jul;64(7):1116–30.
228. Rando HM, Wadlington WH, Johnson JL, Stutchman JT, Trut LN, Farré M, et al. The Red Fox Y-Chromosome in Comparative Context. Genes. 2019 May 28;10(6):409.

229. Rittié L, Fisher GJ. Isolation and Culture of Skin Fibroblasts. In: Varga J, Brenner DA, Phan SH, editors. Fibrosis Research: Methods and Protocols [Internet]. Totowa, NJ: Humana Press; 2005 [cited 2023 Jul 28]. p. 83–98. (Methods in Molecular Medicine). Available from: https://doi.org/10.1385/1-59259-940-0:083
230. Takahashi H, Kato S, Murata M, Carninci P. CAGE (cap analysis of gene expression): a protocol for the detection of promoter and transcriptional networks. Gene Regul Netw Methods Protoc. 2012;181–200.
231. Gordon A, Hannon G, others. Fastx-toolkit. FASTQA Short-Reads Preprocessing Tools Unpubl Httphannonlab Cshl Edufastxtoolkit. 2010;5.
232. Lassmann T. TagDust2: a generic method to extract reads from sequencing data. BMC Bioinformatics. 2015 Jan 28;16(1):24.
233. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Apr;9(4):357–9.
234. Haberle V, Forrest ARR, Hayashizaki Y, Carninci P, Lenhard B. CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. Nucleic Acids Res. 2015 Apr 30;43(8):e51.
235. Thodberg M, Thieffry A, Vitting-Seerup K, Andersson R, Sandelin A. CAGEfightR: analysis of 5′-end data using R/Bioconductor. BMC Bioinformatics. 2019;20(1):1–13.
236. Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, et al. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. Genome Res. 2017 May;27(5):768–77.
237. Vasimuddin Md, Misra S, Li H, Aluru S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. In: 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS). 2019. p. 314–24.
238. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
239. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc. 2016 Sep;11(9):1650–67.
240. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15–21.
241. Zhang Z, Schwartz S, Wagner L, Miller W. A Greedy Algorithm for Aligning DNA Sequences. J Comput Biol. 2000 Feb;7(1–2):203–14.
242. Delcher AL. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res. 2002 Jun 1;30(11):2478–83.
243. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 2016 Jul 1;33(7):1870–4.
244. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021 Feb 1;10(2):giab008.
245. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011 Aug 1;27(15):2156–8.
246. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007 Sep 1;81(3):559–75.
247. Chen S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. iMeta. 2023;2(2):e107.
248. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841–2.
249. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep 1;20(9):1297–303.
250. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods. 2018 Aug;15(8):591–4.
251. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for

annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w ^{1118} ; iso-2; iso-3. Fly (Austin). 2012 Apr;6(2):80–92.
252. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 2019 Dec 16;20(1):275.
253. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 2020 Sep 14;21(1):245.
254. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011 Jun;21(6):974–84.
255. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinforma Oxf Engl. 2014 Apr 1;30(7):923–30.
256. Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics. 2018 Mar 1;34(5):867–8.
257. Krzywinski M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D, et al. Circos: An information aesthetic for comparative genomics. Genome Res. 2009 Sep;19(9):1639–45.
258. Leung AKY, Jin N, Yip KY, Chan TF. OMTools: a software package for visualizing and processing optical mapping data. Bioinformatics. 2017 Sep 15;33(18):2933–5.
259. Wolff J, Bhardwaj V, Nothjunge S, Richard G, Renschler G, Gilsbach R, et al. Galaxy HiCExplorer: a web server for reproducible Hi-C data analysis, quality control and visualization. Nucleic Acids Res. 2018 Jul 2;46(W1):W11–6.
260. Gotoh O. A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. Nucleic Acids Res. 2008 May;36(8):2630–8.
261. Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. KaKs_Calculator 2.0: A Toolkit Incorporating Gamma-Series Methods and Sliding Window Strategies. Genomics Proteomics Bioinformatics. 2010 Mar 1;8(1):77–80.
262. Pei J, Grishin NV. AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinforma Oxf Engl. 2001 Aug;17(8):700–12.
263. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014 May 1;30(9):1312–3.
264. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLOS Comput Biol. 2019 Apr 8;15(4):e1006650.
265. Ahmed IA. IX.—Cytological Analysis of Chromosome Behaviour in Three Breeds of Dogs. Proc R Soc Edinb Sect B Biol Sci. 1941 ed;61(1):107–18.
266. Świtoński M, Reimann N, Bosma AA, Long S, Bartnitzke S, Pieńkowska A, et al. Report on the progress of standardization of the G-banded canine (Canis familiaris) karyotype. Chromosome Res. 1996 Jun;4(4):306–9.
267. Selden JR, Moorhead PS, Oehlert ML, Patterson DF. The Giemsa banding pattern of the canine karyotype. Cytogenet Genome Res. 1975;15(6):380–7.
268. Pathak S, Tuinen PV, Merry DE. Heterochromatin, synaptonemal complex, and NOR activity in the somatic and germ cells of a male domestic dog, Canis familiaris(Mammalia, Canidae). Cytogenet Genome Res. 1982;34(1–2):112–8.
269. Mayr B, Geber G, Auer H, Kalat M, Schleger W. Heterochromatin composition and nucleolus organizer activity in four canid species. Can J Genet Cytol. 1986;28(5):744–53.
270. Mäkinen A, Zijlstra C, Haan NAD, Mellink CHM, Bosma AA. Localization of 18S+28S and 5S ribosomal RNA genes in the dog by fluorescence in situ hybridization. Cytogenet Genome Res. 1997 Jan 1;78(3–4):231–5.
271. Goodpasture C, Bloom SE. Visualization of nucleolar organizer regions im mammalian chromosomes using silver staining. Chromosoma. 1975 Nov 20;53(1):37–

50.
272. Young AC, Kirkness EF, Breen M. Tackling the characterization of canine chromosomal breakpoints with an integrated in-situ/in-silico approach: The canine PAR and PAB. Chromosome Res. 2008 Dec;16(8):1193–202.
273. Fletcher S, Darragh D, Fan Y, Grounds MD, Fisher CJ, Beilharz MW. Specific cloning of DNA fragments unique to the dog Y chromosome. Genet Anal Biomol Eng. 1993 Jun 1;10(3):77–83.
274. Meyers-Wallen VN, Palmer VL, Acland GM, Hershfield B. Sry-negative XX sex reversal in the American cocker spaniel dog. Mol Reprod Dev. 1995;41(3):300–5.
275. Olivier M, Lust G. Two DNA sequences specific for the canine Y chromosome. Anim Genet. 1998 Apr;29(2):146–9.
276. Olivier M, Breen M, Binns MM, Lust G. Localization and characterization of nucleotide sequences from the canine Y chromosome. Chromosome Res Int J Mol Supramol Evol Asp Chromosome Biol. 1999;7(3):223–33.
277. Breen M, Jouquand S, Renier C, Mellersh CS, Hitte C, Holmes NG, et al. Chromosome-Specific Single-Locus FISH Probes Allow Anchorage of an 1800-Marker Integrated Radiation-Hybrid/Linkage Map of the Domestic Dog Genome to All Chromosomes. Genome Res. 2001 Oct 1;11(10):1784–95.
278. Guyon R, Lorentzen TD, Hitte C, Kim L, Cadieu E, Parker HG, et al. A 1-Mb resolution radiation hybrid map of the canine genome. Proc Natl Acad Sci U S A. 2003 Apr 29;100(9):5296–301.
279. Edwards RJ, Field MA, Ferguson JM, Dudchenko O, Keilwagen J, Rosen BD, et al. Chromosome-length genome assembly and structural variations of the primal Basenji dog (Canis lupus familiaris) genome. BMC Genomics. 2021 Mar 16;22(1):188.
280. Smeds L, Kojola I, Ellegren H. The evolutionary history of grey wolf Y chromosomes. Mol Ecol. 2019;28(9):2173–91.
281. Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res. 2016 Nov 2;44(19):e147.
282. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLOS ONE. 2014 Nov 19;9(11):e112963.
283. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinforma Oxf Engl. 2011 Nov 1;27(21):2987–93.
284. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6:11.
285. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004 Jan 30;5(2):R12.
286. Abdennur N, Mirny LA. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. Bioinformatics. 2020 Jan 1;36(1):311–6.
287. Ciccodicola A, D'Esposito M, Esposito T, Gianfrancesco F, Migliaccio C, Miano MG, et al. Differentially regulated and evolved genes in the fully sequenced Xq/Yq pseudoautosomal region. Hum Mol Genet. 2000 Feb 12;9(3):395–401.
288. Smeds L, Kawakami T, Burri R, Bolivar P, Husby A, Qvarnström A, et al. Genomic identification and characterization of the pseudoautosomal region in highly differentiated avian sex chromosomes. Nat Commun. 2014 Nov 7;5:5448.
289. Charlesworth B. The evolution of sex chromosomes. Science. 1991;251(4997):1030–3.
290. Tang W, Mun S, Joshi A, Han K, Liang P. Mobile elements contribute to the uniqueness of human genome with 15,000 human-specific insertions and 14 Mbp sequence increase. DNA Res. 2018 Oct 1;25(5):521–33.
291. Chen S, Zhang G, Shao C, Huang Q, Liu G, Zhang P, et al. Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and

adaptation to a benthic lifestyle. Nat Genet. 2014 Mar;46(3):253–60.
292. Chalopin D, Volff JN, Galiana D, Anderson JL, Schartl M. Transposable elements and early evolution of sex chromosomes in fish. Chromosome Res. 2015 Sep 1;23(3):545–60.
293. Mawaribuchi S, Takahashi S, Wada M, Uno Y, Matsuda Y, Kondo M, et al. Sex chromosome differentiation and the W-and Z-specific loci in Xenopus laevis. Dev Biol. 2017;426(2):393–400.
294. Hua-Van A, Rouzic AL, Maisonhaute C, Capy P. Abundance, distribution and dynamics of retrotransposable elements and transposons: similarities and differences. Cytogenet Genome Res. 2005;110(1–4):426–40.
295. Bachtrog D, Hom E, Wong KM, Maside X, de Jong P. Genomic degradation of a young Y chromosome in Drosophila miranda. Genome Biol. 2008 Feb 12;9(2):R30.
296. Vogt PH. Report of the Third International Workshop on Human Y Chromosome Mapping 1997. Cytogenet Genome Res. 1997;79(1–2):1–20.
297. Fan H, Hu Y, Shan L, Yu L, Wang B, Li M, et al. Synteny search identifies carnivore Y chromosome for evolution of male specific genes. Integr Zool. 2019;14(3):224–34.
298. Krzeminska P, Nowacka-Woszuk J, Switonski M. Copy number variation of the SRY gene showed an association with disorders of sex development in Yorkshire Terrier dogs. Anim Genet. 2022;53(1):152–5.
299. King V, Goodfellow PN, Wilkerson AJP, Johnson WE, O'Brien SJ, Pecon-Slattery J. Evolution of the Male-Determining Gene SRY Within the Cat Family Felidae. Genetics. 2007 Apr 1;175(4):1855–67.
300. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24–6.
301. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 34(17):i884–90.
302. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan;29(1):15–21.
303. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011 Aug 4;12(1):323.
304. Liao BY, Scott NM, Zhang J. Impacts of Gene Essentiality, Expression Pattern, and Gene Compactness on the Evolutionary Rate of Mammalian Proteins. Mol Biol Evol. 2006 Nov 1;23(11):2072–80.
305. Nyakatura K, Bininda-Emonds OR. Updating the evolutionary history of Carnivora (Mammalia): a new species-level supertree complete with divergence time estimates. BMC Biol. 2012 Feb 27;10(1):12.
306. Wang DP, Wan HL, Zhang S, Yu J. γ-MYN: a new algorithm for estimating Ka and Ks with consideration of variable substitution rates. Biol Direct. 2009 Jun 16;4(1):20.
307. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Curr Protoc Bioinforma. 2013;43(1):11.10.1-11.10.33.
308. Yang Z, Bielawski JP. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 2000 Dec 1;15(12):496–503.
309. Bellott DW, Skaletsky H, Pyntikova T, Mardis ER, Graves T, Kremitzki C, et al. Convergent Evolution of Chicken Z and Human X Chromosomes by Expansion and Gene Acquisition. Nature. 2010 Jul 29;466(7306):612–6.
310. Graves JAM. Sex Chromosome Specialization and Degeneration in Mammals. Cell. 2006 Mar 10;124(5):901–14.
311. Lai MC, Chang WC, Shieh SY, Tarn WY. DDX3 Regulates Cell Growth through Translational Control of Cyclin E1. Mol Cell Biol. 2010 Nov 1;30(22):5444–53.
312. Walport LJ, Hopkinson RJ, Vollmar M, Madden SK, Gileadi C, Oppermann U, et al. Human UTY(KDM6C) Is a Male-specific Nε-Methyl Lysyl Demethylase*. J Biol Chem. 2014 Jun 1;289(26):18302–13.
313. Huang Y, Baker RT, Fischer-Vize JA. Control of Cell Fate by a Deubiquitinating Enzyme Encoded by the fat facets Gene. Science. 1995 Dec 15;270(5243):1828–31.

314. Swanson WJ, Vacquier VD. The rapid evolution of reproductive proteins. Nat Rev Genet. 2002 Feb;3(2):137–44.
315. Raudsepp T, Chowdhary BP. The Eutherian Pseudoautosomal Region. Cytogenet Genome Res. 2015;147(2–3):81–94.
316. Ellis NA, Ye TZ, Patton S, German J, Goodfellow PN, Weller P. Cloning of PBDX, an MIC2-related gene that spans the pseudoautosomal boundary on chromosome Xp. Nat Genet. 1994 Apr;6(4):394–400.
317. Weller PA, Critcher R, Goodfellow PN, German J, Ellis NA. The human Y chromosome homologue of XG: transcription of a naturally truncated gene. Hum Mol Genet. 1995 May;4(5):859–68.
318. Galtier N. Recombination, GC-content and the human pseudoautosomal boundary paradox. Trends Genet TIG. 2004 Aug;20(8):347–9.
319. Perry J, Palmer S, Gabriel A, Ashworth A. A Short Pseudoautosomal Region in Laboratory Mice. Genome Res. 2001 Nov 1;11(11):1826–32.
320. Palmer S, Perry J, Kipling D, Ashworth A. A gene spans the pseudoautosomal boundary in mice. Proc Natl Acad Sci. 1997;94(22):12030–5.
321. Van Laere AS, Coppieters W, Georges M. Characterization of the bovine pseudoautosomal boundary: Documenting the evolutionary history of mammalian sex chromosomes. Genome Res. 2008 Dec;18(12):1884–95.
322. Raudsepp T, Chowdhary BP. The horse pseudoautosomal region (PAR): characterization and comparison with the human, chimp and mouse PARs. Cytogenet Genome Res. 2008;121(2):102–9.
323. Das PJ, Mishra DK, Ghosh S, Avila F, Johnson GA, Chowdhary BP, et al. Comparative Organization and Gene Expression Profiles of the Porcine Pseudoautosomal Region. Cytogenet Genome Res. 2013;141(1):26–36.
324. Morgan AP, Bell TA, Crowley JJ, Pardo-Manuel de Villena F. Instability of the Pseudoautosomal Boundary in House Mice. Genetics. 2019 Jun 1;212(2):469–87.
325. Shearn R, Wright AE, Mousset S, Régis C, Penel S, Lemaitre JF, et al. Evolutionary stasis of the pseudoautosomal boundary in strepsirrhine primates. Elife. 2020;9:e63650.
326. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.
327. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 2021 Jul 2;49(W1):W293–6.
328. Raudsepp T, Das PJ, Avila F, Chowdhary BP. The Pseudoautosomal Region and Sex Chromosome Aneuploidies in Domestic Species. Sex Dev. 2012;6(1–3):72–83.
329. Peng C, Niu L, Deng J, Yu J, Zhang X, Zhou C, et al. Can-SINE dynamics in the giant panda and three other Caniformia genomes. Mob DNA. 2018;9(1):1–14.
330. Halo JV, Pendleton AL, Shen F, Doucet AJ, Derrien T, Hitte C, et al. Long-read assembly of a Great Dane genome highlights the contribution of GC-rich sequence and mobile elements to canine genomes. Proc Natl Acad Sci. 2021 Mar 16;118(11):e2016274118.
331. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, et al. Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet. 2006 Jun;38(6):626–35.
332. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science. 2015 Jan 23;347(6220):1260419.
333. Baldarelli RM, Smith CM, Finger JH, Hayamizu TF, McCright IJ, Xu J, et al. The mouse Gene Expression Database (GXD): 2021 update. Nucleic Acids Res. 2021 Jan 8;49(D1):D924–31.
334. Otto SP, Pannell JR, Peichel CL, Ashman TL, Charlesworth D, Chippindale AK, et al. About PAR: The distinct evolutionary dynamics of the pseudoautosomal region. Trends Genet. 2011 Sep 1;27(9):358–67.
335. Kent TV, Uzunović J, Wright SI. Coevolution between transposable elements and recombination. Philos Trans R Soc B Biol Sci. 2017 Nov 6;372(1736):20160458.
336. Xu L, Auer G, Peona V, Suh A, Deng Y, Feng S, et al. Dynamic evolutionary history

and gene content of sex chromosomes across diverse songbirds. Nat Ecol Evol. 2019;3(5):834–44.
337. Bourgeois Y, Boissinot S. On the Population Dynamics of Junk: A Review on the Population Genomics of Transposable Elements. Genes. 2019 Jun;10(6):419.
338. Boissinot S, Davis J, Entezam A, Petrov D, Furano AV. Fitness cost of LINE-1 (L1) activity in humans. Proc Natl Acad Sci. 2006 Jun 20;103(25):9590–4.
339. Xue AT, Ruggiero RP, Hickerson MJ, Boissinot S. Differential Effect of Selection against LINE Retrotransposons among Vertebrates Inferred from Whole-Genome Data and Demographic Modeling. Genome Biol Evol. 2018 May 1;10(5):1265–81.
340. Marais G, Galtier N. Sex chromosomes: how X-Y recombination stops. Curr Biol. 2003 Aug 19;13(16):R641–3.
341. Iwase M, Satta Y, Hirai Y, Hirai H, Imai H, Takahata N. The amelogenin loci span an ancient pseudoautosomal boundary in diverse mammalian species. Proc Natl Acad Sci. 2003 Apr 29;100(9):5258–63.
342. Vicoso B. Molecular and evolutionary dynamics of animal sex-chromosome turnover. Nat Ecol Evol. 2019 Dec;3(12):1632–41.
343. Larson EL, Kopania EEK, Good JM. Spermatogenesis and the Evolution of Mammalian Sex Chromosomes. Trends Genet. 2018 Sep 1;34(9):722–32.
344. Holmlund H, Yamauchi Y, Durango G, Fujii W, Ward MA. Two acquired mouse Y chromosome-linked genes, Prssly and Teyorf1, are dispensable for male fertility‡. Biol Reprod. 2022 Sep 1;107(3):752–64.
345. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015 Jan;12(1):59–60.
346. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. Genome Res. 2002 Jun 1;12(6):996–1006.
347. Jung M, Wells D, Rusch J, Ahmad S, Marchini J, Myers SR, et al. Unified single-cell analysis of testis gene regulation and pathology in five mouse strains. Bourc'his D, Wittkopp PJ, Lukassen S, editors. eLife. 2019 Jun 25;8:e43966.
348. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018 May;36(5):411–20.
349. Laurentino S, Heckmann L, Di Persio S, Li X, Meyer zu Hörste G, Wistuba J, et al. High-resolution analysis of germ cells from men with sex chromosomal aneuploidies reveals normal transcriptome but impaired imprinting. Clin Epigenetics. 2019 Aug 28;11(1):127.
350. Guo J, Grow EJ, Mlcochova H, Maher GJ, Lindskog C, Nie X, et al. The adult human testis transcriptional cell atlas. Cell Res. 2018 Dec;28(12):1141–57.
351. Clark EL, Bush SJ, McCulloch MEB, Farquhar IL, Young R, Lefevre L, et al. A high resolution atlas of gene expression in the domestic sheep (Ovis aries). PLOS Genet. 2017 Sep 15;13(9):e1006997.
352. Choi HMT, Schwarzkopf M, Fornace ME, Acharya A, Artavanis G, Stegmaier J, et al. Third-generation in situ hybridization chain reaction: multiplexed, quantitative, sensitive, versatile, robust. Development. 2018 Jun 26;145(12):dev165753.
353. Pitt JJ. Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry. Clin Biochem Rev. 2009 Feb;30(1):19–34.
354. Giansanti P, Samaras P, Bian Y, Meng C, Coluccio A, Frejno M, et al. Mass spectrometry-based draft of the mouse proteome. Nat Methods. 2022 Jul;19(7):803–11.
355. Tyanova S, Temu T, Cox J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nat Protoc. 2016 Dec;11(12):2301–19.
356. Xia K, He S, Luo P, Dong L, Gao F, Chen X, et al. Transcriptomic landscape and potential therapeutic targets for human testicular aging revealed by single-cell RNA sequencing. bioRxiv. 2022;2022.12. 11.519976.
357. Guo J, Nie X, Giebler M, Mlcochova H, Wang Y, Grow EJ, et al. The Dynamic Transcriptional Cell Atlas of Testis Development during Human Puberty. Cell Stem Cell. 2020 Feb 6;26(2):262-276.e4.

358. Nie X, Munyoki SK, Sukhwani M, Schmid N, Missel A, Emery BR, et al. Single-cell analysis of human testis aging and correlation with elevated body mass index. Dev Cell. 2022 May 9;57(9):1160-1176.e5.
359. Blum M, Chang HY, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, et al. The InterPro protein families and domains database: 20 years on. Nucleic Acids Res. 2021 Jan 8;49(D1):D344–54.
360. Peng Y, Li H, Liu Z, Zhang C, Li K, Gong Y, et al. Chromosome-level genome assembly of the Arctic fox (Vulpes lagopus) using PacBio sequencing and Hi-C technology. Mol Ecol Resour. 2021;21(6):2093–108.
361. Harder AM, Walden KKO, Marra NJ, Willoughby JR. High-Quality Reference Genome for an Arid-Adapted Mammal, the Banner-Tailed Kangaroo Rat (Dipodomys spectabilis). Genome Biol Evol. 2022 Jan 1;14(1):evac005.
362. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018 Apr;36(4):338–45.
363. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376(6588):44–53.
364. Belser C, Baurens FC, Noel B, Martin G, Cruaud C, Istace B, et al. Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing. Commun Biol. 2021 Sep 7;4(1):1–12.
365. Payne A, Holmes N, Clarke T, Munro R, Debebe BJ, Loose M. Readfish enables targeted nanopore sequencing of gigabase-sized genomes. Nat Biotechnol. 2021 Apr;39(4):442–50.
366. Ni Y, Liu X, Simeneh ZM, Yang M, Li R. Benchmarking of Nanopore R10.4 and R9.4.1 flow cells in single-cell whole-genome amplification and whole-genome shotgun sequencing. Comput Struct Biotechnol J. 2023 Jan 1;21:2352–64.
367. Takahashi K, Yamanaka S. A decade of transcription factor-mediated reprogramming to pluripotency. Nat Rev Mol Cell Biol. 2016 Mar;17(3):183–93.
368. Lee CZW, Kozaki T, Ginhoux F. Studying tissue macrophages in vitro: are iPSC-derived cells the answer? Nat Rev Immunol. 2018 Nov;18(11):716–25.
369. Wang GD, Shao XJ, Bai B, Wang J, Wang X, Cao X, et al. Structural variation during dog domestication: insights from gray wolf and dhole genomes. Natl Sci Rev. 2019 Jan 1;6(1):110–22.
370. Acosta MJ, Marchal JA, Romero-Fernández I, Megías-Nogales B, Modi WS, Sánchez A. Sequence Analysis and Mapping of the Sry Gene in Species of the Subfamily Arvicolinae (Rodentia). Sex Dev. 2010;4(6):336–47.
371. Nagamine CM. The testis-determining gene, SRY, exists in multiple copies in Old World rodents. Genet Res. 1994 Dec;64(3):151–9.
372. Lundrigan BL, Tucker PK. Evidence for multiple functional copies of the male sex-determining locus, sry, in African murine rodents. J Mol Evol. 1997 Jul 1;45(1):60–5.
373. Kurtz S, Lucas-Hahn A, Schlegelberger B, Göhring G, Niemann H, Mettenleiter TC, et al. Knockout of the HMG domain of the porcine SRY gene causes sex reversal in gene-edited pigs. Proc Natl Acad Sci. 2021 Jan 12;118(2):e2008743118.
374. Mukherjee A, Dass G, G JM, Gohain M, Brahma B, Datta TK, et al. Absolute copy number differences of Y chromosomal genes between crossbred (Bos taurus × Bos indicus) and Indicine bulls. J Anim Sci Biotechnol. 2013 Apr 4;4(1):15.
375. Sun T, Hanif Q, Chen H, Lei C, Dang R. Copy Number Variations of Four Y-Linked Genes in Swamp Buffaloes. Animals. 2020 Jan;10(1):31.
376. Han H, Zhao X, Xia X, Chen H, Lei C, Dang R. Copy number variations of five Y chromosome genes in donkeys. Arch Anim Breed. 2017 Oct 18;60(4):391–7.
377. Han H, Zhang X, Zhao X, Xia X, Lei C, Dang R. Eight Y chromosome genes show copy number variations in horses. Arch Anim Breed. 2018 Jul 2;61(3):263–70.
378. Stachowiak M, Szczerbal I, Nowacka-Woszuk J, Nowak T, Sowinska N, Lukomska A, et al. Cytogenetic and molecular insight into the genetic background of disorders of sex development in seventeen cats. Sci Rep. 2022 Oct 24;12(1):17807.

379. Geraldes A, Rambo T, Wing RA, Ferrand N, Nachman MW. Extensive Gene Conversion Drives the Concerted Evolution of Paralogous Copies of the SRY Gene in European Rabbits. Mol Biol Evol. 2010 Nov 1;27(11):2437–40.
380. Connallon T, Clark AG. Gene Duplication, Gene Conversion and the Evolution of the Y Chromosome. Genetics. 2010 Sep 1;186(1):277–86.
381. Lange J, Skaletsky H, van Daalen SKM, Embry SL, Korver CM, Brown LG, et al. Isodicentric Y chromosomes and sex disorders as byproducts of homologous recombination that maintains palindromes. Cell. 2009;138(5):855–69.
382. Charlesworth B. The organization and evolution of the human Y chromosome. Genome Biol. 2003 Aug 14;4(9):226.
383. Guiler ER, Heddle RWL. Testicular and body temperatures in the Tasmanian Devil and three other species of marsupial. Comp Biochem Physiol. 1970 Apr 15;33(4):881–91.
384. Prontera P, Donti E. Hypothesis: gonadal temperature influences sex-specific imprinting. Front Genet. 2014 Aug 25;5:294.
385. Kleisner K, Ivell R, Flegr J. The evolutionary history of testicular externalization and the origin of the scrotum. J Biosci. 2010 Mar 1;35(1):27–37.
386. Voosen PA. 500-million-year survey of Earth's climate reveals dire warning for humanity. Science. 2019;
387. Martin RD. Primates. Curr Biol. 2012;22(18):R785–90.
388. Pedersen RØ, Sandel B, Svenning JC. Macroecological evidence for competitive regional-scale interactions between the two major clades of mammal carnivores (Feliformia and Caniformia). PloS One. 2014;9(6):e100553.

APPENDIX 1

Supplementary material in support of Chapter 2 of this thesis.

Supplementary Table 2.1 Sample list for CAGE-Seq analysis.

ID	RIN	Barcode	Lane	Breed	Tissue
LN3	9.4	GTA	ULCAGE06	Bulldog	LYMPH NODE
LN1	8.4	TAG	ULCAGE05	Cane Corso	LYMPH NODE
RE2	7.5	ATG	ULCAGE07	Cane Corso	RETINA
BM1	7	GTA	ULCAGE07	Dogue de Bordeaux	BONE MARROW
CE2	9.1	GCC	ULCAGE06	Dogue de Bordeaux	CEREBELLUM
CL4	7.7	ATG	ULCAGE06	Dogue de Bordeaux	CARTILAGE
HE2	8	TAG	ULCAGE07	Dogue de Bordeaux	HEART
LV2	8.1	TGG	ULCAGE06	Dogue de Bordeaux	LIVER
BM4	7.3	GCC	ULCAGE07	Labrador retriever	BONE MARROW
CE1	9.3	TGG	ULCAGE07	Labrador retriever	CEREBELLUM
CL3	7.4	ACG	ULCAGE07	Labrador retriever	CARTILAGE
HE1	8.1	TGG	ULCAGE05	Labrador retriever	HEART
LV1	9	GAT	ULCAGE07	Labrador retriever	LIVER
RE1	7.5	GCC	ULCAGE05	Labrador retriever	RETINA
HE3	7.9	ACG	ULCAGE06	lhasa apso	HEART
LV3	8.5	GAT	ULCAGE06	lhasa apso	LIVER
RE3	8.9	CTT	ULCAGE05	lhasa apso	RETINA
BM3	9.5	ATG	ULCAGE05	Whip	BONE MARROW
CE4	9.1	CTT	ULCAGE07	Whip	CEREBELLUM
CL1	7.7	CTT	ULCAGE06	Whip	CARTILAGE
LN4	9.3	TAG	ULCAGE06	Whip	LYMPH NODE

APPENDIX 2

Supplementary material in support of Chapter 3 of this thesis.

Supplementary Table 3.1 Assembly accuracy evaluated in the variants-based method by *de novo* variants.

Y scaffold	POS	REF	ALT	Class
chrY1	77793	A	T	A->T
chrY1	236966	G	A	G->A
chrY1	236971	G	A	G->A
chrY1	403485	C	T	C->T
chrY1	456240	A	G	A->G
chrY1	456278	T	C	T->C
chrY1	574052	G	A	G->A
chrY1	737571	C	T	C->T
chrY2	138386	G	T	G->T
chrY2	138440	A	G	A->G
chrY2	194669	A	T	A->T
chrY3	570896	C	A	C->A
chrY1	63035	CT	CTTTTTTTTTTTTTTTTTT	Homopolymeric
chrY1	73466	CAAAA	CAAAAAAAAAAAAAAAAAAAA	Homopolymeric
chrY1	74425	TAAAA	TAAAAAAAAAAAAAAAAAA	Homopolymeric

chrY1	132495	ATTTT	ATTTTTTTTTTTTTTTTTTTTT	Homopolymeric
chrY1	144105	GAAAA	GAAAAAAAAAAAAAAAAAAAA	Homopolymeric
chrY1	211382	TAAAA	TAAAAAAAAAAAAAAAAAAAAAAAA	Homopolymeric
chrY1	236964	GA	GAAAAAAAAAAAAAAAGAAAAGAAAAAAAA	Homopolymeric
chrY1	245404	GAAAA	GAAAAAAAAAAAAAAAAAAA	Homopolymeric
chrY1	267159	CGGGG	CGGGGGG	Homopolymeric
chrY1	273998	TAAAA	TAAAAAAAAAAAAAAAAAAAAAAAA	Homopolymeric
chrY1	317184	GAAAA	GAAAAAAAAAAAAAAAAAA	Homopolymeric
chrY1	328532	TAA	TAAAAAAAAAAAAAAA	Homopolymeric
chrY1	337071	CAA	CAAAAAAAAAAAAAAAAA	Homopolymeric
chrY1	377949	CAAAA	CAAAAAAAAAAAAAAAAAAAAA	Homopolymeric
chrY1	378029	ACCCCCACCTTTTCCAAGTCCC	ACCC	Deletion
chrY1	403800	TCCCC	TCCCCC	Homopolymeric
chrY1	403861	GC	GCCCCCCCCCCCCCC	Homopolymeric
chrY1	487206	GAAAA	GAAAAAAAAAAAAA	Homopolymeric
chrY1	508760	CAAAA	CAAAAAAAAAAAAAAAAAAAA	Homopolymeric
chrY1	544006	CAAAA	CAAAAAAAAAAAAAAAAAAAA	Homopolymeric
chrY1	570671	TG	TGGGGGGGGGGGGG	Homopolymeric
chrY1	638851	TAAAA	TAAAAAAAAAAAAAAAAAAA	Homopolymeric
chrY1	640432	CAAAA	CAAAAAAAAAAAAAAAAAAAAAAA	Homopolymeric
chrY1	664737	TAAAA	TAAAAAAAAAAAAAAAAAAA	Homopolymeric
chrY1	668196	GAAAA	GAAAAAAAAAAAAA	Homopolymeric

chrY1	728366	GAA	GAAAAAAAAAAAAA	Homopolymeric
chrY1	741593	CTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTATT	CTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTATTTTATTTTATTTTATT	Simple repeat
chrY1	1111434	TTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTT	TTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTT	Simple repeat
chrY1	1111494	CTTATTATTATTATTATTATTATTATTATTATTAT	CTTATTATTATTATTATTATTATTATTATTATTATTAT	Simple repeat
chrY1	1206518	GGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGT	GGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGT	Simple repeat
chrY1	1415560	GC	GCCCCCCCC	Homopolymeric
chrY1	1466611	CT	CTTTTTTTTTTTTTTTTTT	Homopolymeric
chrY2	138381	ATTTT	ATTTTT	Homopolymeric
chrY2	173634	GCTCTCTCTCTCTCTCTCTCTC	GCTCTCTCTCTCTCTCTCTC	Simple repeat
chrY2	179126	GCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT	GCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT	Simple repeat

Supplementary Table 3.2 Assembly accuracy evaluated in the variants-based method by alignment-based variants.

Y scaffold	POS	REF	ALT	Class
chrY1	77793	A	T	A->T
chrY1	245321	A	G	A->G
chrY1	245329	T	G	T->G
chrY1	245357	T	C	T->C
chrY1	574052	G	A	G->A
chrY1	781743	A	T	A->T
chrY1	889241	G	A	G->A
chrY2	194669	A	T	A->T
chrY1	63035	C	CTTTTTTTTTTTTTTTTTT	Homopolymeric
chrY1	65826	G	GTTTTTTTTTTTT	Homopolymeric
chrY1	65829	A	ATTT	Homopolymeric
chrY1	73466	C	CAAAAAAAAAAAAAAAAA	Homopolymeric
chrY1	132495	A	ATTTTTTTTTTTTTTTTTT	Homopolymeric
chrY1	144105	G	GAAAAAAAAAAAAAAAAA	Homopolymeric
chrY1	211382	T	TAAAAAAAAAAAAAAAAAAAA	Homopolymeric
chrY1	236964	G	GAAAAAAAAAAAAA	Homopolymeric
chrY1	236971	G	GAAAAAAAAAAAAA	Homopolymeric
chrY1	245404	G	GAAAAAAAAAAAAAAA	Homopolymeric
chrY1	258780	G	GTTTTTTTTTTTTTT	Homopolymeric
chrY1	273998	T	TAAAAAAAAAAAAAAAAAAA	Homopolymeric

chrY1	317184	G	GAAAAAAAAAAAAAA	Homopolymeric
chrY1	328532	T	TAAAAAAAAAAAAA	Homopolymeric
chrY1	337071	C	CAAAAAAAAAAAAAAA	Homopolymeric
chrY1	377949	C	CAAAAAAAAAAAAAAAAA	Homopolymeric
chrY1	438950	C	CTCTT	Simple repeat
chrY1	456240	A	AG	Homopolymeric
chrY1	487206	G	GAAAAAAAAA	Homopolymeric
chrY1	508760	C	CAAAAAAAAAAAAAAAA	Homopolymeric
chrY1	526840	C	CTTTTTTTTTTT	Homopolymeric
chrY1	526854	C	CAAAAAAAAAAAAAAA	Homopolymeric
chrY1	544006	C	CAAAAAAAAAAAAAAAA	Homopolymeric
chrY1	570671	T	TGGGGGGGGGGG	Homopolymeric
chrY1	597447	T	TAAAAAAAAAAAAAAAAAAAAAAAAA	Homopolymeric
chrY1	638851	T	TAAAAAAAAAAAAAAA	Homopolymeric
chrY1	640432	C	CAAAAAAAAAAAAAAAAAA	Homopolymeric
chrY1	664737	T	TAAAAAAAAAAAAAAA	Homopolymeric
chrY1	668196	G	GAAAAAAAAA	Homopolymeric
chrY1	670058	C	CTT	Homopolymeric
chrY1	760251	G	GTTTTTTTTTTTTTT	Homopolymeric
chrY1	760259	C	CTTTTTTTTTTTTT	Homopolymeric
chrY1	816682	A	ATTTTTTTTTTTT	Homopolymeric

chrY1	1111494	C	CTTA	Simple repeat
chrY1	1201823	C	CAAAAAAAAAAAAAAAAAAA	Homopolymeric
chrY1	1201904	C	CAAAAAAAAAAAAAA	Homopolymeric
chrY1	1206518	GGT	G	Simple repeat
chrY1	1466611	C	CTTTTTTTTTTTTTTTT	Homopolymeric
chrY1	1481802	C	CTTTTTTTTTTTTTTTTTTT	Homopolymeric
chrY2	173634	GCT	G	Simple repeat
chrY2	179126	GCTCT	G	Simple repeat
chrY3	272443	TCTTC	T	Simple repeat
chrY3	2614511	GTTTTTCTTTTTTCTTTT	G	Simple repeat

Supplementary Table 3.3 Homology regions between Y chromosomes and autosomes or the X chromosome.

Scaffold	Start	End	Chromosome	Start	End	Length
chrY1	1428897	1441633	chrX	10132020	10143770	12736
chrY3	157727	184640	chr6	73539665	73566384	26913
chrY3	93706	112777	chr6	73515409	73534750	19071
chrY3	275004	293838	chr6	73481335	73499971	18834
chrY3	129514	144594	chr6	73519365	73534750	15080
chrY3	306751	315949	chr37	552671	561553	9198
chrY3	1627699	1638776	chr32	58861	69463	11077
chrY3	3269201	3286694	chr32	7377	25283	17493
chrY3	3193406	3205622	chr32	7377	19769	12216
chrY3	3297718	3306160	chr32	42023	50585	8442
chrY3	1704006	1752524	chr19	21349655	21396817	48518
chrY3	1794135	1828441	chr19	21442992	21476966	34306
chrY3	1753155	1784739	chr19	21396667	21427464	31584
chrY3	1828992	1860043	chr19	21479197	21510267	31051
chrY3	1928128	1949740	chr19	21570820	21591797	21612
chrY3	3839278	3861081	chr19	21570820	21591797	21803
chrY3	1860042	1874961	chr19	21513981	21528571	14919
chrY3	1957285	1970317	chr19	21599542	21612365	13032
chrY3	3821351	3839280	chr19	21551336	21568987	17929
chrY3	3868847	3882101	chr19	21599542	21612365	13254
chrY3	1970302	1982420	chr19	21618632	21630504	12118
chrY3	1903062	1928129	chr19	21550122	21568987	25067
chrY3	3882086	3894386	chr19	21618632	21630504	12300
chrY3	1785044	1794151	chr19	21427838	21436746	9107
chrY3	498150	517209	chr19	21399181	21418382	19059
chrY3	1875133	1885218	chr19	21528738	21538690	10085
chrY3	589548	601300	chr19	21618667	21630495	11752
chrY3	3025798	3039953	chr19	21397418	21409162	14155
chrY3	378080	388320	chr19	21566743	21576240	10240

chrY3	3049677	3057959	chr19	21368268	21376127	8282
chrY3	452040	461072	chr19	21485473	21494783	9032
chrY3	422944	431997	chr19	21485473	21494783	9053
chrY3	485066	494407	chr19	21564403	21573362	9341
chrY3	390368	398740	chr19	21564403	21572317	8372
chrY3	52773	84470	chr17	65420347	65451958	31697
chrY3	206694	216537	chr17	65351460	65360855	9843
chrY3	244605	254148	chr15	17435943	17445777	9543
chrY3	301984	314084	chr15	130107	141078	12100
chrY3	108912	120234	chr15	17377984	17387600	11322
chrY3	244605	253115	chr15	254311	263424	8510
chrY3	263729	272158	chr15	17386177	17394948	8429
chrY3	1667981	1676723	chr15	17467664	17475438	8742

Supplementary Table 3.4 Copy number of 28,800 bp-length LINE_CF1 array estimated by short read data.

Assession	Depth	Breeds	Sex
SRR8614085	99.2093	Alaskan Malamute	Male
SAMC045928	81.9413	Cretan Tracer	Male
SRR13376372	81.9351	Cretan Hound	Male
SRR13376336	78.4809	Cretan Hound	Male
SAMC045927	78.4671	Cretan Tracer	Male
SRR7120191	75.9941	New Guinea Singing Dog	Male
SRR10077572	74.4953	Tibetan Terrier	Male
SRR11193499	65.8019	Greenland dog	Male
ERR5383445	65.0492	West Highland White Terrier	Male
SRR7120150	64.8081	Bull Terrier	Male
SRR10077562	64.3201	Pit bull	Male
SRR7120152	64.0919	Bull Terrier	Male
SRR10752628	63.9613	American Bulldog	Male
SRR8614017	63.8598	Cane Corso	Male
SRR8614037	63.5739	Cane Corso	Male
SRR7107973	62.1593	Berger Picard	Male
SRR7120149	60.6946	Bull Terrier	Male
ERR2196103	58.8829	Bull Terrier	Male
SRR11193494	58.8428	Greenland dog	Male
SRR10441628	58.8251	Staffordshire Bull Terrier	Male
ERR2113145	58.7046	Jagdterrier	Male
SRR7107560	58.1994	Jagdterrier	Male
ERR2196104	57.1972	Japanese Chin	Male
SRR7107521	57.1772	Korean Jindo	Male
SRR8614031	53.4478	Cane Corso	Male
ERR2196105	51.554	Dogue de Bordeaux	Male
SRR10441649	49.342	Boxer	Male
SRR10441650	46.4643	Boxer	Male
SRR8614052	46.2006	Rhodesian Ridgeback	Male

SRR7107588	36.7884	Alpine Dachsbracke	Male
ERR2196280	35.7812	Saluki	Male
ERR3047551	34.1619	Shih Tzu	Male
SRR10752621	33.8025	Chesapeake Bay Retriever	Male
SRR7120212	33.7287	Scottish Terrier	Male
SRR5664964	33.4571	Cocker Spaniel	Male
SRR7107974	33.2365	Chesapeake Bay Retriever	Male
ERR3047552	33.1868	Shih Tzu	Male
SRR7107963	33.0655	Field Spaniel	Male
SRR7107977	33.0469	English Setter	Male
SRR8614029	33.0379	Giant Schnauzer	Male
SRR10752647	33.0069	Kerry Blue Terrier	Male
SRR10752629	32.9221	English Mastiff	Male
SRR8614021	32.7964	Tibetan Terrier	Male
ERR3339009	32.7608	German Shorthaired Pointer	Male
SRR8614028	32.6934	Tibetan Terrier	Male
ERR2750975	32.6606	West Highland White Terrier	Male
ERR3047522	32.3925	Brussels Griffon	Male
SRR8614045	32.3692	Dutch Shepherd	Male
SRR7107969	32.245	Greater Swiss Mountain Dog	Male
SRR8614036	32.1832	Pointer	Male
ERR3339011	32.1557	Lagotto Romagnolo	Male
SRR8614090	32.0613	Labrador Retriever	Male
ERR2113149	31.8838	Australian Cattle Dog	Male
SRR8614087	31.8163	Bouvier des Flandres	Male
SRR7107980	31.7236	Labrador Retriever	Male
SRR7107797	31.6389	Golden Retriever	Male
SRR10752624	31.5011	Standard Poodle	Male
SRR8614055	31.3658	English Springer Spaniel	Male
SRR10752618	31.3506	Flat-Coated Retriever	Male
ERR2750978	31.2096	Curly Coated Retriever	Male
SRR8614022	31.0924	Standard Poodle	Male

SRR7107835	31.0574	Jämthund	Male
ERR2196101	31.0369	Bichon Frise	Male
SRR14750534	31.0035	Braque Français	Male
SRR8614034	30.9475	Golden Retriever	Male
SRR7107966	30.9184	Bouvier des Flandres	Male
SRR10077564	30.7907	Golden Retriever	Male
SRR8163597	30.6674	Standard Poodle	Male
SRR7107791	30.6095	Greyhound	Male
ERR5383420	30.6073	Great Dane	Male
SRR8614076	30.599	Basenji	Male
SRR7120169	30.4932	Irish Water Spaniel	Male
SRR10752617	30.4643	Whippet	Male
SRR14750475	30.4425	Old English Sheepdog	Male
SRR10077570	30.3003	Cane Corso	Male
SRR7107968	30.2326	Border Terrier	Male
SRR7107798	30.1523	Golden Retriever	Male
SRR7107984	30.1398	Basenji	Male
ERR2196102	30.0736	Brussels Griffon	Male
SRR10752622	30.0676	Boston Terrier	Male
SRR7107967	29.9586	Great Dane	Male
SRR7764564	29.8899	Maltese dog	Male
SRR11671234	29.811	Great Dane	Male
SRR7107539	29.7934	Bavarian Mountain Hound	Male
SRR7107796	29.7193	Golden Retriever	Male
ERR3047554	29.6266	Pomeranian	Male
ERR2008786	29.6149	Malinois dog	Male
SRR10752627	29.6068	Pekingese	Male
ERR2759443	29.5847	Greyhound	Male
SRR7120201	29.4405	Portuguese Water Dog	Male
SRR7107789	29.4351	Greyhound	Male
SRR7107793	29.3733	Golden Retriever	Male

ERR2750980	29.359	Greyhound	Male
SRR7120172	29.3136	Komondor	Male
ERR5449487	29.303	Leonberger	Male
SRR8614046	29.2711	Labrador Retriever	Male
ERR3047538	29.1657	Havanese	Male
SRR10752619	29.0963	Dobermann	Male
SRR8163595	28.9992	Standard Poodle	Male
ERR2759440	28.9921	Curly Coated Retriever	Male
SRR7107856	28.9266	Golden Retriever	Male
SRR8614053	28.8121	Dobermann	Male
SRR7120181	28.7933	Labrador Retriever	Male
SRR7120160	28.6127	Golden Retriever	Male
SRR7107534	28.3693	Great Dane	Male
ERR5383416	28.3128	Bernese Mountain Dog	Male
ERR5383440	28.2449	Malinois dog	Male
ERR3478972	28.2409	Norwegian buhund	Male
ERR3047537	28.2081	Havanese	Male
SRR7107855	28.1785	Golden Retriever	Male
SRR8614041	28.1347	Otterhound	Male
ERR4579528	28.0807	Leonberger	Male
SRR5664961	28.0364	Cocker Spaniel	Male
SRR7107964	28.0227	Standard Schnauzer	Male
SRR8614059	27.9766	Irish Setter	Male
SRR7107982	27.9757	Standard Schnauzer	Male
SRR7120177	27.9028	Labrador Retriever	Male
ERR3047553	27.8751	Pomeranian	Male
ERR3047533	27.8687	Chihuahua	Male
ERR2196282	27.8104	Whippet	Male
ERR3284981	27.8098	Giant Schnauzer	Male
ERR5383446	27.8064	Pomeranian	Male
SRR7107565	27.7984	Labrador Retriever	Male

SRR7764562	27.7422	Yorkshire Terrier	Male
SRR8614033	27.666	Pomeranian	Male
SRR8614019	27.5968	Dalmatian	Male
SRR10752620	27.5628	Jack Russell Terrier	Male
SRR10441645	27.4782	Flat-Coated Retriever	Male
ERR3284984	27.4511	Giant Schnauzer	Male
SRR8614027	27.355	Irish Setter	Male
ERR3284982	27.2898	Giant Schnauzer	Male
ERR5383409	27.1583	Bernese Mountain Dog	Male
SRR2095320	27.1367	Golden Retriever	Male
SRR8614044	27.1352	Jack Russell Terrier	Male
SRR14750438	27.1195	French spaniel	Male
ERR5383411	27.0487	Bernese Mountain Dog	Male
SRR7120182	27.0442	Labrador Retriever	Male
SRR10077568	27.0382	Irish Terrier	Male
ERR2196278	27.0165	Pug	Male
SRR10077566	26.9925	Dachshund	Male
SRR8614049	26.9906	Alaskan Klee Kai	Male
SRR7107579	26.9867	Chihuahua	Male
SRR8614057	26.9377	Standard Schnauzer	Male
ERR2196277	26.9284	Pug	Male
SRR8163592	26.8922	Standard Poodle	Male
SRR7107979	26.8784	Irish Terrier	Male
ERR3047534	26.7359	Dachshund	Male
ERR2196098	26.7132	Bichon Frise	Male
ERR5383413	26.5728	Bernese Mountain Dog	Male
SRR7107971	26.5661	Dachshund	Male
SRR10752626	26.5364	Toy Poodle	Male
SRR7107976	26.5292	Beagle	Male
SRR7107790	26.5089	Greyhound	Male
ERR2196096	26.4592	American Bulldog	Male

SRR7120205	26.4388	Rottweiler	Male
SRR10441652	26.062	Bernese Mountain Dog	Male
ERR5449484	25.9056	Leonberger	Male
SRR10441631	25.8863	Rottweiler	Male
ERR5449479	25.8568	Leonberger	Male
ERR5449483	25.8335	Leonberger	Male
ERR5449486	25.792	Leonberger	Male
ERR2764781	25.7204	Leonberger	Male
ERR5449478	25.7062	Leonberger	Male
SRR2094385	25.675	Beagle	Male
SRR7120206	25.6151	Rottweiler	Male
ERR2008784	25.4894	Jack Russell Terrier	Male
SRR10441632	25.3423	Pug	Male
ERR3047548	25.1844	Petit Basset Griffon Vendeen	Male
ERR2196264	24.9517	Pug	Male
SRR7120180	24.9208	Labrador Retriever	Male
SRR7107965	24.8382	Dachshund	Male
ERR2196270	24.7953	Pug	Male
ERR2196272	24.7794	Pug	Male
ERR2196271	24.5648	Pug	Male
SRR8163600	24.5369	Standard Poodle	Male
ERR2196266	24.4377	Pug	Male
ERR2113150	24.1199	German Shepherd	Male
ERR2196269	24.1167	Pug	Male
SRR8614056	23.4738	Standard Schnauzer	Male
SRR10441637	22.6563	Cavalier King Charles Spaniel	Male
SRR8614050	22.0537	Shiba Inu	Male
ERR5383433	21.9543	Labrador Retriever	Male
SRR10752637	21.8677	Shiba Inu	Male
SRR7107916	21.7236	Yorkshire Terrier	Male
ERR5449485	20.6659	Leonberger	Male

ERR1688122	14.6114	Miniature Bull Terrier	Male
ERR2196265	0.013	Pug	Female
ERR2196267	0.014	Pug	Female
ERR2196268	0.016	Pug	Female
ERR2196273	0.003	Pug	Female
ERR2196275	0.006	Pug	Female
ERR2750976	0.009	Alaskan Malamute	Female
ERR2750983	0.01	Chihuahua	Female
ERR2759445	0.016	Lagotto Romagnolo	Female
ERR3047532	0.019	Chihuahua	Female
ERR3047546	0.013	Papillon	Female
ERR3047547	0.011	Papillon	Female
ERR5383434	0.007	Lagotto Romagnolo	Female
ERR5383436	0.011	Lagotto Romagnolo	Female
ERR5383438	0.009	Lagotto Romagnolo	Female
SRR10077545	0.01	Alapaha Blue Blood Bulldog	Female
SRR10752633	0.012	Alapaha Blue Blood Bulldog	Female
SRR10752636	0.011	Boston Terrier	Female
SRR10758785	0.024	Airedale Terrier	Female
SRR2095478	0.011	Chihuahua	Female
SRR2095500	0.01	Pekingese	Female
SRR2095503	0.014	Saluki	Female
SRR7107882	0.026	Airedale Terrier	Female
SRR7107899	0.005	Border Terrier	Female
SRR7107927	0.026	Border Terrier	Female
SRR7107950	0.014	Border Collie	Female
SRR7120144	0.042	Borzoi	Female
SRR7120208	0.019	Rottweiler	Female
SRR7120209	0.014	Rottweiler	Female
SRR8614018	0.157	Parson Russell Terrier	Female

APPENDIX 3

Supplementary material in support of Chapter 4 of this thesis.

Supplementary Table 4.1 Selected RNA-Seq samples for the RosY_1.0 annotation.

Accession	Projects	Tissue	Breed	Stage
ERR1948877	PRJEB17926	Testes	Cavalier King Charles spaniel	Adult
ERR1948878	PRJEB17926	Testes	French bulldog	Adult
ERR1948879	PRJEB17926	Testes	Italian greyhound	Adult
ERR1948880	PRJEB17926	Testes	Papillon	Adult
ERR1948881	PRJEB17926	Testes	Pug	Adult
ERR1948882	PRJEB17926	Testes	Whippet	Adult
ERR1948883	PRJEB17926	Testes	Yorkshire terrier	Adult
SRR5889318	PRJNA396033	Lung	ACS X Beagle	Embryo day 36
SRR5889319	PRJNA396033	Heart	ACS X Beagle	Embryo day 36
SRR5889322	PRJNA396033	Liver	ACS X Beagle	Embryo day 36
SRR5889327	PRJNA396033	Lung	Beagle	Embryo day 39
SRR5889328	PRJNA396033	Kidney	Beagle	Embryo day 39
SRR5889330	PRJNA396033	Liver	Beagle	Embryo day 39
SRR8996953	PRJNA396033	Kidney	Yorkshire Terrier	Adult
SRR8996955	PRJNA396033	Bone marrow	Yorkshire Terrier	Adult
SRR8996956	PRJNA396033	Bladder	Yorkshire Terrier	Adult
SRR8996957	PRJNA396033	Adrenal gland	Yorkshire Terrier	Adult
SRR8996958	PRJNA396033	Adipose	Yorkshire Terrier	Adult
SRR8996959	PRJNA396033	Frontal cortex	Yorkshire Terrier	Adult
SRR8996965	PRJNA396033	Lung	Newfoundland	Adult
SRR8996966	PRJNA396033	Liver	Newfoundland	Adult
SRR8996967	PRJNA396033	Pancreas	Newfoundland	Adult
SRR8996968	PRJNA396033	Lymph node	Newfoundland	Adult
SRR8996970	PRJNA396033	Pituitary gland	Newfoundland	Adult
SRR8996971	PRJNA396033	Skeletal muscle	Newfoundland	Adult
SRR8996972	PRJNA396033	Salivary gland	Newfoundland	Adult

SRR8996977	PRJNA396033	Liver	Yorkshire Terrier	Adult
SRR8996979	PRJNA396033	Lymph node	Yorkshire Terrier	Adult
SRR8996980	PRJNA396033	Lung	Yorkshire Terrier	Adult
SRR8996981	PRJNA396033	Pituitary gland	Yorkshire Terrier	Adult
SRR8996983	PRJNA396033	Adipose	Newfoundland	Adult
SRR8996993	PRJNA396033	Bone marrow	Newfoundland	Adult
SRR8996994	PRJNA396033	Cartilage	Newfoundland	Adult
SRR8996995	PRJNA396033	Adrenal gland	Newfoundland	Adult
SRR8996996	PRJNA396033	Bladder	Newfoundland	Adult
SRR8996997	PRJNA396033	Frontal cortex	Newfoundland	Adult
SRR8997031	PRJNA396033	Small intestine	Belgian Malanois	Adult
SRR8997032	PRJNA396033	Spleen	Belgian Malanois	Adult
SRR8997033	PRJNA396033	Frontal cortex	Belgian Malanois	Adult
SRR8997034	PRJNA396033	Colon	Belgian Malanois	Adult
SRR8997035	PRJNA396033	Cerebellum	Belgian Malanois	Adult
SRR8997036	PRJNA396033	Bone marrow	Belgian Malanois	Adult
SRR8997040	PRJNA396033	Kidney	Belgian Malanois	Adult
SRR8997041	PRJNA396033	Lung	Belgian Malanois	Adult
SRR8997042	PRJNA396033	Liver	Belgian Malanois	Adult
SRR8997043	PRJNA396033	Skin	Newfoundland	Adult
SRR8997044	PRJNA396033	Spleen	Newfoundland	Adult
SRR8997045	PRJNA396033	Stomach	Newfoundland	Adult
SRR8997047	PRJNA396033	Adrenal gland	Belgian Malanois	Adult

Supplementary Table 4.2 94 RNA-Seq samples selected for expression analysis from a wide range of tissues.

Accession	Projects	Tissue	Breed	Stage
SRR8996958	PRJNA396033	Adipose	Yorkshire Terrier	Adult
SRR8996983	PRJNA396033	Adipose	Newfoundland	Adult
SRR8997020	PRJNA396033	Adipose	Belgian Malanois	Adult
SRR8997052	PRJNA396033	Adipose	Belgian Malanois	Adult
SRR8996957	PRJNA396033	Adrenal gland	Yorkshire Terrier	Adult
SRR8996995	PRJNA396033	Adrenal gland	Newfoundland	Adult
SRR8997019	PRJNA396033	Adrenal gland	Belgian Malanois	Adult
SRR8997047	PRJNA396033	Adrenal gland	Belgian Malanois	Adult
SRR8996956	PRJNA396033	Bladder	Yorkshire Terrier	Adult
SRR8996996	PRJNA396033	Bladder	Newfoundland	Adult
SRR8997014	PRJNA396033	Bladder	Belgian Malanois	Adult
SRR8997048	PRJNA396033	Bladder	Belgian Malanois	Adult
SRR8996955	PRJNA396033	Bone marrow	Yorkshire Terrier	Adult
SRR8996993	PRJNA396033	Bone marrow	Newfoundland	Adult
SRR8997013	PRJNA396033	Bone marrow	Belgian Malanois	Adult
SRR8997036	PRJNA396033	Bone marrow	Belgian Malanois	Adult
SRR8996962	PRJNA396033	Cartilage	Yorkshire Terrier	Adult
SRR8996994	PRJNA396033	Cartilage	Newfoundland	Adult
SRR8997016	PRJNA396033	Cartilage	Belgian Malanois	Adult
SRR8996961	PRJNA396033	Cerebellum	Yorkshire Terrier	Adult
SRR8996999	PRJNA396033	Cerebellum	Newfoundland	Adult
SRR8997015	PRJNA396033	Cerebellum	Belgian Malanois	Adult
SRR8997035	PRJNA396033	Cerebellum	Belgian Malanois	Adult
SRR8996960	PRJNA396033	Colon	Yorkshire Terrier	Adult
SRR8997000	PRJNA396033	Colon	Newfoundland	Adult
SRR8997022	PRJNA396033	Colon	Belgian Malanois	Adult
SRR8997034	PRJNA396033	Colon	Belgian Malanois	Adult
SRR8996959	PRJNA396033	Frontal cortex	Yorkshire Terrier	Adult
SRR8996997	PRJNA396033	Frontal cortex	Newfoundland	Adult
SRR8997021	PRJNA396033	Frontal cortex	Belgian Malanois	Adult
SRR8997033	PRJNA396033	Frontal cortex	Belgian Malanois	Adult

SRR5889317	PRJNA396033	Head	ACS X Beagle	Embryo day 36
SRR5889319	PRJNA396033	Heart	ACS X Beagle	Embryo day 36
SRR5889325	PRJNA396033	Heart	Beagle	Embryo day 39
SRR5889320	PRJNA396033	Kidney	ACS X Beagle	Embryo day 36
SRR5889328	PRJNA396033	Kidney	Beagle	Embryo day 39
SRR8996953	PRJNA396033	Kidney	Yorkshire Terrier	Adult
SRR8997001	PRJNA396033	Kidney	Newfoundland	Adult
SRR8997005	PRJNA396033	Kidney	Belgian Malanois	Adult
SRR8997040	PRJNA396033	Kidney	Belgian Malanois	Adult
SRR5889322	PRJNA396033	Liver	ACS X Beagle	Embryo day 36
SRR5889330	PRJNA396033	Liver	Beagle	Embryo day 39
SRR8996966	PRJNA396033	Liver	Newfoundland	Adult
SRR8996977	PRJNA396033	Liver	Yorkshire Terrier	Adult
SRR8997009	PRJNA396033	Liver	Belgian Malanois	Adult
SRR8997042	PRJNA396033	Liver	Belgian Malanois	Adult
SRR5889318	PRJNA396033	Lung	ACS X Beagle	Embryo day 36
SRR5889327	PRJNA396033	Lung	Beagle	Embryo day 39
SRR8996965	PRJNA396033	Lung	Newfoundland	Adult
SRR8996980	PRJNA396033	Lung	Yorkshire Terrier	Adult
SRR8997010	PRJNA396033	Lung	Belgian Malanois	Adult
SRR8997041	PRJNA396033	Lung	Belgian Malanois	Adult
SRR8996968	PRJNA396033	Lymph node	Newfoundland	Adult
SRR8996979	PRJNA396033	Lymph node	Yorkshire Terrier	Adult
SRR8997011	PRJNA396033	Lymph node	Belgian Malanois	Adult
SRR8997029	PRJNA396033	Lymph node	Belgian Malanois	Adult
SRR8996967	PRJNA396033	Pancreas	Newfoundland	Adult
SRR8997003	PRJNA396033	Pancreas	Belgian Malanois	Adult
SRR8997027	PRJNA396033	Pancreas	Belgian Malanois	Adult
SRR8996970	PRJNA396033	Pituitary gland	Newfoundland	Adult
SRR8996981	PRJNA396033	Pituitary gland	Yorkshire Terrier	Adult
SRR8997004	PRJNA396033	Pituitary gland	Belgian Malanois	Adult
SRR8996972	PRJNA396033	Salivary gland	Newfoundland	Adult
SRR8996990	PRJNA396033	Salivary gland	Belgian Malanois	Adult

SRR8997026	PRJNA396033	Salivary gland	Belgian Malanois	Adult
SRR8997053	PRJNA396033	Salivary gland	Yorkshire Terrier	Adult
SRR8996971	PRJNA396033	Skeletal muscle	Newfoundland	Adult
SRR8996989	PRJNA396033	Skeletal muscle	Belgian Malanois	Adult
SRR8997023	PRJNA396033	Skeletal muscle	Belgian Malanois	Adult
SRR8997054	PRJNA396033	Skeletal muscle	Yorkshire Terrier	Adult
SRR8996988	PRJNA396033	Skin	Belgian Malanois	Adult
SRR8997024	PRJNA396033	Skin	Belgian Malanois	Adult
SRR8997043	PRJNA396033	Skin	Newfoundland	Adult
SRR8997055	PRJNA396033	Skin	Yorkshire Terrier	Adult
SRR8996987	PRJNA396033	Small intestine	Belgian Malanois	Adult
SRR8997031	PRJNA396033	Small intestine	Belgian Malanois	Adult
SRR8997056	PRJNA396033	Small intestine	Yorkshire Terrier	Adult
SRR8996986	PRJNA396033	Spleen	Belgian Malanois	Adult
SRR8997032	PRJNA396033	Spleen	Belgian Malanois	Adult
SRR8997044	PRJNA396033	Spleen	Newfoundland	Adult
SRR8997049	PRJNA396033	Spleen	Yorkshire Terrier	Adult
SRR8996985	PRJNA396033	Stomach	Belgian Malanois	Adult
SRR8997018	PRJNA396033	Stomach	Belgian Malanois	Adult
SRR8997045	PRJNA396033	Stomach	Newfoundland	Adult
SRR8997050	PRJNA396033	Stomach	Yorkshire Terrier	Adult
ERR1948875	PRJEB17926	Testes	Bulldog	Adult
ERR1948876	PRJEB17926	Testes	Cavalier King Charles spaniel	Adult
ERR1948877	PRJEB17926	Testes	Cavalier King Charles spaniel	Adult
ERR1948878	PRJEB17926	Testes	French bulldog	Adult
ERR1948879	PRJEB17926	Testes	Italian greyhound	Adult
ERR1948880	PRJEB17926	Testes	Papillon	Adult
ERR1948881	PRJEB17926	Testes	Pug	Adult
ERR1948882	PRJEB17926	Testes	Whippet	Adult
ERR1948883	PRJEB17926	Testes	Yorkshire terrier	Adult

Supplementary Table 4.3 The study cohort for polymorphism analysis. A total of 222 male WGS samples were utilised.

ID	Depth	Accession	Breed	Region
ABUL1510	46.38	SRR10752628	American Bulldog	America
AESK1538	22.89	SRR12330062	American Eskimo Dog	Europe
AFFN0179	27.98	ERR2196022	Affenpinscher	Europe
AFGH0821	16.43	SRR7107829	Afghan Hound	Middle East
AFOX1541	20.60	SRR12330380	American Foxhound	America
AFPV1706	23.59	SRR14750304	Anglo-Français de Petite Vénerie	Europe
AIRT0961	22.57	SRR7107922	Airedale Terrier	Europe
ALDA0255	35.12	SRR7107588	Alpine Dachsbracke	Europe
ALHUSK0240	15.88	SRR7107573	Alaskan Husky	Arctic
AMAL0140	17.08	SRR7107548	Alaskan Malamute	Arctic
AMST0413	24.97	ERR2759437	American Staffordshire Terrier	America
AMWS1542	22.90	SRR12330280	American Water Spaniel	America
APPS1707	25.60	SRR14750315	Appenzeller Sennenhund	Europe
ARGS1708	22.80	SRR14750327	Ariegeois	Europe
ATOL1703	24.42	SRR14750414	Anatolian Shepherd Dog	Europe
AUCD0118	25.64	SRR7107537	Australian Cattle Dog	Europe
AUSS1546	23.33	SRR12330203	Australian Shepherd	America
AUST1547	25.29	SRR12330180	Australian Terrier	Europe
AUVP1712	23.72	SRR14750381	Auvergne pointer	Europe
BAAN1715	23.81	SRR14750385	Basset Artésien Normand	Europe
BASS1548	25.94	SRR12330158	Basset Hound	Europe
BEAG1007	38.91	SRR7107976	Beagle	Europe
BEDT1549	22.21	SRR12330114	Bedlington Terrier	Europe
BELS1129	23.06	SRR7120114	Belgian Shepherd	Europe
BERD0299	20.40	SRR7107609	Bearded Collie	Europe
BFDB1717	21.80	SRR14750512	Basset Fauve de Bretagne	Europe
BICH1720	22.21	SRR14750518	Bichon Frise	Europe
BIET1559	23.25	SRR12330044	Biewer Terrier	Europe

BILL1722	22.37	SRR14750520	Billy	Europe
BLDH1724	21.86	SRR14750523	Bloodhound	Europe
BLGG1719	22.29	SRR14750516	Blue Gascony Griffon	Europe
BMAL0549	35.14	ERR5383440	Malinois dog	Europe
BMD1150	11.16	SRR7120136	Bernese Mountain Dog	Europe
BMHO0122	37.15	SRR7107539	Bavarian Mountain Hound	Europe
BOER1555	24.14	SRR12330074	Boerboel	Africa
BORD0986	25.72	SRR7107949	Border Collie	Europe
BORT0999	43.76	SRR7107968	Border Terrier	Europe
BOST1065	6.68	SRR5311674	Boston Terrier	America
BOUV1161	21.72	SRR7120147	Bouvier des Flandres	Europe
BOX1356	20.86	SRR8541918	Boxer	Europe
BRAF1728	32.89	SRR14750534	Braque Français	Europe
BRIT1092	9.69	SRR5311635	Brittany	Europe
BRJH1733	23.49	SRR14750396	Bruno Jura Hound	Europe
BRZT1735	22.65	SRR14750399	Brazilian Terrier	America
BSJI1399	42.46	SRR8614076	Basenji	Africa
BULD1079	6.35	SRR5311653	Bulldog	Europe
BULM1557	23.84	SRR12330047	Bullmastiff	Europe
BULT1558	21.18	SRR12330046	Bull Terrier	Europe
CANE1429	35.80	SRR10077570	Cane Corso	Europe
CARD1736	24.11	SRR14750404	Cardigan Welsh Corgi	Europe
CCRT0416	32.24	ERR2759440	Curly Coated Retriever	Europe
CESK1573	16.29	SRR12330377	Cesky Terrier	Europe
CHBR1005	39.06	SRR7107974	Chesapeake Bay Retriever	America
CHIH0246	31.11	SRR7107579	Chihuahua	America
CHIN1563	20.88	SRR12330027	Japanese Chin	East Asia
CHOW0609	7.40	SRR7107668	Chow Chow	East Asia
CIOK1569	20.42	SRR12330387	Chinook	America
CKCS1685	5.68	SRR13739072	Cavalier King Charles Spaniel	Europe

CNDE1566	21.73	SRR12330020	Cirneco dell'Etna	Europe
COCK1365	6.12	SRR8614066	Cocker Spaniel	Europe
COLL1044	25.43	SRR5190662	Collie	Europe
COTO1450	25.82	SRR10077546	Coton de Tulear	Africa
CRES0801	29.21	SRR7107801	Chinese Crested	America
CTHD1659	20.71	SRR13376356	Cretan Hound	Europe
CTTC0063	19.48	SAMC045933	Cretan Tracer	Europe
DACH1002	37.09	SRR7107971	Dachshund	Europe
DANE0998	33.94	SRR7107967	Great Dane	Europe
DEER0889	23.78	SRR7107893	Scottish Deerhound	Europe
DGAG1576	20.75	SRR12330365	Dogo Argentino	America
DGYG1020	16.81	SRR5177186	DongGyeoungi	East Asia
DING0862	6.39	SRR7107869	Dingo	East Asia
DOBP1367	12.79	SRR8614074	Dobermann	Europe
DSDO1740	23.73	SRR14750427	Dutch Shepherd	Europe
EFOX1578	22.57	SRR12330357	English Foxhound	Europe
ELO0328	16.91	ERR1688115	Elo	Europe
EMAST1762	26.82	SRR14750465	English Mastiff	Europe
ENTL1580	23.15	SRR12330354	Entlebucher Sennenhund	Europe
ESET1008	39.89	SRR7107977	English Setter	Europe
ESSP0968	27.64	SRR7107929	English Springer Spaniel	Europe
EURS0326	9.88	ERR1688114	Eurasier	Europe
FBUL1348	18.46	SRR8541932	French Bulldog	Europe
FCR1489	37.58	SRR10752618	Flat-Coated Retriever	Europe
FCSP1743	31.78	SRR14750438	French spaniel	Europe
FIEL0994	42.37	SRR7107963	Field Spaniel	Europe
FMAS1739	20.70	SRR14750544	Dogue de Bordeaux	Europe
FOXT1746	26.02	SRR14750444	Fox Terrier	Europe
GAFT1751	23.48	SRR14750301	Grand Anglo-Français Tricolore	Europe

GAWOH1753	21.04	SRR14750310	Great Anglo-French White and Orange Hound	Europe
GBDG1754	21.82	SRR14750311	Grand Bleu de Gascogne	Europe
GJAG0162	37.79	SRR7107560	Jagdterrier	Europe
GOLD1173	29.22	SRR7120161	Golden Retriever	Europe
GPYR0954	29.31	SRR7107914	Great Pyrenees	Europe
GREY0791	63.51	SRR7107791	Greyhound	Middle East
GRFF1788	23.49	SRR14750368	Griffon	Europe
GRFN1757	21.19	SRR14750320	Griffon Nivernais	Europe
GRND1524	34.74	SRR11193499	Greenland dog	Arctic
GSD0955	27.52	SRR7107915	German Shepherd	Europe
GSHP1072	7.50	SRR5311665	German Shorthaired Pointer	Europe
GSMD1758	25.81	SRR14750448	Greater Swiss Mountain Dog	Europe
GWHP1071	5.14	SRR5311666	German Wirehaired Pointer	Europe
HARR1759	21.62	SRR14750452	Harrier	Europe
HUSK0115	28.00	ERR1990016	Siberian Husky	Arctic
IBIZ1589	24.49	SRR12330309	Ibizan Hound	Middle East
ICES1591	22.85	SRR12330301	Icelandic Sheepdog	Europe
INDG0003	14.56	SAMC008992	Indigenous dog	Africa
INDG0032	21.46	SAMC036703	Indigenous dog	East Asia
INDG0034	24.81	SAMC036706	Indigenous dog	East Asia
INDG0036	18.63	SAMC036708	Indigenous dog	Middle East
INDG0040	20.77	SAMC036713	Indigenous dog	Middle East
INDG0071	17.00	SAMC052946	Indigenous dog	Middle East
INDG0617	7.19	SRR7107671	Indigenous Dog	East Asia
INDG0631	6.82	SRR1061841	Indigenous dog	Middle East
INDG0639	6.98	SRR1061958	Indigenous dog	Middle East
INDG0653	5.80	SRR1061967	Indigenous dog	East Asia
INDG0657	7.97	SRR7107690	Indigenous Dog	Middle East
INDG0658	7.73	SRR7107692	Indigenous Dog	Africa

INDG0705	12.81	SRR1138357	Indigenous Dog	East Asia
IRSE1076	6.87	SRR5311659	Irish Setter	Europe
IRTR1010	43.49	SRR7107979	Irish Terrier	Europe
ITGY0539	28.35	ERR5383430	Italian Greyhound	Europe
IWOF1353	24.67	SRR8541931	Irish wolfhound	Europe
IWSP1178	55.19	SRR7120169	Irish Water Spaniel	Europe
JACK0963	24.88	SRR7107924	Jack Russell Terrier	Europe
JAMT0827	37.16	SRR7107835	Jämthund	Europe
KERY1486	42.64	SRR10752647	Kerry Blue Terrier	Europe
KJDO0086	43.22	SRR7107521	Korean Jindo	East Asia
KMLD1760	21.53	SRR14750458	Small Munsterlander	Europe
KOMO1181	51.22	SRR7120172	Komondor	Europe
KROM0234	23.57	SRR7107567	Kromfohrlander	Europe
KUVZ1640	20.76	SRR12330106	Kuvasz	Europe
LAB1364	18.04	SRR8541929	Labrador Retriever	America
LAPPO0828	16.52	SRR7107836	Lapponian herder	Europe
LARO0356	19.39	SRR7107637	Lagotto Romagnolo	Europe
LEON0566	30.23	ERR5449487	Leonberger	Europe
LEOP1593	22.16	SRR12330281	Catahoula Leopard Dog	America
LHSA1597	21.73	SRR12330273	Lhasa Apso	East Asia
LOWC0974	27.19	SRR7107935	Lowchen	Europe
MALT1314	5.79	SRR8541953	Maltese dog	Europe
MANT1599	25.56	SRR12330267	Manchester Terrier	Europe
MAST1600	20.88	SRR12330265	Mastiff	Europe
MBLT1602	20.50	SRR12330263	Miniature Bull Terrier	Europe
MHLD0589	12.97	SRR7107660	Mexican hairless dog	America
MPIN1603	20.46	SRR12330257	Miniature Pinscher	Europe
MSNZ0441	14.02	ERR2865339	Miniature Schnauzer	Europe
MUDI1604	24.21	SRR12330254	Mudi	Europe
NBUH1605	23.37	SRR12330253	Norwegian buhund	Europe
NEAM1766	25.35	SRR14750473	Neapolitan Mastiff	Europe

NELK1608	22.83	SRR12330248	Norwegian Elkhound	Europe
NEWF1080	8.43	SRR5311652	Newfoundland	America
NEWL0236	15.87	SRR7107569	Landseer	America
NGINDG0820	17.30	SRR7107828	Nigerian Indigenous Dog	Africa
NGSD1193	46.09	SRR7120191	New Guinea Singing Dog	East Asia
NORF1610	22.18	SRR12330240	Norfolk Terrier	Europe
NOWT0103	15.38	ERR1990010	Norwich Terrier	Europe
NSDT0857	22.78	SRR7107865	Nova Scotia Duck Tolling Retriever	America
OES1767	33.31	SRR14750475	Old English Sheepdog	Europe
OTTR1612	25.23	SRR12330232	Otterhound	Europe
PBGV0243	29.51	SRR7107576	Petit Basset Griffon Vendeen	Europe
PCSL1771	24.65	SRR14750344	Picardy Spaniel	Europe
PDLP1777	21.06	SRR14750354	Pudelpointer	Europe
PEKE1511	44.05	SRR10752627	Pekingese	East Asia
PEMB0802	24.43	SRR7107802	Pembroke Welsh Corgi	Europe
PICA1768	21.18	SRR14750335	Berger Picard	Europe
PIN1764	20.80	SRR14750469	Pinscher	Europe
PIOD0830	16.95	SRR7107838	Peruvian Inca Orchid	America
PITB1434	22.85	SRR10077565	Pit bull	Europe
POD_MN1189	14.69	SRR7120186	Miniature Poodle	Europe
POD_SD1265	54.20	SRR8163600	Standard Poodle	Europe
POD_TY1512	51.72	SRR10752626	Toy Poodle	Europe
POLS1772	24.91	SRR14750346	Polish Lowland Sheepdog	Europe
POM0555	33.77	ERR5383446	Pomeranian	Europe
POPODG0875	25.68	SRR7107881	Portuguese Podengo	Europe
PTWD1204	15.57	SRR7120202	Portuguese Water Dog	Europe
PUG0212	45.64	ERR2196278	Pug	East Asia
PUMI1613	20.73	SRR12330216	Pumi	Europe
PYRS1780	20.85	SRR14750482	Pyrenean Shepherd	Europe

ROTT1208	32.44	SRR7120206	Rottweiler	Europe
SAAR1781	21.54	SRR14750487	Saarlooswolfdog	Europe
SALU1060	6.60	SRR5311685	Saluki	Middle East
SARP1737	22.49	SRR14750409	Sarplaninac	Europe
SCHP1624	23.29	SRR12330179	Schipperke	Europe
SCOT1299	12.35	SRR8541880	Scottish Terrier	Europe
SEAL1617	20.11	SRR12330202	Sealyham Terrier	Europe
SHAR1785	23.77	SRR14750491	Shar Pei	East Asia
SHIB0972	23.97	SRR7107933	Shiba Inu	East Asia
SHIH1215	17.27	SRR7120214	Shih Tzu	East Asia
SIKK1621	21.52	SRR12330188	Shikoku	East Asia
SLGI0318	27.48	SRR7107619	Sloughi	Middle East
SLHD1782	20.56	SRR14750488	Slovak Hound	Europe
SPAM1786	26.33	SRR14750493	Spanish Mastiff	Europe
SPIN1452	26.19	SRR10077544	Spinone Italiano	Europe
SSHP0901	25.11	SRR7107905	Shetland Sheepdog	Europe
SSNZ1012	38.98	SRR7107982	Standard Schnauzer	Europe
STAB1626	26.90	SRR12330156	Stabyhoun	Europe
STAF1625	23.43	SRR12330161	Staffordshire Bull Terrier	Europe
STBD0873	23.91	SRR7107879	Saint Bernard	Europe
SUSX1627	20.89	SRR12330155	Sussex Spaniel	Europe
SVAL1629	21.25	SRR12330149	Swedish Vallhund	Europe
TIBM0682	14.81	SRR1138364	Tibetan Mastiff	East Asia
TIBS1344	11.49	SRR8541936	Tibetan spaniel	East Asia
TIBT0894	22.55	SRR7107898	Tibetan Terrier	East Asia
TOSA1632	20.69	SRR12330143	Tosa	East Asia
TREE1634	20.69	SRR12330139	Treeing Walker Coonhound	America
TRSP1774	20.75	SRR14750351	Tatra Shepherd Dog	Europe
TRSV1616	24.12	SRR12330211	Teddy Roosevelt Terrier	America
TURV1137	16.07	SRR7120123	Belgian Tervuren	Europe

WEIM1055	10.92	SRR5311690	Weimaraner	Europe
WELT1636	22.38	SRR12330123	Welsh Terrier	Europe
WHIP1094	9.35	SRR5311633	Whippet	Europe
WHWT0883	27.51	SRR7107887	West Highland White Terrier	Europe
WOLF0779	6.14	SRR7107779	Grey Wolf	Europe
WOLF0788	5.61	SRR7107788	Grey Wolf	America
WOLF0840	16.50	SRR7107848	Grey Wolf	East Asia
WOLF0844	12.48	SRR7107852	Grey Wolf	East Asia
WOLF0909	26.10	SRR7107913	Grey Wolf	East Asia
WOLF1647	23.96	SRR13376369	Grey Wolf	Europe
WOLF1670	20.25	SRR13376343	Grey Wolf	Europe
WOLF1801	12.47	SRR8049197	Grey Wolf	America
WOLF1831	28.04	ERR4318106	Grey Wolf	Arctic
WSSD0322	16.37	ERR1688112	White Swiss Shepherd Dog	Europe
XISI0625	7.53	SRR7107674	Xiasi Dog	East Asia
YORK1263	43.87	SRR7764562	Yorkshire Terrier	Europe

Supplementary Table 4.4 Ancient DNA samples' information in BEAST analysis.

SAMPLE	Genus	Species	Material	Site	Country	Estimated YBP[1]	Accession
CTC	Canis	familiaris	Bone	Cherry Tree Cave (Kirschbaumhöhle)	Germany	4716	SAMN04884534
CGG29	Canis	lupus	Bone	Bunge-Toll-1885, Siberia	Russia	48210	SAMEA7538371
HXH	Canis	familiaris	Bone	Herxheim	Germany	7081	SAMN04884535
JK2183	Canis	lupus	N.A.	Höhle Fels	Germany	32366	N.A.
LOW002	Canis	lupus	N.A.	Letniaya River, Siberia	Russia	32781	N.A.
LOW003	Canis	lupus	N.A.	Unnegen site, Siberia (aka Bunge-Toll 1885)	Russia	44450	N.A.
NGDG	Canis	familiaris	Bone	Newgrange	Ireland	4800	PRJEB13070
PortauChoix	Canis	familiaris	N.A.	Port au Choix, Newfoundland	Canada	4157	SAMEA104190273
SC1061	Canis	familiaris	Bone	Chondorko	Ecuador	833	N.A.
SOTN01	Canis	familiaris	Bone	Sotin	Croatia	4900	N.A.
TRF.04.09	Canis	familiaris	Bone	Veretye	Russia	8750	N.A.

[1]YBP, years before present.

Supplementary Table 4.5 List of genes including coding, pseudogenes, and noncoding genes, and their annotation methods.

Scaffold	Gene name	Type	Annotation methods
chrY1	*TETY2*	Coding	PacBio
chrY1	*UTY*	Coding	Trinity
chrY1	*DDX3Y*	Coding	PacBio, Trinity
chrY1	*USP9Y*	Coding	Trinity
chrY1	*HSFY*	Coding	PacBio
chrY1	*EIF1AY*	Coding	PacBio, Trinity
chrY1	*KDM5D*	Coding	Trinity
chrY1	*ZFY*	Coding	Trinity
chrY1	*WWC3Y*	Coding	PacBio
chrY1	*EIF2S3Y*	Coding	Trinity
chrY1	*AMELY*	Coding	Splan
chrY1	*AP1S2Y*	Coding	PacBio, Trinity
chrY1	*TMSB4Y*	Coding	PacBio, Trinity
chrY1	*BCORY2*	Coding	PacBio
chrY1	*UBE1Y_1*	Coding	Splan
chrY1	*OFD1*	Coding	Trinity
chrY1	*TRAPPC2Y*	Coding	Trinity
chrY2	*UBE1Y_2*	Coding	PacBio
chrY2	*BCORY1*	Coding	PacBio, Trinity
chrY2	*CYorf15*	Coding	PacBio, Trinity
chrY2	*RBMYL*	Coding	PacBio
chrY2	*PRSS55Y*	Coding	PacBio
chrY2	*CUL4BY_1*	Coding	PacBio, Splan
chrY2	*TSPY_1*	Coding	PacBio
chrY2	*TSPYL_1*	Coding	Splan
chrY2	*TSPY_2*	Coding	Splan
chrY2	*TSPY_3*	Coding	Splan
chrY2	*TSPY_4*	Coding	Splan

chrY2	*TSPY_5*	Coding	PacBio, Splan
chrY2	*TSPYL_2*	Coding	Splan
chrY2	*SRY*	Coding	Trinity, Splan
chrY2	*CUL4BY_2*	Coding	Splan
chrY2	*CUL4BY_3*	Coding	Splan
chrY2	*TSPY_6*	Coding	Splan
chrY2	*CUL4BY_3*	Coding	PacBio, Trinity
chrY2	*TSPYL_3*	Coding	Splan
chrY1	*ATP5MGL*	pseudogene	Refseq liftover
chrY1	*VDAC3*	Pseudogene	Refseq liftover
chrY1	*PPP2CB*	Pseudogene	Refseq liftover
chrY1	*U6_spliceosomal*	Pseudogene	Refseq liftover
chrY1	*U4_spliceosomal*	Pseudogene	Refseq liftover
chrY1	*RPL23A*	Pseudogene	Refseq liftover
chrY1	*RPS29*	Pseudogene	Refseq liftover
chrY1	*CASP6*	Pseudogene	Refseq liftover
chrY1	*TXNDC12*	Pseudogene	Refseq liftover
chrY1	*LOC119868756*	LncRNA	Refseq liftover
chrY1	*LOC119868758*	LncRNA	Refseq liftover
chrY1	*lncRNA-PB13*	LncRNA	PacBio
chrY1	*DUF1725*	Pseudogene	PacBio
chrY1	*lncRNA-PB14*	LncRNA	PacBio
chrY1	*EEPD1*	Pseudogene	PacBio
chrY1	*GPS2*	Pseudogene	Trinity
chrY2	*LOC119868774*	LncRNA	Refseq liftover
chrY2	*LOC119868775*	LncRNA	Refseq liftover
chrY2	*MITF*	Pseudogene	PacBio, Trinity
chrY2	*PRDM5*	Pseudogene	Trinity
chrY2	*HP*	Pseudogene	Trinity
chrY2	*lnc-PB19*	LncRNA	PacBio
chrY2	*lnc-PB19*	LncRNA	PacBio

chrY3	*SRRM1*	Pseudogene	Refseq liftover
chrY3	*LOC119868722*	LncRNA	Refseq liftover
chrY3	*RPL10A*	Pseudogene	Refseq liftover
chrY3	*TSPYL4*	Pseudogene	Refseq liftover
chrY3	*LOC119868725*	LncRNA	Refseq liftover
chrY3	*LOC119868727*	LncRNA	Refseq liftover
chrY3	*RPL10A*	Pseudogene	Refseq liftover
chrY3	*RPL10A*	Pseudogene	Refseq liftover
chrY3	*RPL10A*	Pseudogene	Refseq liftover
chrY3	*RPL10A*	Pseudogene	Refseq liftover
chrY3	*NME7*	Pseudogene	Refseq liftover
chrY3	*RPL10A*	Pseudogene	Refseq liftover
chrY3	*NME7*	Pseudogene	Refseq liftover
chrY3	*LOC119868732*	LncRNA	Refseq liftover
chrY3	*LOC119868727*	LncRNA	Refseq liftover
chrY3	*RPL10A*	Pseudogene	Refseq liftover
chrY3	*RPL10A*	Pseudogene	Refseq liftover
chrY3	*HP*	Pseudogene	Trinity
chrY3	*lnc-PB29*	LncRNA	PacBio
chrY3	*lnc-PB33*	LncRNA	PacBio

Supplementary Table 4.6 Quatification of MSY gene expression (TPM) in 94 RNA-Seq samples across tissues.

The full table is available on https://github.com/WengangXbio/script_bio/blob/main/Supplementary%20Table%204.6.xlsx

Supplementary Table 4.7 Paralog comparison between Y-linked and X-linked genes.

Y-linked gene	species 1	species 2	Ka	Ks	Ka/Ks	P-Value(Fisher)
AP1S2Y	canine	fox	0	0	NA	0
BCORY2	canine	fox	0.025102	0.032158	0.780588	0.424979
CYorf15	canine	fox	0.015017	0.018828	0.797599	0.655027
DDX3Y	canine	fox	0.000654	0.020234	0.032311	1.21E-05
EIF1AY	canine	fox	0	0.0397	0	0
EIF2S3Y	canine	fox	0.003724	0.033414	0.111453	6.78E-05
HSFY	canine	fox	0.00751	0.028174	0.266567	0.015615
KDM5DY	canine	fox	0.003943	0.027606	1.43E-01	8.16E-11
OFD1Y	canine	fox	0.010437	0.013779	0.757457	0.599911
RBMYL	canine	fox	0.018012	0.035628	0.505568	0.138996
SRY	canine	fox	0.021874	0.045163	0.484329	0.096391
TMSB4Y	canine	fox	0.033374	0.061517	0.542522	0.173824
TRAPPC2Y	canine	fox	0.053836	9.28E-06	5800.26	0.124961
TSPY1	canine	fox	0.430817	1.43643	0.299922	6.24E-10
USP9Y	canine	fox	0.002142	0.032711	6.55E-02	1.38E-23
UTY	canine	fox	0.004538	0.021198	0.214066	1.20E-06
WWC3Y	canine	fox	0.021179	0.033296	0.636077	0.093837
ZFY	canine	fox	0.001072	0.02553	0.041988	1.83E-07
AMELY	canine	cat	0.043997	0.152283	0.288915	0.000123
CUL4BY	canine	cat	0.14511	0.263391	0.55093	5.59E-05
CYorf15	canine	cat	0.259458	0.321482	0.80707	0.527786
DDX3Y	canine	cat	0.019496	0.176567	0.110415	2.43E-26
EIF1AY	canine	cat	0.005793	0.183117	0.031638	2.34E-08
EIF2S3Y	canine	cat	0.005133	0.242071	2.12E-02	8.92E-34
HSFY	canine	cat	0.091917	0.250509	0.36692	2.35E-07

KDM5DY	canine	cat	0.052388	0.26633	1.97E-01	7.50E-59
RBMYL	canine	cat	0.086492	0.352018	0.245703	4.96E-15
SRY	canine	cat	0.209198	0.255456	0.81892	0.349887
TSPY1	canine	cat	0.620166	3.63464	1.71E-01	9.75E-15
UBE1Y1	canine	cat	0.038895	0.239402	0.162467	1.08E-39
UBE1Y2	canine	cat	0.035644	0.236824	0.150509	3.99E-42
USP9Y	canine	cat	0.0179	0.238181	7.52E-02	1.36E-137
UTY	canine	cat	0.033979	0.16475	0.206243	5.08E-30
ZFY	canine	cat	0.017544	0.247323	0.070935	2.22E-44
CYorf15	fox	cat	0.22989	0.463576	0.495906	0.011703
DDX3Y	fox	cat	0.020175	0.178246	0.113188	3.50E-26
EIF1AY	fox	cat	0.005764	0.197932	0.029121	8.83E-11
EIF2S3Y	fox	cat	0.009315	0.236216	0.039434	7.89E-30
HSFY	fox	cat	0.093214	0.252462	0.369217	3.43E-07
KDM5D	fox	cat	0.047533	0.277043	0.171573	1.80E-64
RBMYL	fox	cat	0.085653	0.310363	0.275976	1.25E-12
SRY	fox	cat	0.215234	0.270316	0.796231	0.301076
TMSB4Y	fox	cat	0.011083	0.502173	0.022071	3.29E-06
TSPY	fox	cat	1.02641	3.97569	0.258172	1.49E-06
USP9Y	fox	cat	0.017691	0.245829	0.071964	7.39E-144
UTY	fox	cat	0.033498	0.173086	0.193536	1.15E-32
ZFY	fox	cat	0.018882	0.224497	0.084108	2.79E-40

Supplementary Table 4.7 Ortholog comparison for MSY genes in dogs, foxes and cats.

species	Y-linked gene	X-linked gene	Ka	Ks	Ka/Ks	P-Value(Fisher)
canine	AMELY	AMELX	0.022101	0.092494	0.238947	0.000731
canine	AP1S2Y	AP1S2X	0.013758	0.280269	0.049089	3.06E-12
canine	BCORY1	BCORX	0.191261	0.573539	0.333475	1.17E-18
canine	BCORY2	BCORX	0.191261	0.573539	0.333475	1.17E-18
canine	CUL4BY	CUL4B	0.228629	1.21919	0.187526	3.91E-57
canine	CYorf15	TXLNG	0.148874	0.440795	0.337739	2.03E-12
canine	DDX3Y	DDX3X	0.037166	0.621055	5.98E-02	5.80E-79
canine	EIF1AY	EIF1AX	0.002915	0.855673	0.003406	2.27E-25
canine	EIF2S3Y	EIF2S3X	0.006638	0.434622	1.53E-02	3.63E-61
canine	HSFY	HSFX	0.52492	3.83993	0.1367	1.97E-46
canine	KDM5DY	KDM5CX	0.082464	0.774818	1.06E-01	1.75E-158
canine	OFD1Y	OFD1X	0.085673	0.094784	0.903878	0.503251
canine	RBMYL	RBMX	0.525184	3.52867	0.148833	1.28E-18
canine	SRY	SOX3	0.444046	3.57761	0.124118	4.44E-27
canine	TMSB4Y	TMSB4X	0.011623	0.392783	0.029591	4.39E-05
canine	TRAPPC2Y	TRAPPC2X	0.07593	0.128325	0.591705	0.323121
canine	TSPY1	TSPYL2(X)	0.61641	3.7168	1.66E-01	2.08E-22
canine	UBE1Y1	UBA1	0.053817	0.610874	0.088098	7.17E-97
canine	USP9Y	USP9X	0.026493	0.494633	0.053561	3.44E-246
canine	UTY	UTX	0.072248	0.283215	0.2551	2.11E-42
canine	WWC3Y	WWC3X	0.137118	0.475631	0.288286	4.01E-34
canine	ZFY	ZFX	0.016888	0.269393	0.062689	2.12E-54
cat	CUL4BY	CUL4B	0.153741	1.71695	8.95E-02	1.52E-82
cat	DDX3Y	DDX3X	0.022128	0.439057	0.050398	1.01E-71
cat	EIF2S3Y	EIF2S3	0.008663	0.335513	2.58E-02	7.10E-44

cat	KDM5D	KDM5C	0.068749	0.711845	0.096578	1.19E-142
cat	RBMYL	RBMX	0.49121	3.52968	0.139165	8.81E-18
cat	SRY	SOX3	0.479301	3.68662	0.130011	2.74E-25
cat	TSPY	TSPYL2(X)	0.836727	3.81691	2.19E-01	6.21E-17
cat	USP9Y	USP9X	0.031911	0.448486	0.071152	2.82E-235
cat	AMELY	AMELX	0.036887	0.08245	0.447379	0.044248
cat	CYorf15	TXLNG	0.309391	0.847165	0.365207	0.000143
cat	EIF1AY	EIF1AX	0.002917	0.770994	0.003784	2.10E-25
cat	HSFY	HSFX4	0.975532	4.16624	0.234152	2.67E-13
cat	UBE1Y	UBA1	0.047961	0.638665	0.075096	2.19E-120
cat	UTY	KDM6A	0.076233	0.331754	0.229789	1.64E-46
cat	ZFY	ZFX	0.017687	0.147842	0.119637	8.99E-28
fox	AP1S2Y	AP1S2	0.013758	0.280269	0.049089	3.06E-12
fox	BCORY2	BCOR	0.191483	0.643729	0.297458	3.08E-22
fox	DDX3Y	DDX3X	0.016839	0.566709	2.97E-02	3.29E-87
fox	EIF2S3Y	EIF2S3	0.011853	0.380419	0.031159	5.00E-51
fox	KDM5D	KDM5C	0.074546	0.742656	0.100378	1.43E-158
fox	RBMYL	RBMX	0.503287	3.54874	0.141822	3.88E-16
fox	SRY	SOX3	0.443278	3.45898	0.128153	3.14E-24
fox	TSPY	TSPYL2(X)	0.89541	4.00301	0.223685	4.35E-19
fox	USP9Y	USP9X	0.028286	0.453504	0.062372	3.05E-248
fox	CYorf15	TXLNG	0.156064	0.439073	0.35544	1.01E-11
fox	EIF1AY	EIF1AX	0.002896	0.69611	0.00416	5.63E-25
fox	HSFY	*LOC112912672*	1.17551	3.71427	0.316484	4.18E-07
fox	TMSB4Y	TMSB4X	0.011125	0.370817	0.03	3.67E-06
fox	TRAPPC2Y	TRAPPC2	0.090922	0.148095	0.613945	0.389655

fox	UTY	KDM6A	0.071121	0.291436	0.244035	1.02E-45
fox	ZFY	ZFX	0.018971	0.239091	0.079346	1.82E-44
fox	WWC3Y	WWC3	0.091442	0.364126	0.251128	1.77E-28
fox	OFD1Y	OFD1X	0.056341	0.106641	0.528321	0.197583

Supplementary Table 4.9 The number of heterozygosity sites for each gene locus and their occurrence in 222 male samples. The variants were called in GATK using the under biallelic model.

Gene	Heterozygosity sites	Heterozygosity occurrence
USP9Y	29	3579
CUL4BY_4	6	148
TSPY_6	5	88
TSPY_5	4	21
TSPY_3	4	11
DDX3Y	6	7
CUL4BY_1	2	6
TSPY_1	1	4
TSPY_2	2	4
UTY	2	2
KDM5D	2	2
TSPY_4	2	2
EIF1AY	1	1

Supplementary Table 4.10 103 variants located in coding regions and their annotations (synonymous and nonsynonymous). Ref, reference allele; Alt, alternative allele; MAF, minor allele frequency.

Gene	Position	Ref	Alt	Missing	MAF	Annotation	DNA substitutions	Protein substitution
BCORY1	112380	364	572	7	0.388889	missense_variant	c.1225G>C	p.Glu409Gln
WWC3Y	922372	693	245	5	0.261194	missense_variant	c.1010A>G	p.Lys337Arg
WWC3Y	944600	740	195	8	0.208556	missense_variant	c.1926C>G	p.Ile642Met
WWC3Y	909446	746	182	15	0.196121	missense_variant	c.807A>C	p.Glu269Asp
BCORY1	136501	768	171	4	0.182109	missense_variant	c.185A>T	p.Lys62Met
UBE1Y_1	1398612	753	147	43	0.163333	missense_variant	c.2531C>A	p.Ser844Tyr
OFD1	1449337	790	147	6	0.156884	missense_variant	c.479T>G	p.Phe160Cys
KDM5D	565472	114	820	9	0.122056	missense_variant	c.2615T>C	p.Val872Ala
WWC3Y	889304	893	28	22	0.030402	missense_variant	c.471A>T	p.Gln157His
WWC3Y	906732	905	24	14	0.025834	missense_variant	c.650T>C	p.Val217Ala
WWC3Y	857277	902	22	19	0.02381	missense_variant	c.341G>A	p.Arg114His
UBE1Y_1	1400004	21	915	7	0.022436	missense_variant	c.3065T>C	p.Leu1022Pro
CYorf15	217416	906	19	18	0.020541	missense_variant	c.998A>G	p.Asn333Ser
KDM5D	565442	910	18	15	0.019397	missense_variant	c.2645C>G	p.Ala882Gly
EIF2S3Y	715565	913	17	13	0.01828	missense_variant	c.917A>C	p.Asp306Ala
OFD1	1424796	922	14	7	0.014957	missense_variant	c.2600G>A	p.Arg867Lys
KDM5D	563389	919	12	12	0.012889	missense_variant	c.3533C>T	p.Ser1178Phe

OFD1	1468755	919	11	13	0.011828	missense_variant	c.190G>A	p.Val64Ile
WWC3Y	943721	925	10	8	0.010695	missense_variant	c.1814C>T	p.Ser605Leu
WWC3Y	857276	916	9	18	0.00973	missense_variant	c.340C>T	p.Arg114Cys
KDM5D	562921	924	9	10	0.009646	missense_variant	c.4001T>C	p.Met1334Thr
UBE1Y_2	66403	925	9	9	0.009636	missense_variant	c.3074G>A	p.Arg1025Gln
UBE1Y_2	72178	860	8	75	0.009217	missense_variant	c.1990C>T	p.Arg664Cys
AMELY	1098581	913	7	23	0.007609	missense_variant	c.323C>T	p.Pro108Leu
KDM5D	563293	922	7	14	0.007535	missense_variant	c.3629A>G	p.His1210Arg
KDM5D	592136	923	7	13	0.007527	missense_variant	c.398G>A	p.Arg133Gln
UBE1Y_2	66922	926	7	10	0.007503	missense_variant	c.2807C>A	p.Pro936His
WWC3Y	939708	918	6	19	0.006494	missense_variant	c.1651G>C	p.Gly551Arg
WWC3Y	939493	922	6	15	0.006466	missense_variant	c.1436T>C	p.Leu479Ser
UBE1Y_1	1392183	923	4	16	0.004315	missense_variant	c.1508C>T	p.Ala503Val
UBE1Y_1	1394233	868	3	72	0.003444	missense_variant	c.1974G>A	p.Met658Ile
UTY	184272	916	3	24	0.003264	missense_variant	c.110G>C	p.Ser37Thr
WWC3Y	960388	924	3	16	0.003236	missense_variant	c.2990T>C	p.Phe997Ser
UBE1Y_1	1392246	925	3	15	0.003233	missense_variant	c.1571A>G	p.Glu524Gly
WWC3Y	939427	925	3	15	0.003233	missense_variant	c.1370G>A	p.Arg457His
KDM5D	569038	926	3	14	0.003229	missense_variant	c.2393G>A	p.Arg798Lys

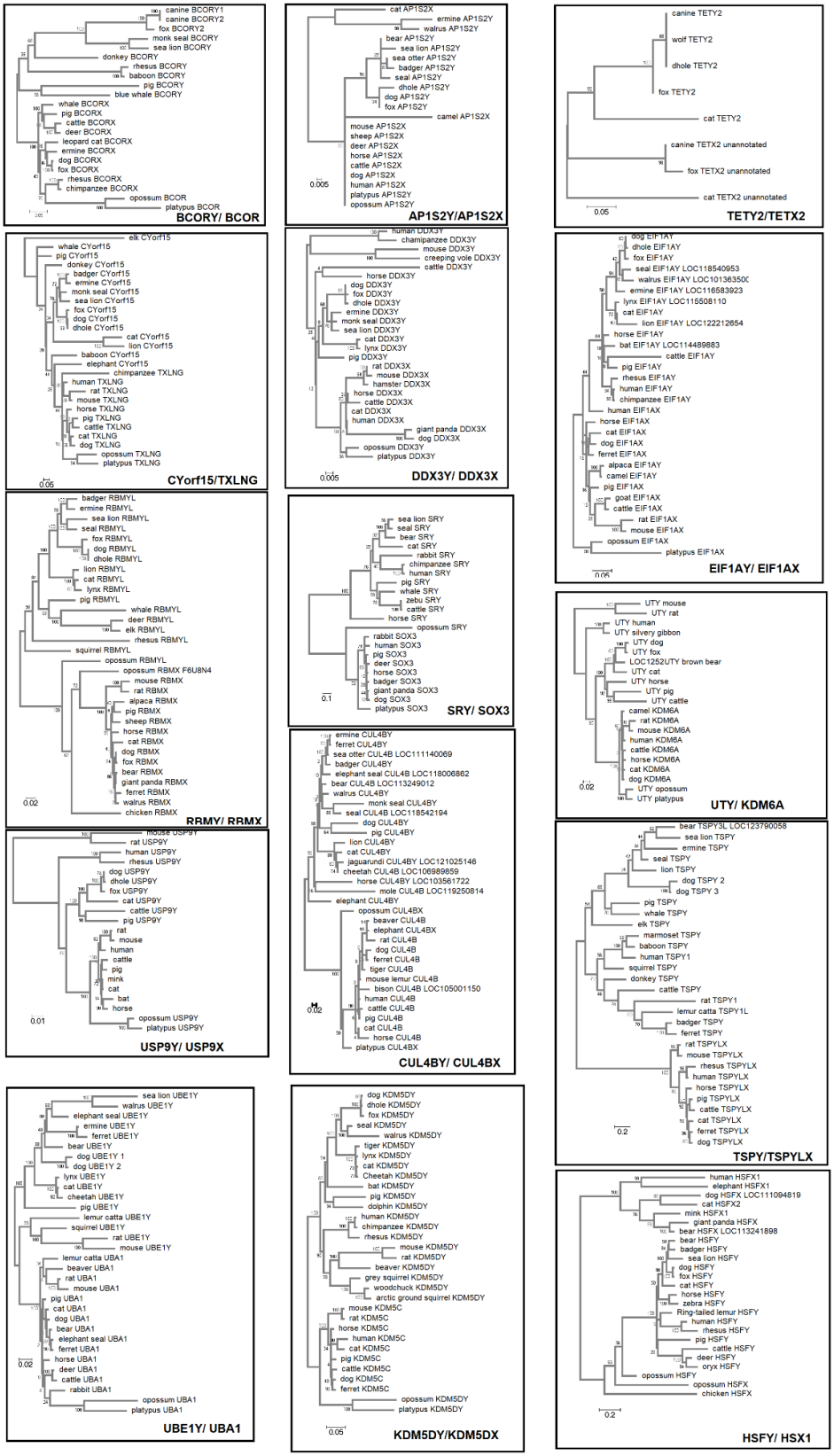
BCORY2	1349218	927	3	13	0.003226	missense_variant	c.1604G>A	p.Arg535Gln
KDM5D	565143	928	3	12	0.003222	missense_variant	c.2944G>A	p.Glu982Lys
OFD1	1427904	929	3	11	0.003219	missense_variant	c.2474A>G	p.Tyr825Cys
HSFY	463565	933	3	7	0.003205	missense_variant	c.47C>T	p.Ser16Leu
SRY	709678	938	3	2	0.003188	missense_variant	c.41T>C	p.Val14Ala
SRY	709194	940	3	0	0.003181	missense_variant	c.525G>C	p.Gln175His
CYorf15	250818	912	2	29	0.002188	missense_variant	c.169G>A	p.Val57Met
WWC3Y	979584	927	2	14	0.002153	missense_variant	c.3019G>A	p.Gly1007Arg
UBE1Y_2	73812	929	2	12	0.002148	missense_variant	c.1829A>G	p.Asn610Ser
CUL4BY_1	529838	931	2	10	0.002144	missense_variant	c.413A>G	p.His138Arg
USP9Y	255234	932	2	9	0.002141	missense_variant	c.6193C>T	p.Pro2065Ser
UBE1Y_1	1399487	933	2	8	0.002139	missense_variant	c.2795G>A	p.Arg932Gln
UBE1Y_1	1388835	871	1	71	0.001147	missense_variant	c.878A>G	p.Lys293Arg
WWC3Y	948399	920	1	22	0.001086	missense_variant	c.2452G>A	p.Glu818Lys
UBE1Y_1	1389438	923	1	19	0.001082	missense_variant	c.1225G>A	p.Ala409Thr
DDX3Y	218048	924	1	18	0.001081	missense_variant	c.76A>C	p.Asn26His
OFD1	1468662	929	1	13	0.001075	missense_variant	c.283G>A	p.Ala95Thr
DDX3Y	212233	931	1	11	0.001073	missense_variant	c.530C>T	p.Pro177Leu
KDM5D	563716	931	1	11	0.001073	missense_variant	c.3386C>T	p.Ala1129Val

OFD1	1433853	934	1	8	0.00107	missense_variant	c.1987C>T	p.Pro663Ser
TMSB4Y	1199456	934	1	8	0.00107	missense_variant	c.191C>T	p.Pro64Leu
USP9Y	335434	934	1	8	0.00107	missense_variant	c.1124C>A	p.Pro375His
BCORY2	1349385	936	1	6	0.001067	missense_variant	c.1771A>G	p.Asn591Asp
CUL4BY_1	512846	936	1	6	0.001067	missense_variant	c.1051G>A	p.Ala351Thr
DDX3Y	220447	729	194	20	0.210184	synonymous_variant	c.27G>A	NA
UBE1Y_2	80834	108	22	813	0.169231	synonymous_variant	c.216G>A	NA
UBE1Y_2	69966	788	142	13	0.152688	synonymous_variant	c.2241T>C	NA
UBE1Y_1	1398809	829	90	24	0.097933	synonymous_variant	c.2604C>G	NA
KDM5D	592657	913	25	5	0.026652	synonymous_variant	c.327C>T	NA
UBE1Y_2	66876	914	22	7	0.023504	synonymous_variant	c.2853G>A	NA
UBE1Y_1	1390188	910	21	12	0.022556	synonymous_variant	c.1488A>G	NA
EIF1AY	519349	905	18	20	0.019502	synonymous_variant	c.36G>A	NA
WWC3Y	979601	910	18	15	0.019397	synonymous_variant	c.3036C>T	NA
USP9Y	248454	911	18	14	0.019376	synonymous_variant	c.6795G>A	NA
KDM5D	563424	913	18	12	0.019334	synonymous_variant	c.3498A>G	NA
USP9Y	255274	913	17	13	0.01828	synonymous_variant	c.6153C>G	NA
AP1S2Y	1182451	916	13	14	0.013994	synonymous_variant	c.129T>C	NA
USP9Y	347627	919	12	12	0.012889	synonymous_variant	c.312T>C	NA

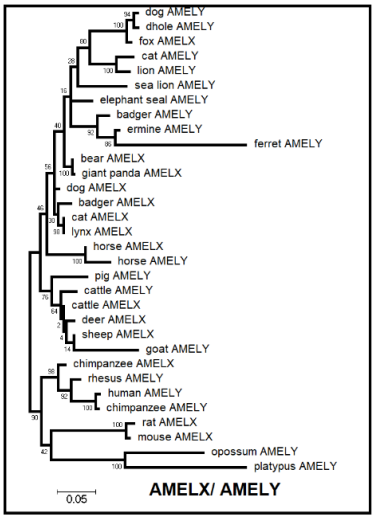
UBE1Y_1	1389108	917	10	16	0.010787	synonymous_variant	c.1023G>C	NA
WWC3Y	939719	914	9	20	0.009751	synonymous_variant	c.1662C>T	NA
UTY	45006	916	9	18	0.00973	synonymous_variant	c.3798G>A	NA
USP9Y	320566	925	9	9	0.009636	synonymous_variant	c.1863G>T	NA
UBE1Y_2	72098	876	8	59	0.00905	synonymous_variant	c.2070C>A	NA
WWC3Y	959321	923	7	13	0.007527	synonymous_variant	c.2874T>C	NA
EIF2S3Y	733300	925	7	11	0.007511	synonymous_variant	c.51T>C	NA
WWC3Y	886272	914	6	23	0.006522	synonymous_variant	c.411G>A	NA
DDX3Y	210954	920	6	17	0.006479	synonymous_variant	c.744C>G	NA
USP9Y	324705	927	6	10	0.006431	synonymous_variant	c.1623C>T	NA
UTY	35313	927	6	10	0.006431	synonymous_variant	c.4053A>T	NA
USP9Y	256034	929	6	8	0.006417	synonymous_variant	c.6003A>T	NA
UBE1Y_2	74117	931	4	8	0.004278	synonymous_variant	c.1635C>T	NA
UBE1Y_1	1388857	880	3	60	0.003398	synonymous_variant	c.900G>A	NA
UBE1Y_2	77214	914	3	26	0.003272	synonymous_variant	c.1047G>A	NA
BCORY2	1349216	928	3	12	0.003222	synonymous_variant	c.1602G>A	NA
USP9Y	272740	932	3	8	0.003209	synonymous_variant	c.4350T>C	NA
DDX3Y	208990	934	3	6	0.003202	synonymous_variant	c.1578A>G	NA
TETY2	10928	935	3	5	0.003198	synonymous_variant	c.39C>T	NA
DDX3Y	210995	920	2	21	0.002169	synonymous_variant	c.703C>T	NA
WWC3Y	939740	926	2	15	0.002155	synonymous_variant	c.1683C>A	NA

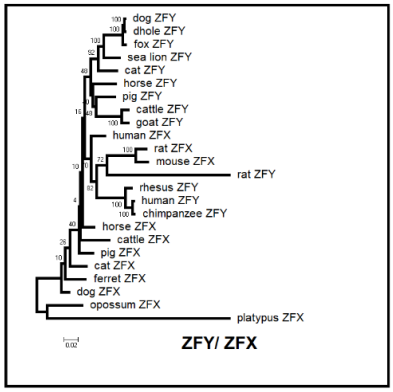
AP1S2Y	1173409	935	2	6	0.002134	synonymous_variant	c.393A>G	NA
KDM5D	562364	927	1	15	0.001078	synonymous_variant	c.4167G>A	NA
DDX3Y	210936	930	1	12	0.001074	synonymous_variant	c.762G>A	NA
CUL4BY_1	529846	932	1	10	0.001072	synonymous_variant	c.405A>G	NA
OFD1	1434016	932	1	10	0.001072	synonymous_variant	c.1824A>G	NA
UTY	82558	932	1	10	0.001072	synonymous_variant	c.1347T>G	NA
HSFY	464995	934	1	8	0.00107	synonymous_variant	c.1038A>G	NA
OFD1	1436878	934	1	8	0.00107	synonymous_variant	c.1494G>A	NA

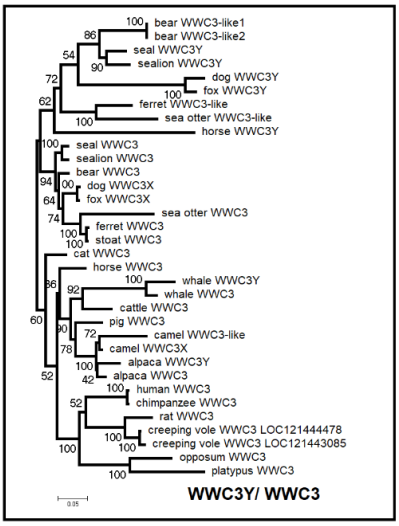
Supplementary Figure 4.1. Maximum likelihood phylogenies with bootstrap support values were constructed for monophyletic MSY genes. (A high resolution figure is available on https://github.com/WengangXbio/script_bio/blob/main/Supplementary%20Figure%204.1.png)

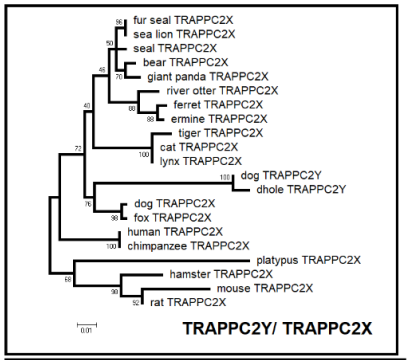


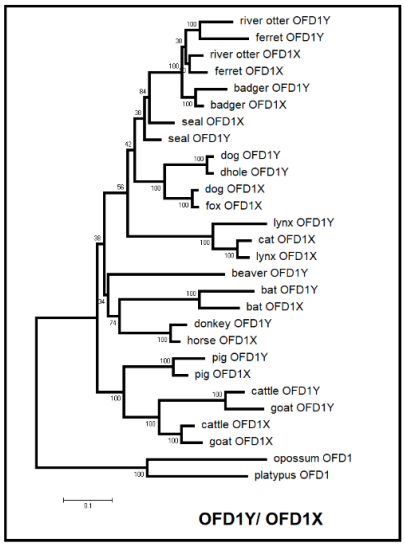
Supplementary Figure 4.2. Maximum likelihood phylogenies with bootstrap support values were constructed for polyphyletic MSY genes. (A high resolution figure is available on https://github.com/WengangXbio/script_bio/blob/main/Supplementary%20Figure%204.2.png)

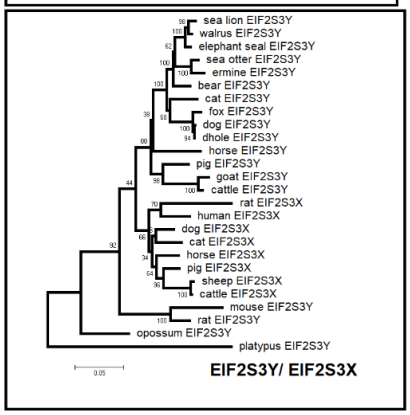


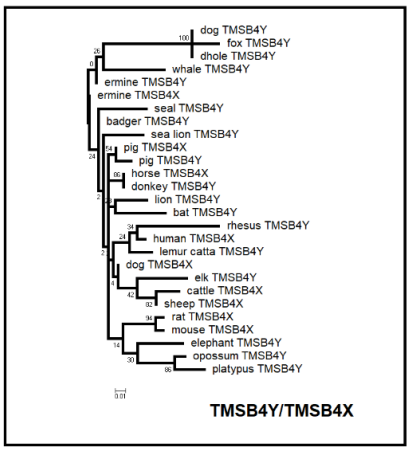












APPENDIX 4

Supplementary material in support of Chapter 5 of this thesis.

Supplementary Table 5.1 The study cohort for sex-specific variants analysis.

Accession	Sex	Coverage	Breed	Accession	Sex	Coverage	Breed
SRR7120237	Female	26.69	Yorkshire Terrier	SRR8614057	Male	34.04	Standard Schnauzer
ERR3339005	Female	29.76	Whippet	SRR8614022	Male	38.96	Standard Poodle
SRR7120234	Female	27.7	West Highland White Terrier	SRR10752624	Male	39.02	Standard Poodle
SRR7120232	Female	42.51	West Highland White Terrier	SRR7107874	Male	22.75	Standard Poodle
SRR7107890	Female	29.94	West Highland White Terrier	SRR7107900	Male	21.6	Standard Poodle
SRR7107889	Female	23.81	West Highland White Terrier	SRR8163600	Male	54.2	Standard Poodle
SRR7107888	Female	26.86	West Highland White Terrier	SRR8163597	Male	40.73	Standard Poodle
SRR7107538	Female	36.26	West Highland White Terrier	SRR8163595	Male	40.06	Standard Poodle
ERR1688126	Female	20.49	West Highland White Terrier	SRR8163591	Male	29.61	Standard Poodle
SRR8614024	Female	37.31	Welsh Terrier	SRR8163592	Male	44.95	Standard Poodle
SRR7107634	Female	22.97	Weimaraner	SRR8163590	Male	20.29	Standard Poodle

SRR10752641	Female	35.18	Weimaraner	SRR10077544	Male	26.19	Spinone Italiano
SRR10752640	Female	43.69	Weimaraner	SRR14750488	Male	20.56	Slovak Hound
SRR10441626	Female	53.03	Weimaraner	SRR7107619	Male	27.48	Sloughi
SRR2095540	Female	53.41	Toy Poodle	ERR1990016	Male	28	Siberian Husky
SRR10077569	Female	38.04	Tibetan Terrier	ERR3047551	Male	45.19	Shih Tzu
SRR7107983	Female	38.22	Standard Schnauzer	ERR3047552	Male	41.94	Shih Tzu
ERR2263462	Female	20.52	Standard Schnauzer	SRR8614050	Male	38.2	Shiba Inu
SRR8614062	Female	39.52	Standard Poodle	SRR10752637	Male	31.67	Shiba Inu
SRR8614047	Female	38.79	Standard Poodle	SRR7107933	Male	23.97	Shiba Inu
SRR8163602	Female	23.49	Standard Poodle	SRR7107905	Male	25.11	Shetland Sheepdog
SRR8163598	Female	55.98	Standard Poodle	SRR7120212	Male	43.66	Scottish Terrier
SRR8163596	Female	39.27	Standard Poodle	SRR7107893	Male	23.78	Scottish Deerhound
SRR8163594	Female	39.63	Standard Poodle	ERR2196280	Male	42.89	Saluki
SRR8163593	Female	55.98	Standard Poodle	ERR2750971	Male	26.73	Saint Bernard
SRR7107970	Female	44.14	Standard Poodle	ERR2750972	Male	23.14	Saint Bernard
SRR10441627	Female	46.29	Standard Poodle	SRR7107879	Male	23.91	Saint Bernard
SRR7107972	Female	48.51	Spinone Italiano	SRR7120205	Male	35.05	Rottweiler

SRR14750489	Female	22.26	Slovak Hound	SRR7120206	Male	32.44	Rottweiler
SRR7107584	Female	21.61	Sloughi	SRR10441631	Male	34.97	Rottweiler
SRR2095539	Female	62.35	Siberian Husky	SRR10441630	Male	24.74	Rottweiler
SRR7107655	Female	24.14	Shih Tzu	SRR8614052	Male	39.62	Rhodesian Ridgeback
ERR2196281	Female	39.32	Shih Tzu	SRR14750478	Male	25.86	Pyrenean Shepherd
SRR8614091	Female	52.26	Shiba Inu	SRR14750480	Male	23.28	Pyrenean Shepherd
SRR7107955	Female	36.96	Shiba Inu	SRR14750482	Male	20.85	Pyrenean Shepherd
SRR8614040	Female	35.35	Shetland Sheepdog	ERR2196264	Male	50.03	Pug
SRR7107550	Female	37.81	Shetland Sheepdog	ERR2196266	Male	55.16	Pug
SRR4011154	Female	52.33	Shetland Sheepdog	ERR2196269	Male	46.43	Pug
ERR2759431	Female	36.4	Shetland Sheepdog	ERR2196270	Male	50.98	Pug
SRR7120213	Female	24	Scottish Terrier	ERR2196271	Male	50.77	Pug
SRR7107886	Female	30.31	Scottish Terrier	ERR2196272	Male	49.19	Pug
SRR2094409	Female	49.85	Scottish Terrier	ERR2196276	Male	45.04	Pug
SRR10250963	Female	54.99	Scottish Deerhound	ERR2196277	Male	43.78	Pug
SRR10250962	Female	51.93	Scottish Deerhound	ERR2196278	Male	45.64	Pug
SRR10250961	Female	51.51	Scottish Deerhound	SRR10441632	Male	32.62	Pug
SRR10250960	Female	45.01	Scottish Deerhound	SRR7120197	Male	23.61	Portuguese Water Dog

SRR10250959	Female	59.23	Scottish Deerhound	SRR7120199	Male	21.73	Portuguese Water Dog
SRR10250958	Female	52.93	Scottish Deerhound	SRR7120201	Male	34.23	Portuguese Water Dog
SRR2095503	Female	64.11	Saluki	SRR8614036	Male	39.57	Pointer
SRR2095502	Female	46.9	Saint Bernard	SRR10077565	Male	22.85	Pit bull
SRR7120209	Female	51.01	Rottweiler	SRR10077562	Male	52.41	Pit bull
SRR7120208	Female	22.33	Rottweiler	ERR3047548	Male	42.47	Petit Basset Griffon Vendeen
SRR7120204	Female	31.54	Rottweiler	SRR7107576	Male	29.51	Petit Basset Griffon Vendeen
SRR7107864	Female	25.04	Rottweiler	SRR7107802	Male	24.43	Pembroke Welsh Corgi
SRR2095501	Female	43.88	Rottweiler	SRR10752627	Male	44.05	Pekingese
SRR7107985	Female	21.74	Rhodesian Ridgeback	SRR12330246	Male	24.21	Norwegian Elkhound
ERR1990011	Female	25.37	Rhodesian Ridgeback	SRR12330248	Male	22.83	Norwegian Elkhound
SRR10752644	Female	32.5	Pug	ERR3478972	Male	57.13	Norwegian buhund
SRR10077551	Female	58.44	Pug	SRR12330253	Male	23.37	Norwegian buhund
ERR3047544	Female	41.61	Pug	ERR2263461	Male	22.44	Newfoundland
ERR3047543	Female	41.84	Pug	ERR2750970	Male	23.05	Newfoundland
ERR3047542	Female	43.73	Pug	SRR12330252	Male	26.22	Neapolitan Mastiff

ERR2196279	Female	49.21	Pug	SRR12330251	Male	22.72	Neapolitan Mastiff
ERR2196275	Female	49.31	Pug	SRR14750473	Male	25.35	Neapolitan Mastiff
ERR2196273	Female	46.4	Pug	ERR2759433	Male	24.89	Mix
ERR2196268	Female	54.09	Pug	ERR2196028	Male	29.64	Mix
ERR2196267	Female	45.31	Pug	ERR2865340	Male	29.48	Miniature Schnauzer
ERR2196265	Female	50.44	Pug	ERR2008786	Male	33.61	Malinois dog
ERR2196095	Female	27.5	Pug	ERR2750973	Male	26.29	Malinois dog
SRR7120203	Female	21.75	Portuguese Water Dog	ERR5383440	Male	35.14	Malinois dog
SRR7120194	Female	54.56	Portuguese Water Dog	SRR12330273	Male	21.73	Lhasa Apso
SRR7120193	Female	42.68	Portuguese Water Dog	SRR12330272	Male	21.19	Lhasa Apso
SRR10077577	Female	28.52	Portuguese Water Dog	ERR2750954	Male	20.69	Leonberger
SRR7107877	Female	23.54	Pointer	ERR2750960	Male	23.09	Leonberger
SRR8614063	Female	30.02	Pit bull	ERR2750961	Male	23.05	Leonberger
SRR8614038	Female	28.5	Pit bull	ERR2750962	Male	21.75	Leonberger
ERR3047550	Female	44.3	Petit Basset Griffon Vendeen	ERR2750963	Male	24.23	Leonberger
ERR3047549	Female	42.65	Petit Basset Griffon Vendeen	ERR2750964	Male	21.96	Leonberger
SRR7107804	Female	22.78	Pembroke Welsh Corgi	ERR2750965	Male	25.93	Leonberger
SRR7107803	Female	24.06	Pembroke Welsh Corgi	ERR2750966	Male	22.99	Leonberger

SRR2095500	Female	59.95	Pekingese	ERR2750968	Male	26.16	Leonberger
SRR7120192	Female	39.36	Norwegian Elkhound	ERR2759446	Male	29.67	Leonberger
ERR3478973	Female	56.26	Norwegian buhund	ERR2764781	Male	34.93	Leonberger
SRR10077575	Female	36.46	Newfoundland	ERR4579527	Male	31.72	Leonberger
SRR14750472	Female	20.78	Neapolitan Mastiff	ERR4579528	Male	30.52	Leonberger
SRR10752625	Female	34.17	Mix	ERR5449478	Male	38.02	Leonberger
ERR2008436	Female	30.51	Mix	ERR5449479	Male	37.45	Leonberger
ERR1990029	Female	30.32	Mix	ERR5449483	Male	41.4	Leonberger
ERR1688118	Female	27.22	Mix	ERR5449484	Male	34.14	Leonberger
ERR2503985	Female	22.71	Miniature Schnauzer	ERR5449485	Male	40.29	Leonberger
ERR2503984	Female	30.92	Miniature Schnauzer	ERR5449486	Male	40.45	Leonberger
ERR2503983	Female	27.37	Miniature Schnauzer	ERR5449487	Male	30.23	Leonberger
ERR5383441	Female	26.49	Malinois dog	ERR3339011	Male	36.33	Lagotto Romagnolo
ERR1688123	Female	20.11	Malinois dog	ERR5383435	Male	32.43	Lagotto Romagnolo
SRR10752638	Female	33.16	Lhasa Apso	ERR5383437	Male	25.79	Lagotto Romagnolo
ERR5449482	Female	37.69	Leonberger	ERR5383439	Male	28.55	Lagotto Romagnolo
ERR5449481	Female	32.58	Leonberger	SRR7107608	Male	21.35	Lagotto Romagnolo
ERR5449480	Female	37.69	Leonberger	ERR5383433	Male	37.31	Labrador Retriever

ERR4579526	Female	25.64	Leonberger	SRR8614046	Male	34.32	Labrador Retriever
ERR2750967	Female	21.77	Leonberger	SRR8614090	Male	32.11	Labrador Retriever
ERR2357313	Female	23.84	Leonberger	SRR7107565	Male	41.73	Labrador Retriever
ERR5383438	Female	26.01	Lagotto Romagnolo	SRR7107937	Male	25.2	Labrador Retriever
ERR5383436	Female	28.48	Lagotto Romagnolo	SRR7107980	Male	37.47	Labrador Retriever
ERR5383434	Female	30.53	Lagotto Romagnolo	SRR7120177	Male	36.69	Labrador Retriever
ERR2759445	Female	29.5	Lagotto Romagnolo	SRR7120178	Male	29.99	Labrador Retriever
ERR2759439	Female	25.21	Lagotto Romagnolo	SRR7120180	Male	31.11	Labrador Retriever
ERR2113147	Female	21.09	Lagotto Romagnolo	SRR7120181	Male	30.86	Labrador Retriever
SRR7120176	Female	33.86	Labrador Retriever	SRR7120182	Male	40.88	Labrador Retriever
SRR7107934	Female	21.87	Labrador Retriever	SRR8541909	Male	21.26	Labrador Retriever
SRR7107891	Female	29.17	Labrador Retriever	SRR10752647	Male	42.64	Kerry Blue Terrier
SRR2095323	Female	32.6	Labrador Retriever	ERR2008784	Male	32.57	Jack Russell Terrier
ERR5383432	Female	27.58	Labrador Retriever	SRR8614044	Male	40.35	Jack Russell Terrier
SRR7107880	Female	24.65	Kerry Blue Terrier	SRR10752620	Male	46.99	Jack Russell Terrier
ERR5383431	Female	29.38	Kerry Blue Terrier	SRR7107878	Male	25	Jack Russell Terrier
ERR2196263	Female	39.44	Kerry Blue Terrier	SRR7107924	Male	24.88	Jack Russell Terrier
SRR7107896	Female	25.54	Jack Russell Terrier	ERR5383430	Male	28.35	Italian Greyhound

SRR7107892	Female	36.12	Jack Russell Terrier	SRR2094401	Male	23.5	Italian Greyhound
SRR10077557	Female	54.12	Jack Russell Terrier	SRR8614030	Male	25.11	Irish wolfhound
SRR10077556	Female	49.29	Jack Russell Terrier	SRR8541930	Male	24.18	Irish wolfhound
ERR3047540	Female	45.19	Jack Russell Terrier	SRR8541931	Male	24.67	Irish wolfhound
SRR2095322	Female	28.41	Italian Greyhound	SRR7120169	Male	55.19	Irish Water Spaniel
SRR2094400	Female	49.61	Irish wolfhound	SRR8614088	Male	27.71	Irish Terrier
SRR10441635	Female	39.51	Irish wolfhound	SRR10077568	Male	34.66	Irish Terrier
SRR10441634	Female	42.68	Irish wolfhound	SRR7107582	Male	20.11	Irish Terrier
SRR7120170	Female	52.9	Irish Water Spaniel	SRR7107979	Male	43.49	Irish Terrier
SRR7120168	Female	52.43	Irish Water Spaniel	SRR8614059	Male	42.14	Irish Setter
SRR10077553	Female	51	Irish Water Spaniel	SRR8614027	Male	40.11	Irish Setter
ERR2113152	Female	30.9	Irish Terrier	SAMC036703	Male	21.46	Indigenous dog
SRR8614086	Female	43.25	Irish Setter	SAMC036706	Male	24.81	Indigenous dog
SRR7120167	Female	47	Irish Setter	SAMC036709	Male	27.24	Indigenous dog
SRR7107649	Female	20.72	Indigenous Dog	SAMC036713	Male	20.77	Indigenous dog
SAMC036705	Female	20.58	Indigenous dog	SAMC036714	Male	24.12	Indigenous dog
ERR3047536	Female	44.33	Havanese	ERR3047537	Male	46.71	Havanese
SRR7107800	Female	39.86	Greyhound	ERR3047538	Male	46.04	Havanese

SRR7107799	Female	61.89	Greyhound	ERR2750980	Male	35.76	Greyhound
SRR7107795	Female	41.53	Greyhound	ERR2759443	Male	41.62	Greyhound
SRR7107794	Female	35.22	Greyhound	SRR7107789	Male	39.78	Greyhound
ERR5383429	Female	44.63	Greyhound	SRR7107790	Male	62.55	Greyhound
ERR2759442	Female	31.78	Greyhound	SRR7107791	Male	63.51	Greyhound
ERR2750981	Female	37.01	Greyhound	ERR2357315	Male	25.88	Greater Swiss Mountain Dog
ERR2357314	Female	27.43	Greater Swiss Mountain Dog	ERR2359900	Male	27.7	Greater Swiss Mountain Dog
ERR1683867	Female	23.15	Greater Swiss Mountain Dog	ERR2359901	Male	23.27	Greater Swiss Mountain Dog
SRR2095497	Female	59.58	Great Dane	ERR4579529	Male	27.33	Greater Swiss Mountain Dog
SRR11671233	Female	31.47	Great Dane	SRR7107969	Male	35.15	Greater Swiss Mountain Dog
SRR11671232	Female	33.08	Great Dane	SRR7120166	Male	42.1	Greater Swiss Mountain Dog
SRR11671231	Female	31.63	Great Dane	SRR12330315	Male	24.62	Greater Swiss Mountain Dog
SRR10441636	Female	44.35	Great Dane	SRR14750448	Male	25.81	Greater Swiss Mountain Dog

SRR12330331	Female	23.87	Grand Basset Griffon Vendeen	ERR5383420	Male	42.85	Great Dane
SRR12330330	Female	22.58	Grand Basset Griffon Vendeen	SRR7107534	Male	33.44	Great Dane
SRR12330329	Female	22.06	Grand Basset Griffon Vendeen	SRR7107967	Male	33.94	Great Dane
ERR4579525	Female	26.86	Grand Basset Griffon Vendeen	SRR11671234	Male	31.78	Great Dane
SRR8614060	Female	34.13	Golden Retriever	SRR11671229	Male	27.99	Great Dane
SRR8614020	Female	39.18	Golden Retriever	SRR11671228	Male	27.11	Great Dane
SRR7120162	Female	34.63	Golden Retriever	SRR14750298	Male	20.64	Grand Anglo-Français Tricolore
SRR7107857	Female	29.99	Golden Retriever	SRR14750300	Male	22.28	Grand Anglo-Français Tricolore
SRR7107854	Female	31.04	Golden Retriever	SRR14750301	Male	23.48	Grand Anglo-Français Tricolore
SRR10077559	Female	50.66	Golden Retriever	SRR14750297	Male	20.84	Grand Anglo-Français Tricolore
SRR10077549	Female	56.61	Golden Retriever	ERR1943610	Male	21.36	Golden Retriever
SRR10077542	Female	39.54	Golden Retriever	ERR1688124	Male	26.16	Golden Retriever
ERR5383428	Female	29.02	Golden Retriever	ERR1688854	Male	26.69	Golden Retriever
ERR1683868	Female	25.63	Golden Retriever	ERR2759434	Male	22.69	Golden Retriever
ERR3284983	Female	54.79	Giant Schnauzer	SRR8614034	Male	34.96	Golden Retriever

SRR4011155	Female	38.78	German Shepherd	SRR10077564	Male	47.62	Golden Retriever
SRR10752632	Female	49.95	German Shepherd	SRR10752631	Male	29.53	Golden Retriever
SRR10441638	Female	40.89	German Shepherd	SRR2095320	Male	39.45	Golden Retriever
SRR10077554	Female	31.25	German Shepherd	SRR7107793	Male	44.76	Golden Retriever
ERR3339004	Female	30.41	German Shepherd	SRR7107796	Male	43.79	Golden Retriever
ERR1688120	Female	28.57	German Shepherd	SRR7107797	Male	45.41	Golden Retriever
SRR7120158	Female	39.99	Flat-Coated Retriever	SRR7107798	Male	42.87	Golden Retriever
SRR2095487	Female	67.99	Flat-Coated Retriever	SRR7107855	Male	36.01	Golden Retriever
SRR2095480	Female	51.26	Flat-Coated Retriever	SRR7107856	Male	31.09	Golden Retriever
SRR10441648	Female	36.41	Flat-Coated Retriever	SRR7107894	Male	28.16	Golden Retriever
SRR10077567	Female	56.98	English Springer Spaniel	SRR7107926	Male	27.08	Golden Retriever
SRR8614054	Female	40.34	English Setter	SRR7120160	Male	35.46	Golden Retriever
SRR8614035	Female	36.05	English Setter	SRR7120161	Male	29.22	Golden Retriever
SRR7107962	Female	27.05	English Setter	ERR3284981	Male	56.83	Giant Schnauzer
SRR2094396	Female	70.46	English Mastiff	ERR3284984	Male	55.41	Giant Schnauzer
SRR7107606	Female	41.18	Dobermann	ERR3284982	Male	48.59	Giant Schnauzer
SRR2095318	Female	30.12	Dobermann	SRR8614029	Male	39.97	Giant Schnauzer

SRR10441641	Female	92.42	Dobermann	ERR2113150	Male	37.65	German Shepherd
SRR10077558	Female	53.22	Dobermann	ERR1816257	Male	21.01	German Shepherd
SRR10752623	Female	36.29	Dobermann	ERR5383424	Male	24.56	German Shepherd
SRR8614032	Female	26.83	Dalmatian	SRR8614058	Male	21.57	German Shepherd
SRR8614073	Female	50.83	Dachshund	SRR7107884	Male	26.44	German Shepherd
SRR10752643	Female	28.4	Dachshund	SRR7107915	Male	27.52	German Shepherd
SRR10752642	Female	38.12	Dachshund	SRR10752618	Male	37.58	Flat-Coated Retriever
SRR10752613	Female	37.44	Dachshund	SRR10441645	Male	53.12	Flat-Coated Retriever
SRR10441642	Female	32.96	Dachshund	ERR4579523	Male	21.78	English Springer Spaniel
ERR5383422	Female	35.34	Dachshund	SRR8614055	Male	36.87	English Springer Spaniel
ERR3047535	Female	43.31	Dachshund	SRR10077547	Male	25.54	English Springer Spaniel
SRR7107533	Female	21.19	Curly Coated Retriever	SRR7107883	Male	27.35	English Springer Spaniel
ERR2759441	Female	29.41	Curly Coated Retriever	SRR7107929	Male	27.64	English Springer Spaniel
ERR2750979	Female	35.88	Curly Coated Retriever	SRR7107930	Male	23.73	English Setter
SAMC045934	Female	20.7	Cretan Tracer	SRR7107977	Male	39.89	English Setter
SAMC045932	Female	22.65	Cretan Tracer	SRR10077576	Male	25.82	English Mastiff
SAMC045930	Female	21.55	Cretan Tracer	SRR10752629	Male	52.84	English Mastiff

SAMC045929	Female	20.7	Cretan Tracer	SRR14750465	Male	26.82	English Mastiff
SAMC045926	Female	20.6	Cretan Tracer	SRR8614053	Male	38.88	Dobermann
SAMC045925	Female	22.85	Cretan Tracer	SRR8614015	Male	21.73	Dobermann
SRR13376371	Female	20.7	Cretan Hound	SRR10752619	Male	33.91	Dobermann
SRR13376370	Female	21.55	Cretan Hound	SRR7107901	Male	27.47	Dobermann
SRR13376355	Female	22.65	Cretan Hound	SRR8614019	Male	44.58	Dalmatian
SRR13376352	Female	20.74	Cretan Hound	ERR3047534	Male	45.51	Dachshund
SRR13376337	Female	20.6	Cretan Hound	SRR10077566	Male	54.04	Dachshund
SRR13376353	Female	20.7	Cretan Hound	SRR7107544	Male	25.55	Dachshund
SRR5190663	Female	25.57	Collie	SRR7107965	Male	45.17	Dachshund
SRR5190660	Female	25.92	Collie	SRR7107971	Male	37.09	Dachshund
SRR5664965	Female	35.17	Cocker Spaniel	ERR2750978	Male	47.88	Curly Coated Retriever
SRR5664959	Female	33.94	Cocker Spaniel	ERR2759440	Male	32.24	Curly Coated Retriever
SRR5664958	Female	46.38	Cocker Spaniel	SAMC045924	Male	21.97	Cretan Tracer
SRR2095479	Female	52.72	Cocker Spaniel	SAMC045927	Male	21.17	Cretan Tracer
SRR10441651	Female	49.36	Cocker Spaniel	SAMC045928	Male	22.33	Cretan Tracer
SRR10441644	Female	48.76	Cocker Spaniel	SAMC045931	Male	20.71	Cretan Tracer

SRR10441643	Female	59.12	Cocker Spaniel	SRR13376336	Male	21.17	Cretan Hound
SRR2094392	Female	62.15	Chow Chow	SRR13376372	Male	22.33	Cretan Hound
SRR10752648	Female	36.42	Chinese Crested	SRR13376356	Male	20.71	Cretan Hound
SRR2095478	Female	52.06	Chihuahua	SRR13376354	Male	21.61	Cretan Hound
ERR3047532	Female	43.94	Chihuahua	ERR2759436	Male	28.02	Collie
ERR2750983	Female	26.23	Chihuahua	SRR5190662	Male	25.43	Collie
SRR8614039	Female	31.99	Chesapeake Bay Retriever	SRR5190661	Male	23.08	Collie
SRR7107547	Female	28.5	Cavalier King Charles Spaniel	SRR12330394	Male	23.39	Collie
SRR7107546	Female	30.72	Cavalier King Charles Spaniel	SRR12330392	Male	23.42	Collie
ERR2196025	Female	28.26	Cavalier King Charles Spaniel	SRR10752614	Male	27.9	Cocker Spaniel
SRR7107936	Female	26.98	Cane Corso	SRR5664961	Male	37.13	Cocker Spaniel
ERR5383418	Female	29.05	Cane Corso	SRR5664964	Male	49.63	Cocker Spaniel
ERR5383417	Female	29.65	Bullmastiff	SRR7107563	Male	23.94	Cocker Spaniel
SRR2095477	Female	49.91	Bulldog	SRR7107632	Male	26.6	Cocker Spaniel
SRR7120153	Female	88.68	Bull Terrier	SRR12330022	Male	23.23	Chow Chow
SRR7120148	Female	37.84	Bull Terrier	SRR12330021	Male	21.9	Chow Chow
ERR3047531	Female	43.61	Brussels Griffon	SRR7107775	Male	20.02	Chinese Crested

SRR7107773	Female	28.26	Boxer	SRR7107801	Male	29.21	Chinese Crested
ERR2196023	Female	29.01	Boxer	ERR3047533	Male	47.27	Chihuahua
SRR7120146	Female	62.23	Bouvier des Flandres	SRR7107579	Male	31.11	Chihuahua
SRR7120145	Female	51.35	Boston Terrier	SRR10752621	Male	37.69	Chesapeake Bay Retriever
SRR10752636	Female	32.15	Boston Terrier	SRR7107974	Male	39.06	Chesapeake Bay Retriever
SRR7107927	Female	27.35	Border Terrier	SRR10441637	Male	32.81	Cavalier King Charles Spaniel
SRR7107899	Female	22.17	Border Terrier	SRR8614031	Male	32.81	Cane Corso
SRR7107950	Female	23.68	Border Collie	SRR8614037	Male	32.85	Cane Corso
SRR7107590	Female	23.22	Border Collie	SRR8614017	Male	38.5	Cane Corso
SRR7107578	Female	30.08	Border Collie	SRR10077570	Male	35.8	Cane Corso
SRR7107554	Female	26.43	Border Collie	ERR2196024	Male	28.42	Bullmastiff
SRR7107535	Female	30.62	Border Collie	SRR12330047	Male	23.84	Bullmastiff
SRR10752615	Female	33.73	Border Collie	ERR2750977	Male	29.71	Bulldog
SRR10077552	Female	44.97	Border Collie	ERR2196103	Male	51.11	Bull Terrier
ERR3339008	Female	38.13	Border Collie	SRR7120149	Male	42.57	Bull Terrier
SRR12330075	Female	22.5	Boerboel	SRR7120150	Male	49.21	Bull Terrier
SRR2095469	Female	49.94	Bloodhound	SRR7120151	Male	47.54	Bull Terrier

SRR7120127	Female	48.24	Bernese Mountain Dog	SRR7120152	Male	60.57	Bull Terrier
SRR2095463	Female	48.82	Bernese Mountain Dog	SRR12330046	Male	21.18	Bull Terrier
ERR5383414	Female	34.03	Bernese Mountain Dog	ERR3047522	Male	46.45	Brussels Griffon
ERR5383407	Female	27.44	Bernese Mountain Dog	ERR2196102	Male	49.76	Brussels Griffon
SRR7107932	Female	27.97	Berger Picard	ERR5383419	Male	27.16	Boxer
SRR7107921	Female	23.93	Berger Picard	SRR8541918	Male	20.86	Boxer
SRR7107897	Female	23.19	Berger Picard	SRR8541911	Male	20	Boxer
SRR10077571	Female	38.1	Berger Picard	SRR10441650	Male	52.14	Boxer
SRR2095368	Female	55.67	Belgian Shepherd	SRR10441649	Male	34.95	Boxer
SRR7107642	Female	45.07	Beagle	SRR12330069	Male	23.69	Boxer
SRR2094386	Female	21.89	Beagle	SRR8614087	Male	36.32	Bouvier des Flandres
ERR3026883	Female	24.73	Beagle	SRR7107966	Male	41.22	Bouvier des Flandres
SRR10077555	Female	55.84	Basset Hound	SRR7120147	Male	21.72	Bouvier des Flandres
ERR3047521	Female	41.68	Basset Hound	SRR10752622	Male	40	Boston Terrier
ERR1688855	Female	22.95	Basset Hound	SRR7107928	Male	26.97	Border Terrier
SRR8614023	Female	38.92	Basenji	SRR7107968	Male	43.76	Border Terrier
SRR7107885	Female	24.16	Basenji	ERR2008772	Male	26.13	Border Collie

SRR10752639	Female	31.34	Basenji	ERR2008775	Male	23.46	Border Collie
SRR10752616	Female	39.32	Basenji	SRR2094388	Male	26.43	Border Collie
ERR2113155	Female	25.44	Basenji	SRR7107946	Male	24.58	Border Collie
ERR2113154	Female	25.56	Basenji	SRR7107947	Male	23.3	Border Collie
ERR2113153	Female	24.58	Basenji	SRR7107948	Male	25.56	Border Collie
SRR7107923	Female	21.44	Australian Cattle Dog	SRR7107949	Male	25.72	Border Collie
SRR7107867	Female	24.64	Australian Cattle Dog	SRR12330074	Male	24.14	Boerboel
ERR5383406	Female	24.73	Australian Cattle Dog	SRR14750521	Male	23.57	Bloodhound
ERR3339007	Female	25.28	Australian Cattle Dog	SRR14750523	Male	21.86	Bloodhound
SRR14750434	Female	20.56	Anglo-Français de Petite Vénerie	ERR5383408	Male	29.6	Bernese Mountain Dog
SRR8614051	Female	38.07	American Staffordshire Terrier	ERR5383409	Male	39.28	Bernese Mountain Dog
SRR10077573	Female	48.02	American Staffordshire Terrier	ERR5383410	Male	26.04	Bernese Mountain Dog
SRR10077543	Female	55.97	American Staffordshire Terrier	ERR5383411	Male	32.85	Bernese Mountain Dog
SRR7107992	Female	45.01	Alaskan Malamute	ERR5383412	Male	23.95	Bernese Mountain Dog
ERR2750976	Female	36.7	Alaskan Malamute	ERR5383413	Male	30.78	Bernese Mountain Dog
ERR2196097	Female	41.18	Airedale Terrier	ERR5383415	Male	27.69	Bernese Mountain Dog
SRR7107882	Female	30.88	Airedale Terrier	ERR5383416	Male	30.16	Bernese Mountain Dog

SRR10758785	Female	34.39	Airedale Terrier	SRR10441652	Male	39.86	Bernese Mountain Dog
SRR7107586	Male	20.85	Yorkshire Terrier	SRR7107973	Male	57.61	Berger Picard
SRR7107916	Male	31.08	Yorkshire Terrier	SRR14750335	Male	21.18	Berger Picard
SRR7764562	Male	43.87	Yorkshire Terrier	SRR7120114	Male	23.06	Belgian Shepherd
ERR2196282	Male	51.29	Whippet	ERR3026880	Male	26.36	Beagle
SRR10752617	Male	36.96	Whippet	SRR2094385	Male	52.42	Beagle
ERR2750975	Male	33.55	West Highland White Terrier	SRR7107976	Male	38.91	Beagle
ERR5383443	Male	24.64	West Highland White Terrier	SRR8614061	Male	21.36	Basset Hound
ERR5383444	Male	28.96	West Highland White Terrier	SRR12330158	Male	25.94	Basset Hound
ERR5383445	Male	30.68	West Highland White Terrier	ERR2113156	Male	24.3	Basenji
SRR7107887	Male	27.51	West Highland White Terrier	ERR2113157	Male	20.06	Basenji
SRR12330123	Male	22.38	Welsh Terrier	SRR8614076	Male	42.46	Basenji
SRR12330122	Male	23.41	Welsh Terrier	SRR8614026	Male	42.97	Basenji
ERR2113151	Male	27.54	Weimaraner	SRR7107984	Male	39.06	Basenji
SRR10752626	Male	51.72	Toy Poodle	ERR2113149	Male	37.81	Australian Cattle Dog
SRR8614028	Male	39.39	Tibetan Terrier	ERR2263463	Male	21.97	Australian Cattle Dog
SRR8614021	Male	40.22	Tibetan Terrier	SRR7107537	Male	25.64	Australian Cattle Dog

SRR10077572	Male	49.07	Tibetan Terrier	SRR12330225	Male	21.42	Australian Cattle Dog
SRR7107895	Male	25.49	Tibetan Terrier	SRR14750293	Male	23.3	Anglo-Français de Petite Vénerie
SRR7107898	Male	22.55	Tibetan Terrier	SRR14750304	Male	23.59	Anglo-Français de Petite Vénerie
SRR8614056	Male	47.53	Standard Schnauzer	ERR2759437	Male	24.97	American Staffordshire Terrier
SRR7107964	Male	41.3	Standard Schnauzer	SRR8614085	Male	34.88	Alaskan Malamute
SRR7107982	Male	38.98	Standard Schnauzer	SRR7107630	Male	24.42	Airedale Terrier
SRR10758786	Male	26.92	Airedale Terrier	SRR7107922	Male	22.57	Airedale Terrier

Supplementary Table 5.2 Primer sequences for amplifying segments spanning PAB SINEs. The difference of PCR productions in length between males and females due to the existence of PAB SINEs on the Y chromosome.

Position (Rosy_1.0)	Position (X)	Forward Primer	Products length (Rosy_1.0)	Products length (X chromosome)	Annealing temperature
748	6591392	CAGACCTAGCTGCTGTACTGCCAC	607	403	68
1354	6591794	GCTATCCGGCAGGCCTCGCCTTTG			
3445	6593900	CATGCGCACCTTGCCACAGCTTACTTC	659	453	68
4103	6594352	GTCCCCACATTCTGTTACCCATTGTCC			

Supplementary Table 5.3 TPM value for *TETY2* and *CLDN34* across 94 RNA-Seq samples, and the abundance of expression at the exon level.

Accession	Tissue	TETY2 total	CLDN34-exon1	CLDN34-exon2	CLDN34-exon3	CLDN34-exon4	CLDN34-exon5	CLDN34-exon6	CLDN34 total
CKCS.SID00012	Testis	110.05	0.00	0.00	0.73	26.03	38.69	97.74	29.81
FBUL.SID00016	Testis	95.87	0.09	0.00	0.00	27.25	31.68	89.81	26.93
ITGY.SID00003	Testis	70.58	0.00	0.00	0.00	23.29	30.67	62.78	21.07
PAPI.SID00019	Testis	52.15	0.86	0.46	0.74	16.00	21.82	58.71	18.06
PUG.SID00020	Testis	37.67	0.07	0.00	0.00	15.17	11.24	38.41	11.90
YORK.SID00007	Testis	17.34	0.00	0.00	0.77	12.48	5.55	20.07	7.04
SRR8996995	Adrenal gland	0.60	0.00	0.00	0.00	0.00	0.85	0.57	0.24
SRR8997050	Stomach	0.41	2.88	4.00	0.71	0.00	1.60	6.71	2.87
BULD.SID00010	Testis	0.40	0.24	0.52	0.00	0.47	1.50	1.31	0.72
CKCS.SID00017	Testis	0.40	0.29	0.46	0.00	0.41	1.33	2.23	0.83
SRR5889319	Heart	0.37	2.31	1.10	2.65	0.98	2.38	2.81	2.11
SRR8996988	Skin	0.34	3.20	0.96	3.08	0.00	0.34	2.68	1.89
SRR8997035	Cerebellum	0.34	2.91	2.50	4.00	1.11	0.90	2.97	2.48
SRR8997041	Lung	0.33	8.69	5.44	8.27	3.15	3.71	6.07	6.13
SRR8997021	Cerebellum	0.31	3.56	3.62	0.97	1.08	0.00	3.72	2.34
SRR8997033	Cerebellum	0.31	4.10	3.77	2.01	4.48	0.90	5.08	3.55
SRR8996961	Cerebellum	0.28	3.59	1.98	0.53	0.00	2.14	5.03	2.46
SRR8996953	Kidney cortex	0.28	3.40	2.09	2.09	1.16	0.75	2.76	2.17
SRR8996957	Adrenal gland	0.27	1.76	3.30	1.32	0.92	1.33	3.45	2.11
SRR8997034	Colon	0.25	2.02	0.65	1.04	1.74	1.64	3.45	1.87
SRR8997004	Pituitary gland	0.24	3.72	1.71	0.92	0.51	1.64	3.52	2.24
SRR5889317	Head	0.24	5.74	2.42	4.43	1.54	1.99	6.72	4.10
SRR8996972	Salivary gland	0.23	1.02	0.94	0.75	0.00	1.18	1.86	1.02
SRR8996966	Liver	0.22	0.04	0.46	0.00	0.00	0.17	0.04	0.11
SRR5889328	Kidney_cortex	0.22	3.57	3.54	2.12	0.59	3.02	4.64	3.10
SRR8996981	Pituitary gland	0.21	2.27	3.22	2.34	0.26	0.42	2.76	1.98
SRR8997032	Spleen	0.20	0.52	0.28	0.00	0.00	0.40	0.27	0.26
SRR8997019	Adrenal gland	0.19	3.69	2.53	0.00	0.90	1.45	4.97	2.55
SRR8997016	Cartilage	0.19	0.74	2.37	0.00	0.42	0.00	0.18	0.64
SRR8996959	Cerebellum	0.19	3.86	2.19	3.50	2.68	1.38	5.37	3.40
SRR8996965	Lung	0.17	1.51	0.33	0.00	0.60	0.00	1.61	0.79
SRR5889320	Kidney cortex	0.17	5.45	1.85	2.54	0.94	2.09	5.03	3.26
SRR8997047	Adrenal gland	0.16	5.14	4.26	3.67	1.46	2.12	6.12	4.02
SRR8997055	Skin	0.15	2.20	1.99	1.19	0.44	1.61	2.77	1.82
SRR8997010	Lung	0.15	6.13	1.76	4.92	1.57	1.26	6.08	3.94
SRR8997053	Salivary gland	0.15	2.57	2.05	1.44	0.69	1.20	2.81	1.91
SRR8996967	Pancreas	0.13	0.00	0.00	0.00	0.00	0.21	0.14	0.06
SRR8996987	Small intestine	0.11	1.18	0.00	0.00	0.48	0.78	2.50	0.96

SRR5889327	Lung	0.11	3.31	1.90	5.41	1.32	2.58	3.94	3.19
SRR8997024	Skin	0.10	3.21	1.42	1.81	1.26	1.63	3.16	2.20
SRR8996996	Bladder	0.10	1.01	0.00	0.00	0.00	0.00	1.03	0.41
SRR8997005	Kidney cortex	0.09	2.92	0.97	0.77	0.86	1.74	2.51	1.77
SRR8997001	Kidney cortex	0.09	0.80	0.00	0.00	0.00	0.19	1.69	0.52
SRR8996955	Bone marrow	0.09	0.03	0.00	0.00	0.00	0.00	0.07	0.02
SRR8996980	Lung	0.08	1.04	1.30	0.00	1.02	1.40	1.94	1.17
SRR8997040	Kidney cortex	0.08	1.94	1.92	1.54	1.43	1.61	3.45	2.07
SRR8996970	Pituitary gland	0.08	0.99	0.00	0.43	0.00	0.38	0.92	0.52
SRR8996968	Lymph node	0.07	0.26	0.00	0.00	0.00	0.00	0.32	0.11
SRR8997000	Colon	0.06	0.21	0.00	0.00	0.00	0.00	0.06	0.05
SRR8996960	Colon	0.06	1.65	1.58	1.45	1.21	1.30	1.78	1.54
SRR8996977	Liver	0.06	0.39	0.00	0.00	0.15	0.24	0.55	0.25
SRR8996990	Salivary gland	0.05	1.50	0.90	0.00	0.00	0.00	1.68	0.79
SRR8997056	Small intestine	0.05	0.99	0.61	0.00	0.81	0.22	0.93	0.63
SRR8997036	Bone marrow	0.05	0.69	1.27	0.51	0.57	0.23	1.59	0.84
SRR8997026	Salivary gland	0.05	1.70	1.37	1.56	1.04	1.68	1.88	1.60
SRR8996956	Bladder	0.05	1.02	0.27	0.88	0.24	0.59	1.85	0.90
SRR8997052	Adipose	0.04	0.82	0.29	0.00	0.26	0.21	0.56	0.40
SRR8997045	Stomach	0.04	0.34	0.00	0.64	0.71	0.00	0.54	0.39
SRR5889318	Lung	0.03	3.19	0.31	3.00	0.56	2.46	2.76	2.18
SRR8997015	Cerebellum	0.00	1.90	1.11	3.57	1.99	0.80	2.25	2.00
SRR8997031	Small intestine	0.00	2.51	0.34	1.10	0.92	0.49	2.98	1.53
SRR5889322	Liver	0.00	0.73	1.08	1.44	0.80	0.78	0.97	0.96
SRR5889325	Heart	0.00	1.56	0.84	0.67	0.75	1.20	2.25	1.32
SRR8997020	Adipose	0.00	0.26	0.00	1.33	0.74	0.59	0.32	0.53
SRR8997018	Stomach	0.00	2.61	0.59	0.94	0.70	0.98	1.81	1.38
SRR8997048	Bladder	0.00	0.73	1.00	0.00	0.60	0.48	1.54	0.77
SRR8997027	Pancreas	0.00	0.63	0.44	0.23	0.52	0.42	0.78	0.53
SRR8997042	Liver	0.00	0.95	1.59	1.28	0.47	0.57	0.87	0.97
SRR8996999	Cerebellum	0.00	1.15	0.46	0.74	0.41	1.00	1.25	0.88
SRR8997009	Liver	0.00	1.18	0.00	0.00	0.32	0.00	0.63	0.42
SRR8997029	Lymph_node	0.00	0.22	1.09	0.00	0.32	0.52	0.28	0.40
SRR8996983	Adipose	0.00	0.81	0.33	0.00	0.29	0.24	0.76	0.46
SRR8996979	Lymph node	0.00	0.22	0.00	0.00	0.21	0.17	0.54	0.22
SRR8996958	Adipose	0.00	0.63	0.00	0.00	0.18	0.00	0.47	0.25
SRR5889330	Liver	0.00	0.40	0.18	0.29	0.16	0.00	0.35	0.25
SRR8996989	Skeletal muscle	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SRR8996994	Cartilage	0.00	0.00	0.59	0.47	0.00	0.21	0.62	0.31
SRR8996993	Bone marrow	0.00	0.03	0.00	0.00	0.00	0.56	0.26	0.15
SRR8997023	Skeletal_muscle	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.01
SRR8997054	Skeletal_muscle	0.00	0.07	0.00	0.00	0.00	0.00	0.09	0.03

SRR8997044	Spleen	0.00	0.08	0.00	0.00	0.00	0.00	0.39	0.09
SRR8996971	Skeletal_muscle	0.00	0.09	0.00	0.00	0.00	0.00	0.06	0.03
SRR8997043	Skin	0.00	0.09	0.30	0.00	0.00	0.65	0.29	0.23
SRR8996986	Spleen	0.00	0.13	0.41	0.66	0.00	0.00	0.08	0.21
SRR8997011	Lymph_node	0.00	0.14	0.00	0.00	0.00	0.00	0.61	0.15
SRR8997049	Spleen	0.00	0.20	0.00	0.00	0.00	0.60	0.20	0.18
SRR8996997	Cerebellum	0.00	0.31	0.00	1.06	0.00	0.24	0.57	0.37
SRR8996962	Cartilage	0.00	0.39	0.00	0.40	0.00	0.00	0.34	0.21
SRR8997003	Pancreas	0.00	0.54	0.00	0.00	0.00	0.18	0.24	0.18
SRR8997013	Bone_marrow	0.00	0.63	0.00	0.00	0.00	0.36	0.68	0.33
SRR8996985	Stomach	0.00	0.66	0.48	0.00	0.00	0.34	0.55	0.38
SRR8997014	Bladder	0.00	1.13	0.97	0.00	0.00	0.70	1.03	0.70
SRR8997022	Colon	0.00	1.47	0.00	2.89	0.00	0.00	2.00	1.16

Supplementary Table 5.4 TPM value for *TETY2* and *CLDN34* in 13 female samples, the abundance of expression at the exon level.

Accession	Tissue	*CLDN34*-exon1	*CLDN34*-exon2	*CLDN34*-exon3	*CLDN34*-exon4	*CLDN34*-exon5	*CLDN34*-exon6	*CLDN34*_total
ERR2034262	Unknown	4.12	3.80	3.04	1.69	1.82	5.97	3.41
SRR2960340	Temporal lobe	7.05	9.32	0.00	6.75	3.77	7.40	5.71
SRR2960344	Occipital lobe	13.04	13.37	0.00	6.24	0.92	1.47	5.84
SRR2960345	Temporal lobe	9.74	7.44	1.70	4.74	1.91	4.50	5.01
SRR2960346	Occipital lobe	17.34	18.50	5.15	16.49	5.78	13.00	12.71
SRR5889315	Liver	3.77	1.84	1.18	0.33	2.64	4.85	2.43
SRR5889316	Heart	5.26	1.38	5.18	0.62	2.32	5.60	3.39
SRR5889323	Lung	8.61	3.15	5.99	3.33	4.53	6.82	5.41
SRR5889324	Head	10.25	3.10	6.11	3.83	4.80	12.17	6.71
SRR5889331	Kidney	9.32	3.93	3.14	2.27	3.95	10.35	5.49
SRR5889332	Heart	2.35	1.93	5.35	0.94	1.64	3.11	2.55
SRR5889333	Lung	6.16	2.88	3.29	2.20	5.02	6.69	4.37
SRR5889334	Liver	1.02	0.17	0.81	0.15	0.85	1.07	0.68

APPENDIX 5

Supplementary material in support of Chapter 6 of this thesis.

Supplementary Table 6.1 List of available Iso-Seq data for testis in mammals.

Accession	Taxon name	Common name	Age
ERR9764402	Equus caballus	Horse	3 Y
ERR9764403	Equus caballus	Horse	4 Y
SRR10821774	Ovis aries	Sheep	2 Y
SRR11232018	Callithrix jacchus	Common marmoset	1.5 Y
SRR11570895	Trichosurus vulpecula	Common brushtail possum	Adult
SRR5571257	Bos taurus	Cattle	9 Y
SRR17664326	Sus scrofa	Pig	Adult
SRR18652230	Molossus molossus	Velvety free-tailed bat	Adult

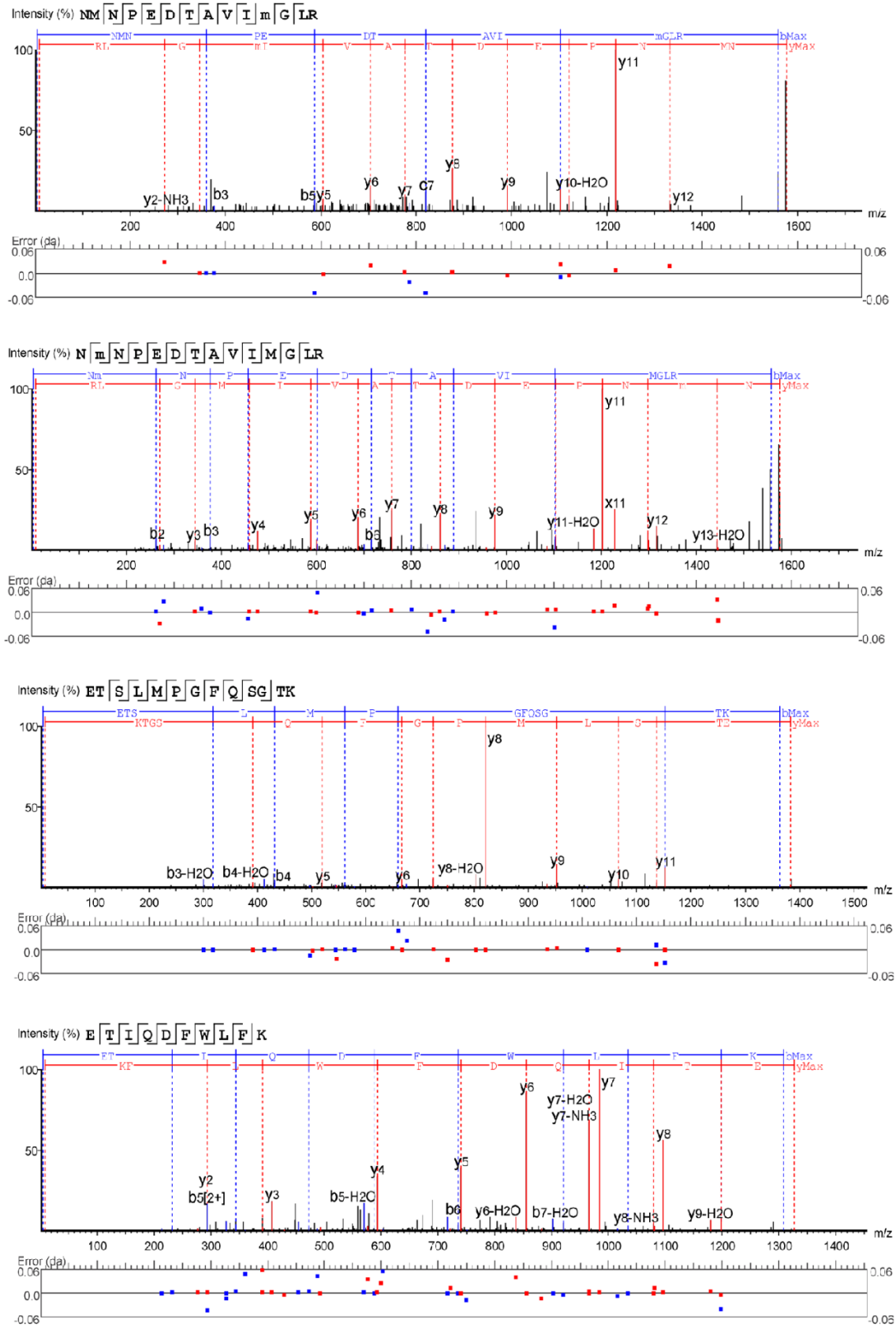
Supplementary Table 6.2 Mice testis RNA-Seq data at different developmental stages.

Project	Accession	Stage	Sex	Tissue
PRJEB26869	ERR2588380	E10	Male	Testis
PRJEB26869	ERR2588381	E10	Male	Testis
PRJEB26869	ERR2588400	E11	Male	Testis
PRJEB26869	ERR2588401	E11	Male	Testis
PRJEB26869	ERR2588420	E12	Male	Testis
PRJEB26869	ERR2588421	E12	Male	Testis
PRJEB26869	ERR2588449	E13	Male	Testis
PRJEB26869	ERR2588450	E13	Male	Testis
PRJEB26869	ERR2588471	E14	Male	Testis
PRJEB26869	ERR2588472	E14	Male	Testis
PRJEB26869	ERR2588495	E15	Male	Testis
PRJEB26869	ERR2588496	E15	Male	Testis
PRJEB26869	ERR2588519	E16	Male	Testis
PRJEB26869	ERR2588520	E16	Male	Testis
PRJEB26869	ERR2588543	E17	Male	Testis
PRJEB26869	ERR2588544	E17	Male	Testis
PRJEB26869	ERR2588567	E18	Male	Testis
PRJEB26869	ERR2588590	0dpb	Male	Testis
PRJEB26869	ERR2588591	0dpb	Male	Testis
PRJEB26869	ERR2588613	2wpb	Male	Testis
PRJEB26869	ERR2588614	2wpb	Male	Testis
PRJEB26869	ERR2588637	4wpb	Male	Testis
PRJEB26869	ERR2588638	4wpb	Male	Testis
PRJEB26869	ERR2588661	3dpb	Male	Testis
PRJEB26869	ERR2588662	3dpb	Male	Testis
PRJEB26869	ERR2588685	9wpb	Male	Testis
PRJEB26869	ERR2588686	9wpb	Male	Testis

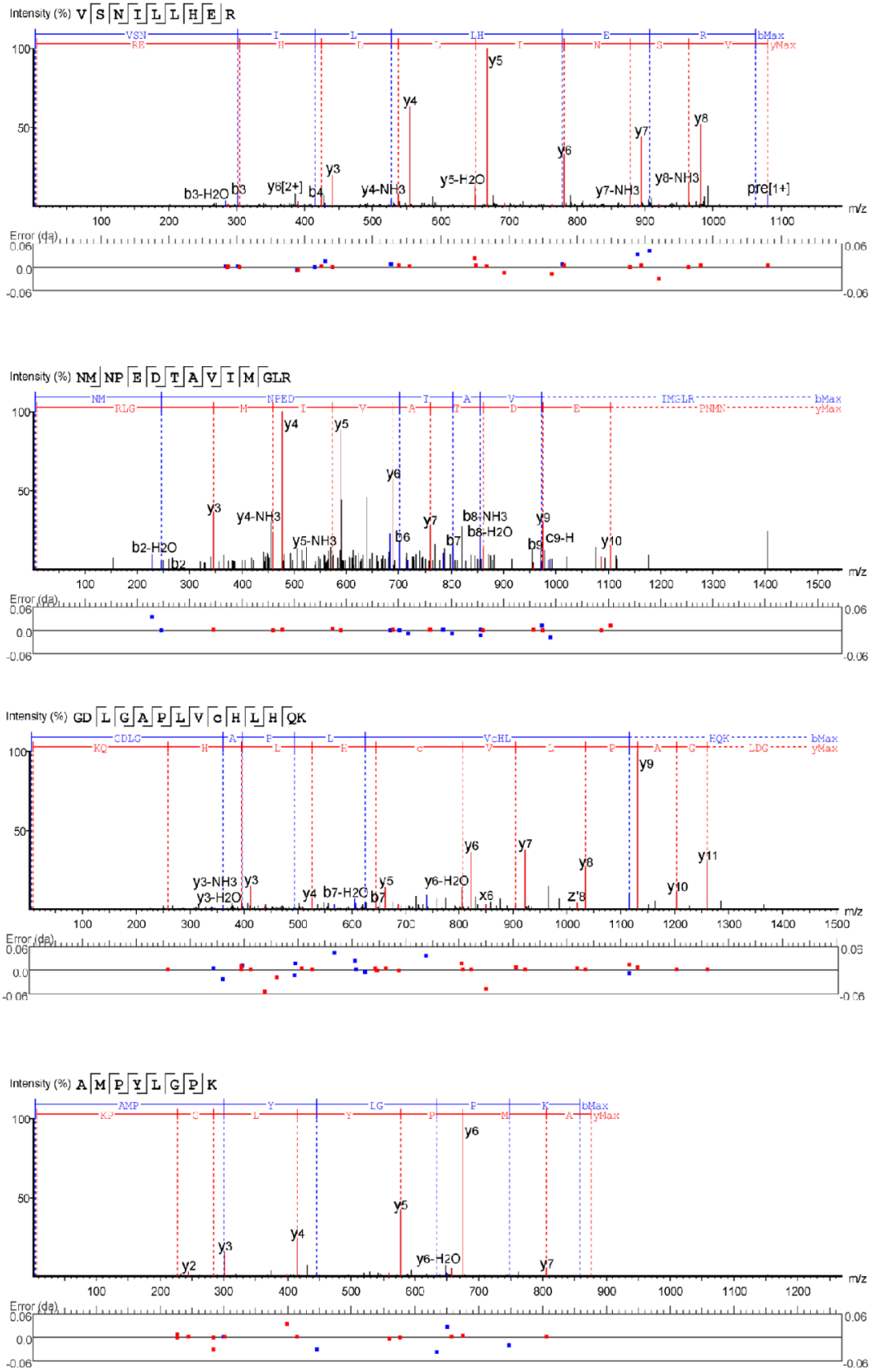
Supplementary Table 6.3 Primer sequences for cDNA of *PRSSLY* in dogs.

Name	Forward Primer	Products length	Annealing temperature
Pair 1-F	ACTGCTGTGGTGCCACATTT	1179	60
Pair 1-R	CCCAAGAGCTGCTAACAATGG		
Pair 2-F	GCTTCTATTCGGCCACAGTTC	505	60
Pair 2-R	TGTCAGGATAGAGGTGGGCA		

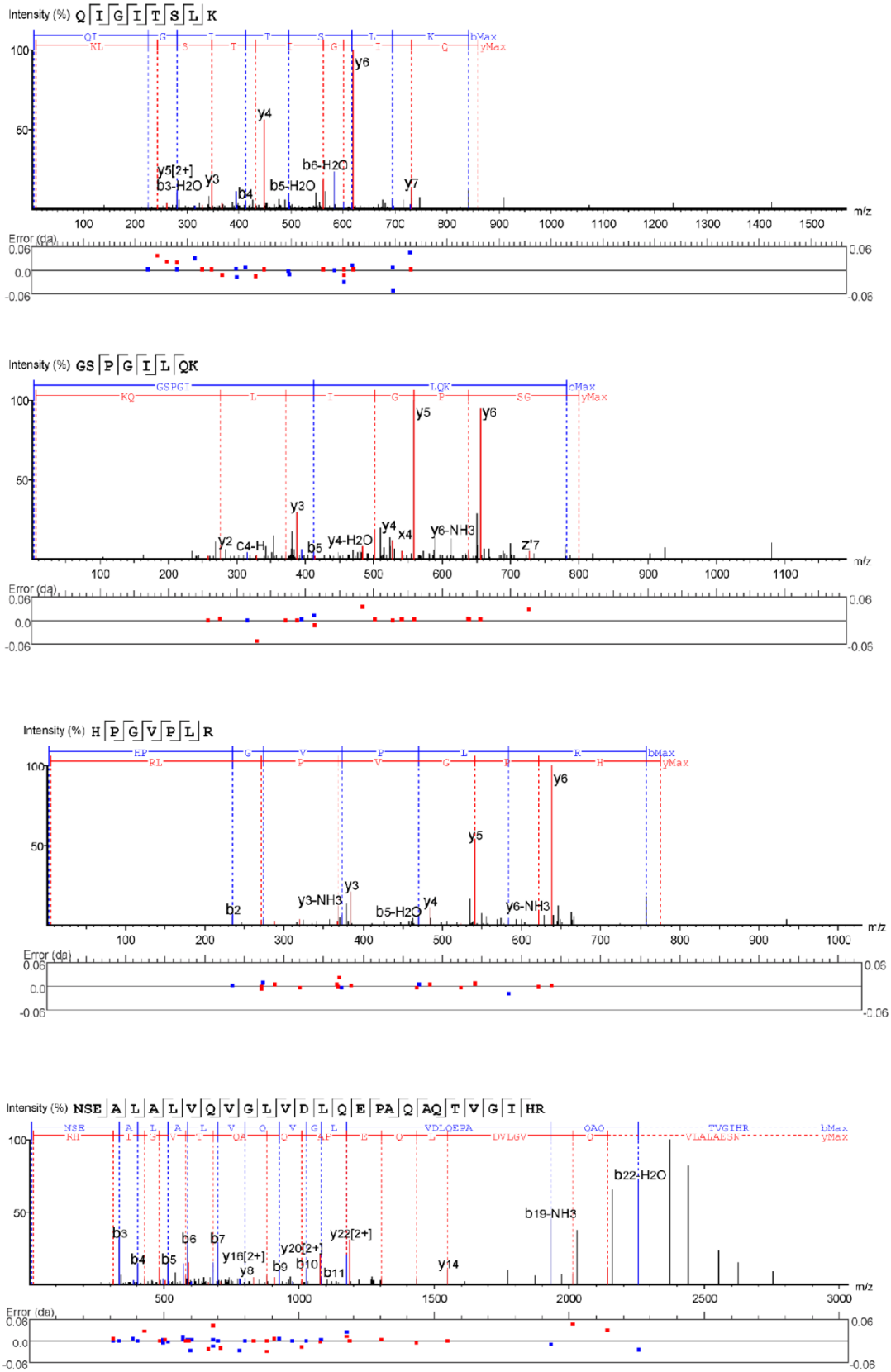
Supplementary Figure 6.1



Supplementary Figure 6.1 Continued

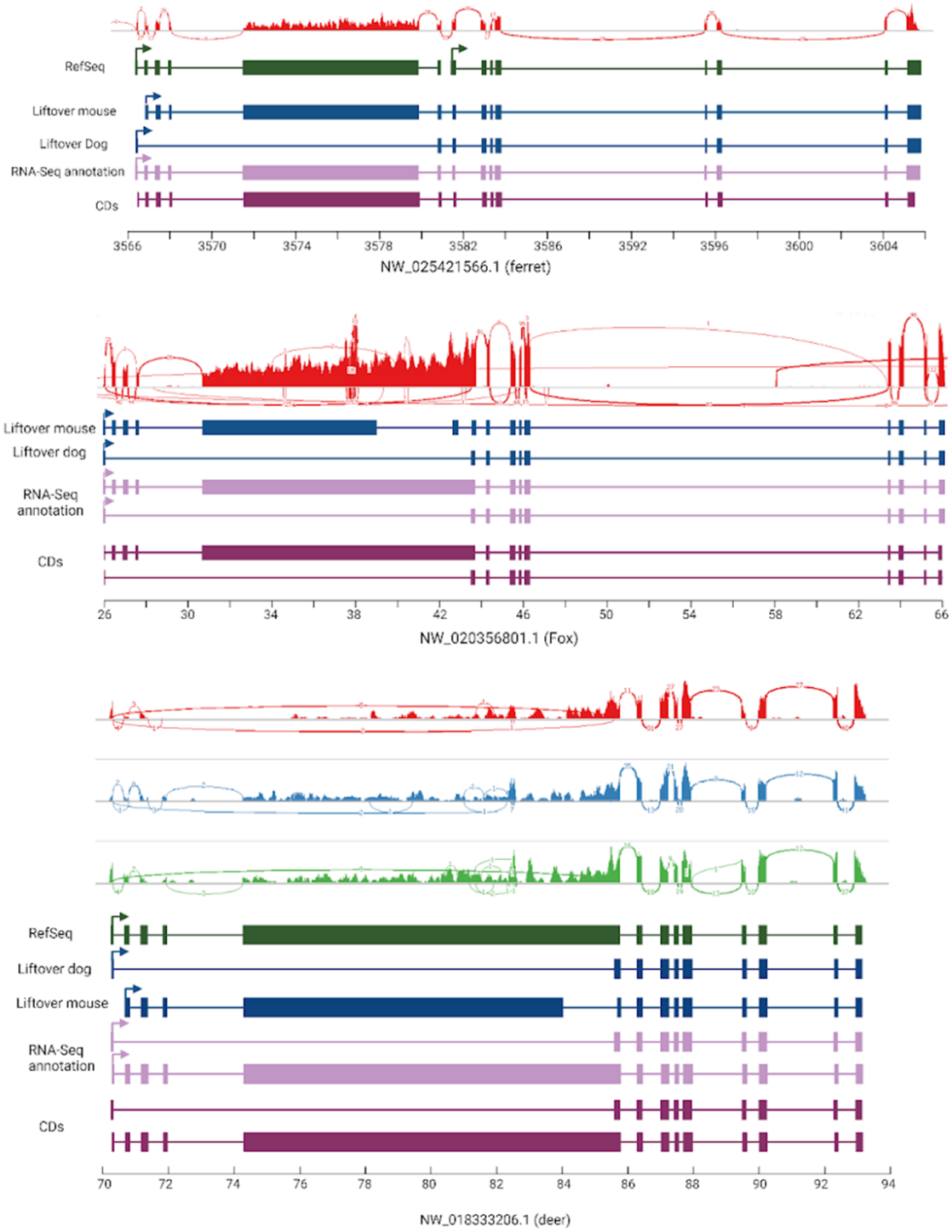


Supplementary Figure 6.1 Continued

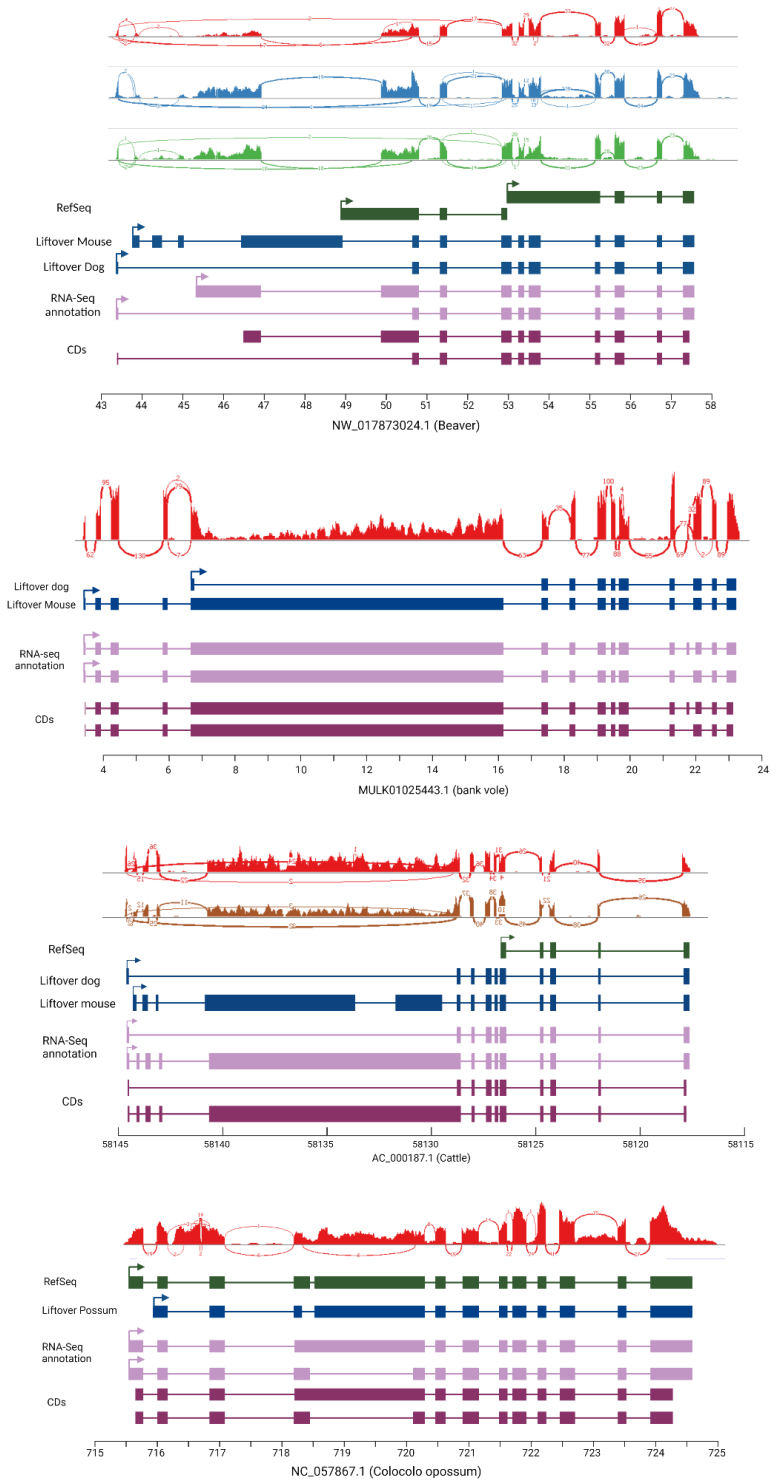


Supplementary Figure 6.1. MS spectra of detected peptides for PRSSLY.

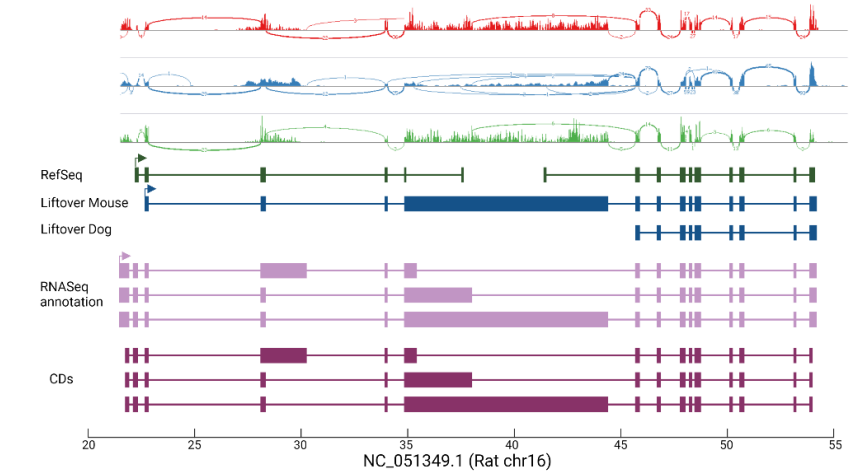
Supplementary Figure 6.2

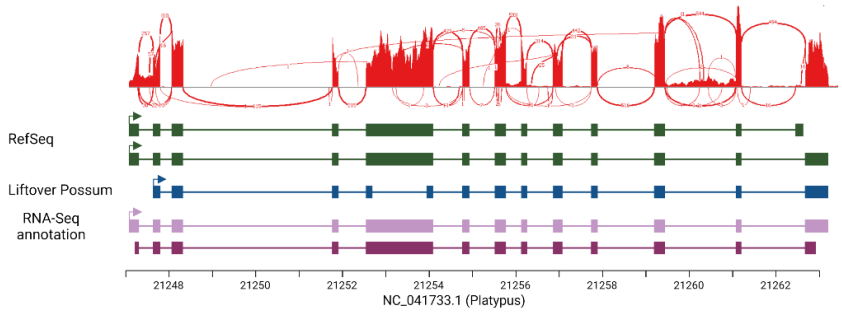


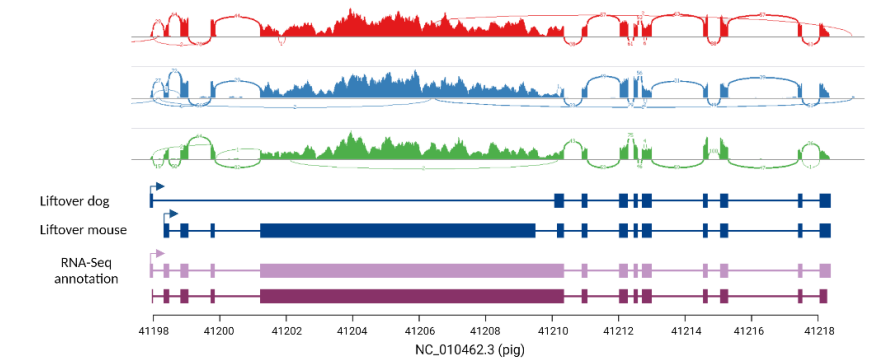
Supplementary Figure 6.2 continued

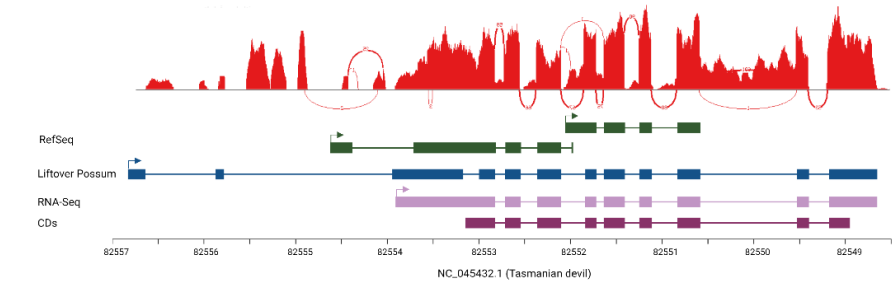


Supplementary Figure 6.2 continued

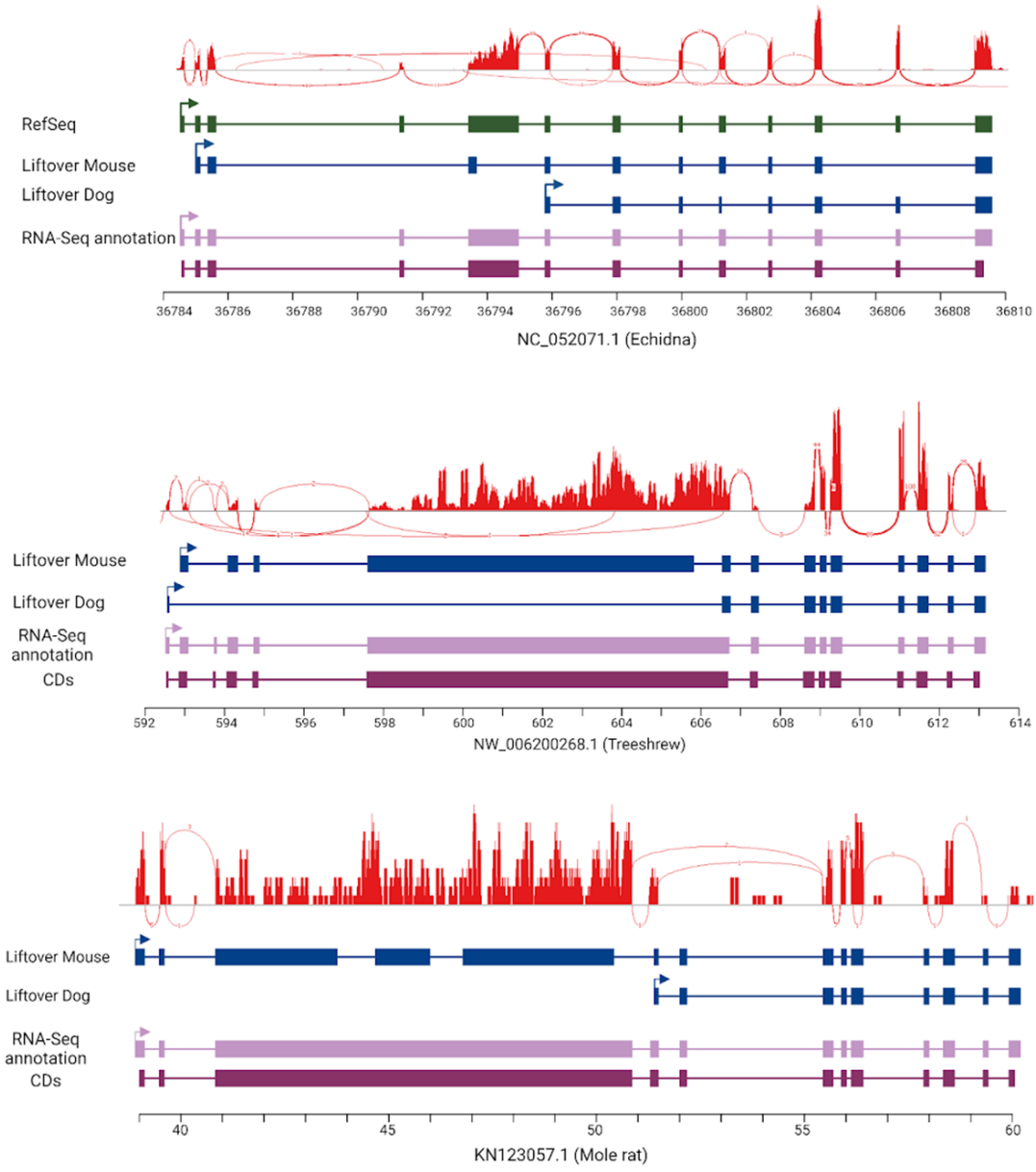




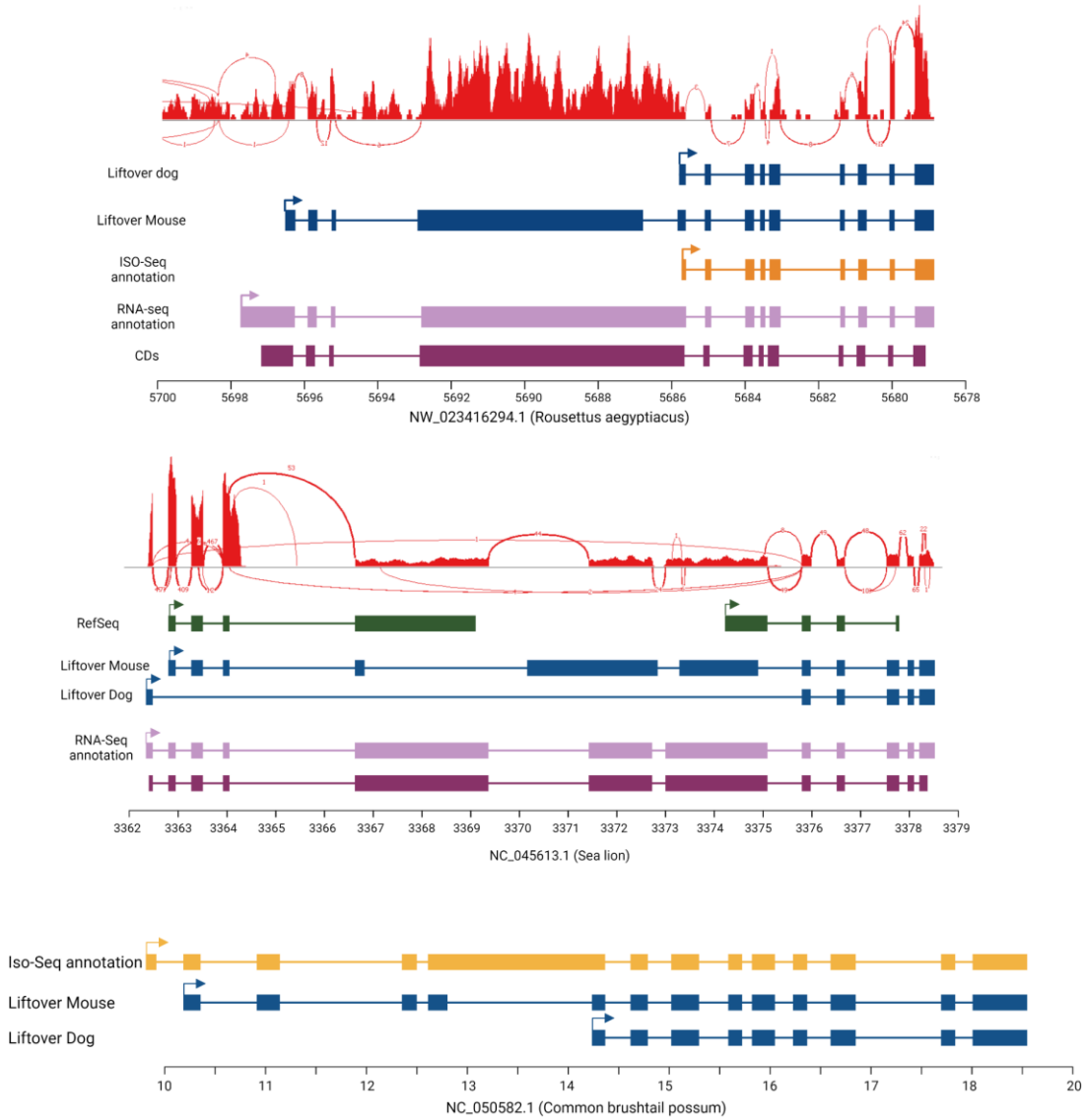




Supplementary Figure 6.2 continued



Supplementary Figure 6.2 continued



Supplementary Figure 6.2 Annotation of *PRSSLY* in mammals. The alignment of RNA-Seq of testis and liftover annotation based on the PRSSLY protein of dogs and mice are shown.