

Tracing genome evolution: from ancestral karyotypes to current genomes in three mammalian clades

Cristina Arias Sardá

School of Biosciences
University of Kent
2023



A thesis submitted to the University of Kent for the Degree of Doctor
of Philosophy in Genetics

Declaration

No part of this thesis has been submitted in support of an application for any degree or other qualification of the University of Kent, or any other University or Institution of learning.

Cristina Arias Sardá

March, 2023

To my grandparents, who will always be with me.
To my dad and my mum.

A mis abuelos, que estarán siempre conmigo.
A mi padre y a mi madre.

"I was taught that the way of progress was neither swift nor easy."

Marie Curie (1867-1934)

Acknowledgements/Agradecimientos

About three years and six months I moved to Canterbury to live one of the most incredible experiences of my life. It has been a time with many changes, both personally and professionally, and without forgetting a global pandemic! However, I've always been accompanied by wonderful people who have been crucial to finish this thesis.

First, I want to thank the person who is responsible of all of this. Thank you, Marta, for choosing me as your first PhD student, even though I didn't know how to open the computer's terminal on my first day. Thanks for guiding me during this process, for all your advice, and for teaching me how to be a better scientist. It's been a pleasure working with you!!

Thanks to all members of Farré, Ellis and Griffin labs for their help and friendship, specially, thanks to Corey for all his support, and you're welcome to everyone for the Spanish lessons.

Thanks to the most similar and different person, both at the same time, that I've ever met. Thank you Lucía for your advice, for helping me every time I needed it, and for sharing my pessimism in any circumstance. PS. Good luck finishing your thesis!!

Obviously, my best friends have been my best supporters. Thanks to Bea, Inez and María for being part of all chapters of my life, even by videocall. Thanks to Alba and Carla for sending me extremely long audios that make me feel that you are here with me. Thank you, Cris, for not killing me every single time I have forgotten to answer one of your texts. Love you lots!!

Thanks to my Spanish/Portuguese family in Canterbury, for being always by my side. You are the truly responsible people for both the thesis and me having succeeded. You are my best achievement without any doubt!! Ángela, Carla, Paula and Sam, love you, my dears!

Finally, thanks to my family, for loving me and supporting me so much, and for looking after me from the distance. Specially, thanks to my parents for always believe in me and respect all my decisions. You are and will always be my best example, my role models, my best supporters and the people I love the most.

Agradecimientos/Acknowledgements

Hace tres años y seis meses que me mudé a Canterbury para vivir una de las experiencias más increíbles de mi vida. Ha sido un tiempo de muchos cambios, tanto personalmente como profesionalmente, y ¡no olvidemos una pandemia a nivel global! Aun así, siempre he estado acompañada de personas maravillosas sin las que esta tesis no hubiera llegado a su fin.

En primer lugar, quiero darle las gracias a la persona responsable de todo esto. Gracias Marta por escogerme como tu primera estudiante de doctorado, a pesar de que el primer día no supiera ni cómo abrir la terminal del ordenador. Gracias por guiarme durante el proceso, por tus consejos y por enseñarme a ser mejor científica. Ha sido un placer trabajar contigo.

Gracias a todos los miembros de Farré, Ellis y Griffin labs por su ayuda y su amistad, en especial, gracias a Corey por todo su apoyo, y de nada a todos por las clases de español.

Gracias a la persona más parecida y, al mismo tiempo, más opuesta a mí que he conocido nunca. Gracias Lucía por tus consejos, por echarme una mano siempre que lo he necesitado y por compartir mi pesimismo ante cualquier circunstancia. PD. ¡Mucho ánimo con tu tesis!

Evidentemente, mis mejores amigas han sido mi mejor apoyo. Gracias Bea, Inez y María por estar en todos los capítulos de mi vida, aunque sea vía videollamada. Gracias Alba y Carla por enviarme esos audios larguísimos que hace que me parezca que estáis aquí conmigo. Gracias Cris por no matarme cuando se me olvida contestarte a un mensaje. ¡Os quiero mucho!

Gracias a mi familia española/portuguesa en Canterbury, por estar a mi lado siempre. Sois los verdaderos responsables de que tanto la tesis como yo misma hayamos salido adelante. Sin duda sois mi mejor descubrimiento. Ángela, Carla, Paula y Sam, ¡OKM, queridas!

Finalmente, quiero dar las gracias a mi familia, por quererme y apoyarme tanto, y por seguir cuidándome desde la distancia. Especialmente, gracias a mis padres por creer siempre en mí y respetar todas mis decisiones. Sois y seréis siempre mi mejor ejemplo, mi modelo a seguir, mi mayor apoyo y las personas a las que más quiero.

Table of contents

Declaration.....	- 3 -
Acknowledgements/Agradecimientos.....	- 9 -
Agradecimientos/Acknowledgements.....	- 10 -
List of Figures	- 15 -
List of tables	- 17 -
List of supplementary material	- 19 -
Abstract	- 21 -
Abbreviations	- 23 -
1. General introduction	- 25 -
1.1 Genetics: The origin	- 27 -
1.2 Genomics, a new step to study speciation	- 28 -
1.3 Chromosomics, the dynamics of chromosome evolution	- 29 -
1.3.1 Genome organization	- 29 -
1.3.2 Chromosome's three-dimensional (3D) structure	- 31 -
1.3.3 Chromosome mutations.....	- 32 -
1.3.3.1 Change in chromosome number: aneuploidies and polyploidies	- 33 -
1.3.3.2 Change in chromosome structure: Chromosome rearrangements (CRs)	- 34 -
1.3.3.2.1 Origin of CRs.....	- 34 -
1.3.3.2.2 Types of chromosome rearrangements (CRs)	- 35 -
1.3.3.2.3 Comparing genomes between species: distribution of CRs among different species ...	- 37 -
1.3.3.2.4 Consequences of CRs	- 40 -
1.4 Studying chromosome evolution: From cytogenetics to genomics.....	- 44 -
1.4.1 The origin of the first cytogenetic techniques	- 44 -
1.4.2 Comparative genomics: from past to current techniques	- 46 -
1.4.2.1 Genetic maps	- 46 -
1.4.2.2 Genome sequencing	- 47 -
1.4.2.3 Genome assembly and annotation.....	- 49 -
1.4.2.4 Computational approaches: The origin of bioinformatics	- 49 -
1.4.3 Ancestral karyotype reconstructions	- 50 -
1.5 Mammals: our target species	- 54 -
1.5.1 Phylogeny	- 55 -
1.5.2 Genome size	- 57 -
1.5.3 Chromosome number	- 58 -
1.5.4 Chromosome structure and evolution	- 59 -
1.5.5 Looking at the extremes: ruminants, marsupials and afrotherians	- 61 -
1.5.5.1 Methods to detect convergent evolution	- 63 -
1.6 Project aims	- 65 -
2. Materials and methods	- 67 -

2.1 Genomes used for the ancestral genome reconstructions	69 -
2.2 Filtering and masking genomes	72 -
2.3 Whole-genome pairwise alignments	72 -
2.4 Ancestral karyotypes reconstruction	74 -
2.5 Detection of Evolutionary Breakpoint Regions (EBRs) and Homologous Syntenic Blocks (HSBs).....	76 -
2.6 African elephant genome annotation.....	79 -
2.7 Multi-associations, statistical analysis and gene ontology enrichment analysis (GOEA) ..	81 -
2.8 Detection of orthologs.....	82 -
2.9 Detection of convergence rates	83 -
2.10 Detection of positive selection	84 -
3. The Ruminantia lineage	87 -
3.1 Introduction	89 -
3.2 Results	94 -
3.2.1 Pairwise alignments.....	94 -
3.2.2 Ancestral karyotype reconstructions	97 -
3.2.3 Assessing the reconstructions	102 -
3.2.4 Detection and study of evolutionary breakpoint regions (EBRs).....	105 -
3.2.5 Detection and study of homologous syntenic blocks (HSBs).....	111 -
3.3 Discussion	117 -
4. The Marsupialia lineage.....	123 -
4.1 Introduction	125 -
4.2 Results	128 -
4.2.1 Pairwise alignments.....	128 -
4.2.2 Ancestral karyotype reconstructions	130 -
4.2.3 Detection and study of evolutionary breakpoint regions (EBRs).....	134 -
4.2.4 Detection and study of homologous syntenic blocks (HSBs).....	138 -
4.3 Discussion	140 -
5. The Afrotheria lineage	145 -
5.1 Introduction	147 -
5.2 Results	150 -
5.2.1 Pairwise alignments.....	150 -
5.2.2 Ancestral karyotype reconstructions	152 -
5.2.4 Detection and study of evolutionary breakpoint regions (EBRs).....	156 -
5.2.5 Detection and study of homologous syntenic blocks (HSBs).....	162 -
5.3 Discussion	165 -
6. Evolution in mammals: Looking for convergence traits in extreme chromosomes ...	169 -
6.1 Introduction	171 -
6.2 Results	174 -
6.2.1 Identification of orthologous genes in mammals	174 -
6.2.2 Estimation of convergence rates	175 -
6.2.3 Detection of positive selection.....	178 -

6.3 Discussion	- 181 -
7. General discussion	- 185 -
7.1. The importance of a good genome assembly.....	- 187 -
7.2 Reconstructing ancestral karyotypes and CRs	- 188 -
7.3 Contrasting functional roles in EBRs and HSBs	- 191 -
7.4 The role of DSBs repair mechanisms in the convergent extreme reduction of diploid number in mammals.....	- 195 -
7.5 Future directions	- 197 -
8. Bibliography	- 199 -
9. Supplementary data	- 225 -

List of Figures

Figure 1.1. Chromatin condensation process.	- 30 -
Figure 1.2. Compartmentalization of genomes into different levels of organization.	- 32 -
Figure 1.3. A human karyotype which consists of 46 chromosomes.	- 33 -
Figure 1.4. Sequence of balanced CRs.	- 36 -
Figure 1.5. A. Schematic representation of EBRs and HSBs when dog and human genomes are compared, and multispecies alignment to human chromosome 7.	- 38 -
Figure 1.6. FISH performed in human chromosomes to detect complex CRs related to disease.	- 45 -
Figure 1.7. Evolutionary relationships of eutherian mammals.	- 56 -
Figure 1.8. Karyotypes corresponding to a male individual of Indian muntjac and a male individual of Red viscacha rat.	- 58 -
Figure 1.9. Chromosomal number variation among mammals.	- 59 -
Figure 2.1. EH image of cattle chromosome 13 aligned with the five different ruminant ancestors: RAK, PAK, BOCE, BOMO, and BOVIN.	- 77 -
Figure 2.2. African elephant HSBs classification comparing to the afrotherian ancestor.	- 78 -
Figure 2.3. The three selected networks from GeneMANIA analysis.	- 85 -
Figure 2.4. Phylogenetic tree used as input tree in EasyCodeML.	- 86 -
Figure 3.1. Phylogeny of ruminants among the Geological Time Scale (GTM) of Earth.	- 90 -
Figure 3.2. Diploid chromosome number in ruminants.	- 92 -
Figure 3.3. EH image of chromosomes 20 and 13 of cattle aligned with 11 of the ruminant species.	- 96 -
Figure 3.4. Phylogenetic tree of all species included in the ruminant ancestral reconstruction.	- 97 -
Figure 3.5. Pairwise comparison between RAK and PAK.	- 100 -
Figure 3.6. Pairwise comparison between PAK and BOVIN.	- 100 -
Figure 3.7. Ideograms comparing RAK, PAK, and BOVIN with cattle chromosomes.	- 101 -
Figure 3.8. Ideograms comparing RAK2, PAK2, and BOVIN2 with cattle chromosomes.	- 104 -
Figure 3.9. Correlation between the number of well-defined (A) and well-defined and not well-defined (B) EBRs and cattle chromosome sizes in Kb.	- 107 -
Figure 3.10. Permutation tests performed to establish the association between gene content and cattle EBRs.	- 108 -
Figure 3.11. GOEA of cattle EBRs.	- 109 -
Figure 3.12. TEs associated with cattle EBRs.	- 111 -
Figure 3.13. GOEA of cattle HSBs according to their classification.	- 116 -
Figure 4.1. Phylogeny of marsupials among the Geological Time Scale (GTM) of Earth.	- 126 -
Figure 4.2. EH image of chromosome 4 of Tasmanian devil aligned to four marsupial species.	- 129 -
Figure 4.3. Phylogenetic tree of all species used for the marsupial ancestral reconstruction.	- 130 -
Figure 4.4. Marsupial ancestor comparing to human chromosomes.	- 132 -
Figure 4.5. Pairwise comparison between MAK, MAUK, and Tasmanian devil.	- 133 -

Figure 4.6. Correlation between the number of well-defined (A) and well-defined and not well-defined (B) EBRs and Tasmanian devil chromosome sizes in Kb.	- 136 -
Figure 4.7. Permutation tests performed to establish the correlation between gene content and Tasdmanian devil EBRs.....	- 137 -
Figure 4.8. GOEA of Tasmanian devil HSBs according to their classification.....	- 140 -
Figure 5.1. Phylogeny of afrotherians among the Geological Time Scale (GTM) of Earth.	- 148 -
Figure 5.2. EH image of chromosome 24 of African elephant aligned to six afrotherian species..	- 152 -
Figure 5.3. Phylogenetic tree of all species included in the afrotherian ancestral reconstruction	- 153 -
Figure 5.4. Ideogram comparing AFAK and African elephant chromosomes.	- 155 -
Figure 5.5. Correlation between the number of well-defined (A) and will-defined and not well-defined (B) EBRs and African elephant chromosomes.	- 158 -
Figure 5.6. Permutation tests performed to establish the association between gene content and African elephant EBRs.	- 159 -
Figure 5.7. Repetitive elements associated with African elephant EBRs	- 162 -
Figure 5.8. GOEA of African elephant HSBs according to their classification.	- 165 -
Figure 6.1. Phylogenetic tree inferred for OrthoFinder.	- 175 -
Figure 6.2. RERs of species using Group 1 as ingroup (or foreground) species: JMJD1C and MROH9.....	- 177 -
Figure 6.3. RERs of species using Group 1 as ingroup (or foreground) species: IQSEQ2.....	- 180 -
Figure 7.1. Representation of genomic architectural features detected in lineage-specific chromosomal reorganizations in the Tasmanian devil and African elephant.	- 191 -

List of tables

Table 2.1. Genomes used for the ancestral genome reconstructions.	- 70 -
Table 2.2. Genomes used for the detection of orthologs.....	- 82 -
Table 3.1. Coverage of net files and HSBs in the alignments of ruminant species against cattle.....	- 94 -
Table 3.2. DESCHRAMBLER results for ruminant ancestors.....	- 99 -
Table 3.3. Summary table of type of CR differentiating every ancestor from cattle.....	-102-
Table 3.4. DESCHRAMBLER results with second set of species for ruminant ancestors.	- 103 -
Table 3.5. Classification of cattle EBRs.	- 105 -
Table 3.6. Classification of HSBs in cattle.	- 112 -
Table 4.1. Coverage of net files and HSBs in the alignment of all marsupial species against Tasmanian devil.	- 128 -
Table 4.2. DESCHRAMBER results for marsupial ancestors.....	- 131 -
Table 4.3. Summary table of type of CR differentiating every ancestor from Tasmanian devil (TD).....	- 134 -
Table 4.4. Classification of Tasmanian devil EBRs.....	- 134 -
Table 4.5. Classification of HSBs in Tasmanian devil genome.....	- 138 -
Table 5.1. Coverage of net files and HSBs of the alignments of afrotherian species against African elephant.....	- 151 -
Table 5.2. DESCHRAMBER results for afrotherian ancestors.....	- 154 -
Table 5.3. Summary table of type of CR differentiating every ancestor from African elephant.....	- 156 -
Table 5.4. Classification of African elephant EBRs.	- 156 -
Table 5.5. Summary of African elephant HSBs.....	- 163 -
Table 6.1. EasyCodeML results, showing P-values for every gene in the specific branches of interest.....	- 179 -

List of supplementary material

Table S1. Information of all ruminant species.	- 227 -
Figure S1. Ideogram comparing BOCE with cattle chromosomes.	- 233 -
Figure S2. Ideogram comparing BOMO with cattle chromosomes.	- 234 -
Table S2. Genes found in Tasmanian devil well-defined EBRs expanded up- and downstream until reach 50 Kb.	- 235 -
Table S3. Genes found in Tasmanian devil no well-defined EBRs.	- 235 -
Table S4. Classification of HSBs in African elephant.	- 236 -
Table S5. Genes showing significant P-values in Group 1 after running RERConverge.	- 238 -

Abstract

Mammals show a wide range of chromosomal diversity. Their karyotypes range from $2n=6/7$ in the Indian muntjac (Ruminantia) to $2n=102$ in the Viscacha rat (Rodentia). Although the mammalian genome size is round 3.5 Gb, the mammalian average chromosome size is highly variable, including species with a high number of small chromosomes such as cattle (a ruminant with $2n=60$ and an average chromosome size of 87 Mb), and species with a low number of big chromosomes such as Tasmanian devil (a marsupial with $2n=14$ and an average chromosome size of 440 Mb). Identifying and timing when and where gross genomic rearrangements occurred during evolution will help to explain changes in genome structure with functional consequences that might eventually lead to speciation. Here we used DESCHRAMBLER to reconstruct eight ancestral genomes from three different lineages: Ruminantia, Marsupialia and Afrotheria. We classified the rearrangement events occurring in each lineage and identified the Evolutionary Breakpoints Regions (EBRs) and Homologous Synteny Blocks (HSBs). Cattle and African elephant showed the same number of well-defined EBRs, 32, while Tasmanian devil only showed 19. Marsupial and ruminant genomes are characterised by inversions, while interchromosomal rearrangements are also important in the oldest ancestor of ruminants and are the main rearrangement force in the afrotherians. These EBRs are located in gene-rich regions of the genome, and in ruminants and afrotherians, they are also located in regions with a high density of transposable elements (TEs). Moreover, we identified signatures of convergent evolution in mammals with an extreme range of both, diploid numbers and chromosome sizes, using representative species of the three lineages and searching for positive selection in their orthologous genes. We found that some genes under convergent evolution and positive selection were related to processes such as chromatin structure or DNA repair, such as *MSH6* or *RNF123*, that, at the same time, are related to causes and consequences of chromosome rearrangements (CRs). Overall, our results significantly expand knowledge of genome evolution and will facilitate greater understanding of the role of CRs in mammalian evolution.

Abbreviations

3D - Three-Dimensional	Mb - Mega Base Pairs
3D-FISH – 3D-Fluorescence In Situ Hybridization	MDD - Mammal Diversity Database
BACs - Bacterial Artificial Chromosomes	MER - Medium Reiterated Repeats
Bp - Base Pairs	MHC - Major Histocompatibility Complex
Cm - Centimetre	ML - Maximum Likelihood
CNV - Copy number variants	My - Million Years
CRs - Chromosomal Rearrangements	Mya - Million Years Ago
CTCF - CCCTC-Binding Factor	NAHR - Non-Allelic Homologous Recombination
DFTD - Devil Facial Tumour Disease	NCBI - National Center for Biotechnology Information
DFT1 - Devil Facial Tumour 1	NGS - Next-Generation Sequencing,
DFT2 - Devil Facial Tumour 2	ONT - Oxford Nanopore Technology
DNA - Desoxyribonucleic Acid	PacBio - Pacific Biosciences
DSBs - Double-Strand Breaks	RACFs - Reconstructed Ancestral Chromosome Fragments
DToL - Darwin Tree of Life Project	RERs - Relative Evolutionary Rates
EBP - Earth BioGenome Project	RGD - Ruminant Genome Database
EBRs - Evolutionary Breakpoint Regions	RNA - Ribonucleic Acid
EH - Evolution Highway	SDs - Segmental Duplications
ERVs - Endogenous Retroviruses	SFs - Synteny Fragments
FDR – False Discovery Rate	SINEs - Short Interspersed Nuclear Elements
Gb - Giga Base Pairs	SNVs - Single-Nucleotide Variants
GO - Gene Ontology	SSBs - Single-Strand Breaks
GOEA - Gene Ontology Enrichment Analysis	SSBR - Single-Strand Break Repair
GTM - Geological Time Scale	STAG - Species Tree inference from All Genes
HSBs - Homologous Synteny Blocks	TADs - Topologically Associated Domains
IGV - Integrative Genomics Viewer	TEs - Transposable Elements
IUCN - International Union for Conservation of Nature	TRs - Tandem Repeats
Kb - Kilo Base Pairs	UCSC - University of California, Santa Cruz
Kg - Kilogram	VGP - Vertebrates Genome Project
LCRs - Low Copy Repeats	
LINEs - Long Interspersed Nuclear Elements	
LTRs - Long Terminal Repeats	



1. General introduction

1.1 Genetics: The origin

Genetics, the study of inheritance and how organisms can show characteristics that have been inherited from their progenitors, arose in 1865 when Gregor Mendel published his memoir on plant hybridization. Using *Pisum sativum* (common pea) as a study model, he hypothesized that each parent contributes some particulate matter to the offspring and he called this heritable substance “elementen” (German word for “element”). This was the first description of genes.

However, as reviewed in (Gayon, 2016), it was not until 1905 when the English biologist William Bateson used the term “genetics” to describe “the science of heredity and variation”, and it comes from the Greek word *genetikos*, which at the same time comes from the word *genesis*, meaning “origin”.

After 1906, when it is considered the official creation of Genetics as a discipline, many other terms that are very common nowadays were created, such as “gene”, “genotype” or “phenotype”. The origin of the word “gene” is attributed to the Danish botanist and geneticist, Wilhelm Ludvig Johannsen, who used it in 1909 to describe “the units of heredity”. Although it is thought that Johannsen shortened the German word “Pangenes”, which was used previously by the Dutch botanist Hugo de Vries (Gayon, 2016).

At the end of the 19th century, while the Mendelian theory was beginning to be relatively well known, the description of the chromosomal morphology (see **Introduction section 1.3.1**) and the study of processes such as mitosis and meiosis began to be more popular and originated a new way of understanding evolution with the “chromosomal theory of inheritance”. These two theories were fused by Thomas Hunt Morgan, an American evolutionary biologist who in 1908 began to work with *Drosophila melanogaster* (fruit fly) and in 1915 published, together with other colleagues, one of the most important book on the history of evolutionary biology and genetics: *The mechanism of Mendelian heredity* (Morgan et al., 1915).

Although it was Mendel who first set the foundations of modern genetics, the fusion of his theory together with the chromosomal theory, was the essential step to set up the basis for the study of heredity and the role that genes play in inheritance and in evolution.

1.2 Genomics, a new step to study speciation

The definition of gene has been changing during the last decades. In the early 1960s, every gene was defined as a discrete sequence of a deoxyribonucleic acid (DNA) molecule that encoded a polypeptide chain. Nowadays, a gene is considered a DNA sequence that specifies for one or more sequence-related ribonucleic acids (RNAs) and/or proteins (Portin & Wilkins, 2017).

The term "genome" was used to describe a complete set of genes and chromosomes in an organism. The genome has been very often described as the "information repository" of an organism, containing millions of letters of DNA and RNA (adenine (A), thymine (T), cytosine (C), guanine (G) and uracil (U)), whose transmission across generations allows the inheritance of different traits. The definition of what a genome is has been controversial, due to the continuous development of new information; however, a most recent genome definition could be: "an informational entity, often but not always manifest as DNA, encoding a broad set of functional possibilities that, together with other sources of information, produce and maintain the organism" (Goldman & Landweber, 2016).

In the same way that the study of all aspects of genes is encompassed by the term "genetics", the study of whole genomes of organisms is referred as "genomics". This term, genomics, was coined in 1986, by the American geneticist Thomas Roderick, to describe the scientific discipline of mapping, sequencing, and analysing genomes (Hieter & Boguski, 1997). Moreover, genome analysis can be divided into two categories, "structural genomics" and "functional genomics"; where the first one aims to reconstruct a high-resolution map containing the complete DNA sequence of an organism, while the second one aims to study single genes or proteins of an organism to explore their specific function.

Studies describing genome characteristics such as genome size and function or chromosome number and shape (karyotype), allowed us to compare differences between species and hence, to trace back genome histories and their evolution. The process of speciation was first described by the English biologist Charles Darwin in his book "On the origin of species" (Darwin, 1859) while studying Galapagos finches. Darwin viewed evolution by natural selection as a very gradual mechanism of change within populations and postulated that speciation occurs when a

group within a species separates from other members of its species and develops its own unique characteristics differentiating them from their ancestors.

Only within the past few decades, genomic studies have progressed from the analysis of a single or a small number of genes to the investigation of thousands of genes at once; and this has promptly expanded from the determination of DNA sequences to the study of the expression profiles and the roles of both genes and proteins. As we know that even individuals belonging to same species could possess not identical genomes, the completeness of a large amount of genome sequences of several organisms would allow the study of species evolution with a high level of accuracy.

1.3 Chromosomics, the dynamics of chromosome evolution

1.3.1 Genome organization

In order to understand how genomes change, first we have to understand how genomes are organised. In most prokaryotic cells the genome is organised in one or more DNA molecules within the cell and usually in a circular shape. In contrast, in eukaryotic cells the genome is distributed within the nuclei among multiple linear structures, the chromosomes.

As the total length of cellular DNA is up to a hundred thousand times a cell's diameter, the packing of DNA is crucial to cell architecture, to prevent the long DNA molecules from getting knotted or tangled with each other during cell division – when they must be precisely segregated to daughter cells (Lodish, 2008). The packing of eukaryotic DNA (its tertiary chromosomal structure) is performed by abundant nuclear proteins called histones. The complex of histones, nonhistone proteins and DNA constitutes the chromatin.

Chromatin can be found on cells in two different states: (i) euchromatin, which undergoes the normal process of condensation and decondensation during the cell cycle; and (ii) heterochromatin, which remains in a highly condensed state throughout the cell cycle. The majority of the chromosomal material is constituted by euchromatin (where more of the transcription takes place), although all chromosomes have permanent heterochromatin, called constitutive heterochromatin, at the centromeres and telomeres, reviewed in (Pierce, 2020).

The basic repeating structural (and functional) unit of chromatin is the nucleosome, containing eight histone proteins and about 146 base pairs (bp) of DNA (A. Wolffe, 1998; A. P. Wolffe & Matzke, 1999). Today scientists describe the nucleosome as a structure in which four histones (H2A, H2B, H3, and H4) come together to form a histone octamer containing two of each. This octamer binds and wraps approximately 1.7 turns of DNA (or which is the same, about 146 bp). After the addition of one H1 protein wraps another 20 base pairs, resulting in two full turns around the octamer, and forming a structure called “chromatosome”. The resulting 166 bp structure is not very long, therefore, every chromosome contains hundreds of thousands of nucleosomes, and these nucleosomes are joined by the DNA that runs between them, and that is referred as linker DNA. Each chromosome is thus a long chain of nucleosomes, which gives the appearance of a string of beads when viewed using an electron microscope (Pierce, 2020). All this process of chromatin condensation until forming a chromosome is explained in detail in (**Figure 1.1**).

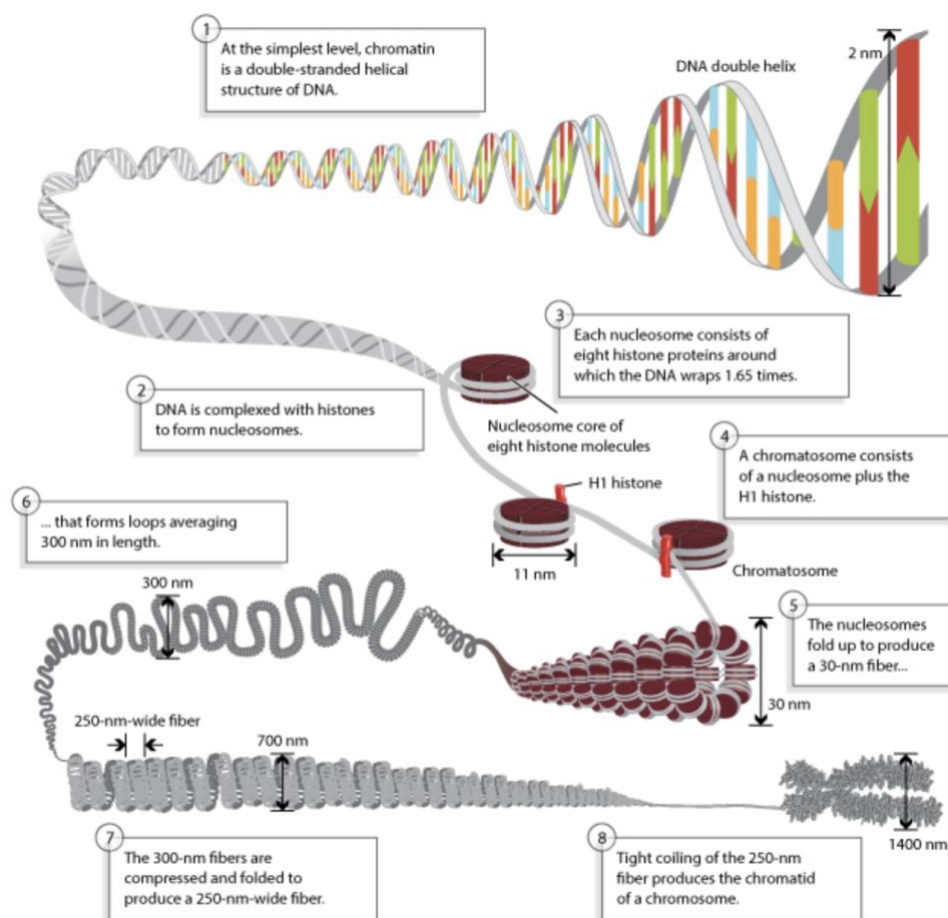


Figure 1.1. Chromatin condensation process. Description of the different levels of compaction of the chromatin, from the simplest level as a double-stranded helical structure of DNA to the totally compressed chromatin forming chromosomes. Figure from Nature Education, adapted from (Pierce, 2020).

DNA condensation has implications not only in cell structure but also in gene expression and ensuring the correct segregation during cell division. Depending on the state of chromatin, DNA inside the interphase nuclei is organised in heterochromatic regions, which are believed to contain transcriptionally inactive genes, and euchromatic regions, where most transcription process takes place (Lodish, 2008).

1.3.2 Chromosome's three-dimensional (3D) structure

Gene regulation is not only affected by chromatin condensation. The three-dimensional (3D) morphological changes in chromosomes are also essential elements in gene regulation. To link the plasticity of chromosomes to the 3D positions of genes, the term “chromosomics” was introduced.

Chromosomics includes research into chromatin-modification-mediated changes in the architecture of chromosomes, which may influence the functions and lifespans of cells, tissues, organs, and individuals. It also addresses the occurrence and prevalence of chromosomal gaps and breaks (Claussen, 2005).

It has been probed that chromosomes are not randomly positioned in the nucleus, but organised into specific areas called chromosomal territories (Cremer & Cremer, 2006) that change during the cell cycle and development. Maintenance of these territories is important for proper cell functioning, replication, and the accurate division and differentiation of cells. Chromatin remodelling and changes in chromatin conformation affect interactions between sequences in different genomic regions and can influence gene regulation (Deakin et al., 2019).

When looking in detail the structure of chromosomal territories, they possess specific partitions called chromosomal compartments, which at the same time are partitioned into topologically associated domains (TADs) and chromatin loops mediated by CCCTC-binding factor (CTCF) (**Figure 1.2**). The chromosomal compartments are classified in two groups: A and B. Sequences in the A compartment generally contain transcribed genes and active histone modifications, although some regions are not transcribed, while the B compartment generally contains inactive genes with histone modifications associated to a transcriptionally repressed state but also some transcribed genes. In the same way, TADs are separated by boundaries enriched in CTCF sites and highly transcribed genes (Rowley & Corces, 2018).

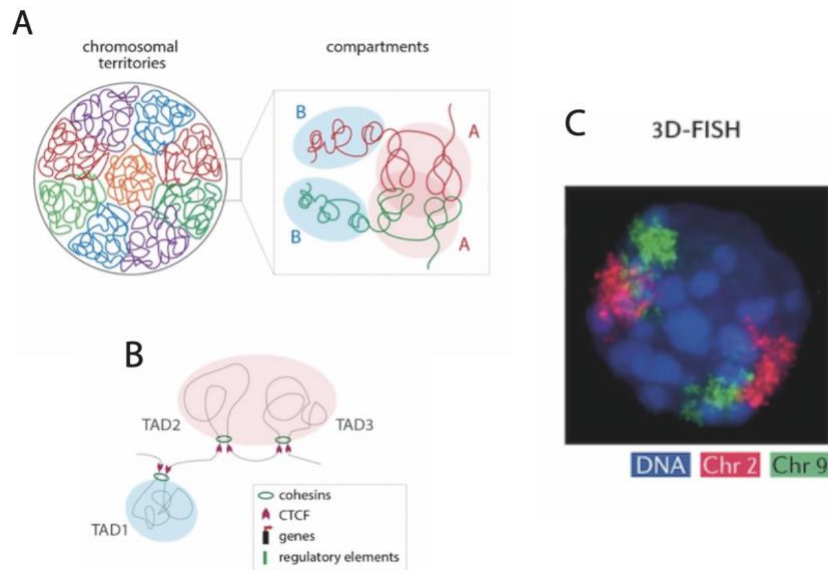


Figure 1.2. Compartmentalization of genomes into different levels of organization. A. Chromosome territories and chromosome compartments (A and B) inside them. B. Topological associated domains (TADs). C. Chromosomal territories captured with 3D-FISH. The 3D-fluorescence in situ hybridization (3D-FISH) image shows the positions of the chromosome territories of chromosome 2 (red) and chromosome 9 (green) within DAPI-stained nuclei (blue) from mouse embryonic stem cells. A and B images have been adapted from (Deakin et al., 2019) and C has been adapted from (Kempfer & Pombo, 2020).

In terms of chromatin interaction patterns within chromosomal territories, A and B compartments correspond to domains that interact preferentially with sequences in other A or B compartment regions, while TADs correspond to sequences that interact preferentially with themselves, rather than with other regions of the genome (Rowley & Corces, 2018). In addition, transcriptional control is mediated through physical contacts between enhancers and target genes, which occur via loop formation between the respective DNA elements. Functional loops between regulatory regions and genes are thought to occur predominantly within TADs (Kempfer & Pombo, 2020). This remarks the fact that the expression of genes can be influenced by their position within the 3D chromatin structure, and occasionally, patterns of proximity which were assembled by chance, may have provided functional advantageous and, consequently, being favoured by natural selection (Cremer & Cremer, 2010).

1.3.3 Chromosome mutations

Most species have a characteristic number of chromosomes, each with a distinct size and structure, nevertheless, variations in chromosome number or size do periodically arise. These

variations are called chromosome mutations and they might play an important role in evolution and speciation.

Chromosome mutations can be grouped into three basic categories: (i) chromosome rearrangements (CRs), which alter the structure of the chromosome; (ii) aneuploidies, which are produced when one or more chromosomes are duplicated or deleted (only cause alteration in chromosome number); and (iii) polyploidies, which are the result of one or more duplications of a complete set of chromosomes (Pierce, 2020).

1.3.3.1 Change in chromosome number: aneuploidies and polyploidies

A complete set of chromosomes possessed by an organism is called its karyotype and is usually presented as a picture of metaphase chromosomes lined up in descending order of their size, prepared from actively dividing cells (**Figure 1.3**).

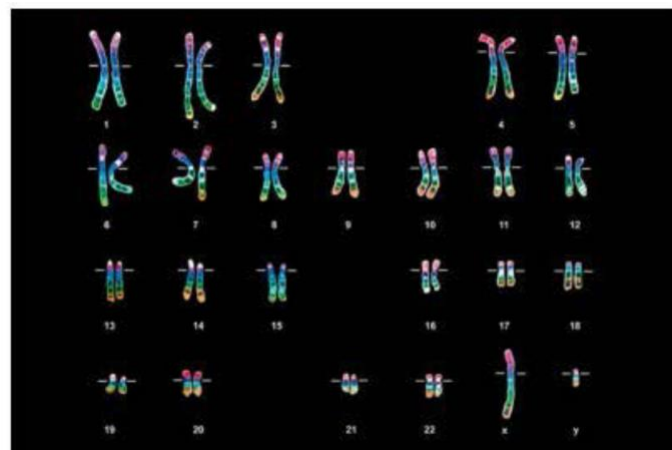


Figure 1.3. A human karyotype which consists of 46 chromosomes. This karyotype shows a male.
Figure from (Pierce, 2020).

Karyotypes are not completely immutable, and during cell division can mutate and suffer the gain or loss of one or more chromosomes (aneuploidy), or even the gain of a whole set of chromosomes (polyploidy). Aneuploidies can arise in several ways during the course of mitosis or meiosis and in human, can cause diseases such as the Down syndrome (when there is an extra chromosome 21). Polyploidy is most common in plants and is documented as the major mechanisms by which new plant species have evolved (Pierce, 2020).

These changes in chromosome number, aneuploidies, have been documented to be associated with cancer. Many tumour cells have extra or missing chromosomes, which can increase the rates of mutation at the DNA sequence or chromosomal level, causing a phenomenon collectively known as genomic instability (Baudoin & Bloomfield, 2021).

1.3.3.2 Change in chromosome structure: Chromosome rearrangements (CRs)

1.3.3.2.1 Origin of CRs

The other type of chromosome mutations, which alter the structure of individual chromosomes, are chromosome rearrangements (CRs). They usually are originated by erroneous repair of double-strand breaks (DSBs). This type of lesion is generated either by exogenous factors (i.e. ionizing radiation and/or chemical agents), endogenous agents (i.e. free radicals or stalls of the replication fork) or by highly specialised cellular processes such as meiosis. When the mechanisms in charge of their repair fail, the chance of CRs to appear is very high, causing an increase of chromosome instability in the genome. Those mechanisms are direct joining of incorrect DSBs and recombination of non-allelic homologous sequences (Schubert & Lysak, 2011).

Furthermore, specific DNA features have been associated with the occurrence of CRs, either as a cause or as a result of them: repetitive elements and telomeric repeats. Repetitive sequences of DNA can act as templates for non-allelic homologous recombination (NAHR). Within the repetitive elements, there are three specific cases that have been reported to play an important role in genome reshuffling:

- (i) Segmental duplications (SDs), also known as low copy repeats (LCRs). They consist on duplicated DNA fragments larger than 1 Kb in size that spread across the genome in a non-random distribution (She et al., 2008). They show >90% of sequence homology (Bailey & Eichler, 2006), explaining the propensity of these regions of promoting non-allelic homologous recombination (NAHR). Given their repetitive nature, SDs are an important source of genome instability by promoting the diversification and expansion of gene families or birthing new genes (Wilson et al., 2006). Furthermore, SDs are also involved in the occurrence of CRs implicated in human disorders (Lupski, 1998; Sharp et al., 2006) and chromosomal evolution (Murphy et al., 2005; Elsie et al., 2009). In

fact, they have been associated with gene families related to innate immunity, olfaction, and digestion (among others) that are commonly duplicated in mammals (Beckmann et al., 2007; G. E. Liu et al., 2009).

- (ii) Transposable elements (TEs). They were discovered by Barbara McClintock in 1984 when studying maize, and since then, they have been found in almost all organisms, independently if they are prokaryotes or eukaryotes. TEs consist of stretches of DNA which move around the genome of a cell, creating interspersed repeats throughout the genome (Biémont & Vieira, 2006). Due to their interspersed sequence homology, TEs favour genome instability related to genomic disorders and species speciation (Deininger & Batzer, 1999; Schibler et al., 2006; J. Lee et al., 2008; M. Farré et al., 2011). The formation of CRs by TEs can be due to the result of either direct or indirect DNA repair mechanisms (Wicker et al., 2007; Richard et al., 2008). TEs, as well as SDs, possess the ability of influencing genome plasticity by altering gene function and regulation or by creating new genes (Cordaux & Batzer, 2009). In the case of mammals, TEs related to CRs specific of different lineages have been studied, such as in Cetartiodactyla (Groenen et al., 2012) and Marsupialia (Longo et al., 2009).
- (iii) Tandem repeats (TRs). They consist of clusters of repeated sequences with sizes ranging from one to hundreds of base pairs per repeat. Along with TEs, they also represent a major source of DNA variation and mutation (Armour, 2006). It has been shown in human and primates studies, an association of CRs triggered by TRs and specific regions of the genome which are more susceptible to suffer breaks and re-joins (Ruiz-Herrera et al., 2006; M. Farré et al., 2011).

1.3.3.2.2 Types of chromosome rearrangements (CRs)

CRs encompass different types of events, and among the different ways of classifying them, they can be divided depending on their effect on the final dosage of a genome, in balanced or unbalanced CRs (Pierce, 2020).

Balanced rearrangements include those that do not change the final gene dosage of a genome: inversions, reciprocal and nonreciprocal translocations, fusions and fissions. In **Figure 1.4**, there is a sequence of rearrangements that explained how balanced rearrangements can completely change karyotypes in shape, size, and number.

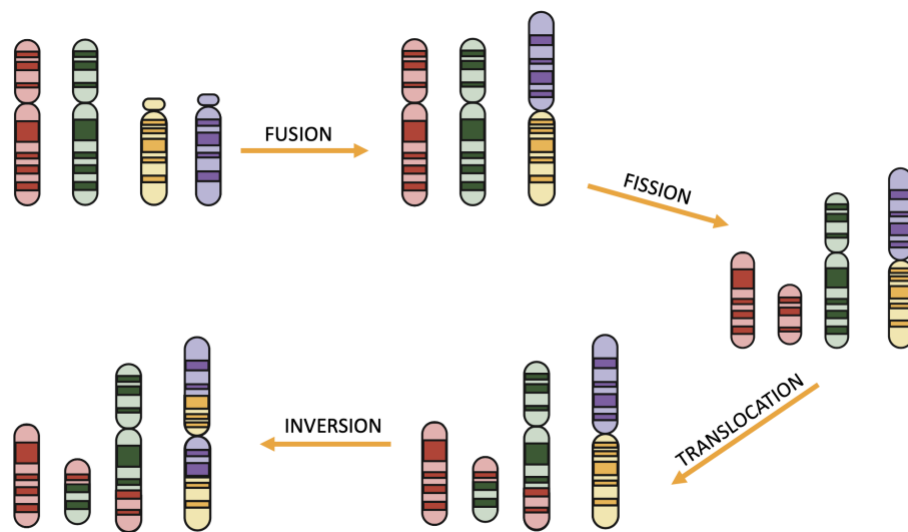


Figure 1.4. Sequence of balanced CRs. The figure illustrates how karyotypes can be modified in number, size or shape by CRs.

Fusions occur when two non-homologous chromosomes are combined to form a new chromosome. One of the most common types of chromosomal fusions found in nature is the Robertsonian fusion (also called centric fusion), that occurs when two acrocentric chromosomes fused by its centromeric region form a new single metacentric chromosome. Fissions occur when one single chromosome breaks into two different pieces. These fragments resulted from a fission, could join to another chromosome or could remain in the genome as a chromosome per se. Reciprocal translocations occur when there is an exchange of two genomic regions belonging to two different chromosomes. Translocations occur between non-homologous chromosomes and should not be confused with crossing overs during meiosis, which is the exchange of genomic material between homologous chromosomes. Instead, non-reciprocal translocations occur when the genetic material moves from one chromosome to another without any reciprocal exchange. Finally, inversions occur when an internal region of a chromosome changes its orientation (turns 180°) with respect to flanking regions. Depending on whether these changes involve or not the centromere, the inversions could be pericentric or paracentric, respectively.

On the other side, unbalanced rearrangements are those which involve an alteration in the final gene dosage of a genome: duplications and deletions. Duplications occur when a part of a chromosome is doubled. There are different types of duplications depending on the positioning of the duplicated fragment. When the duplicated region is immediately adjacent to

the original region, it is called tandem duplication. When the duplicated fragment is located some distance from the original segment, either on the same chromosome or on a different one, it is called displaced duplication. Deletions, contrary to duplications, occur when one fragment of a chromosome or even one entire chromosome is lost.

1.3.3.2.3 Comparing genomes between species: distribution of CRs among different species

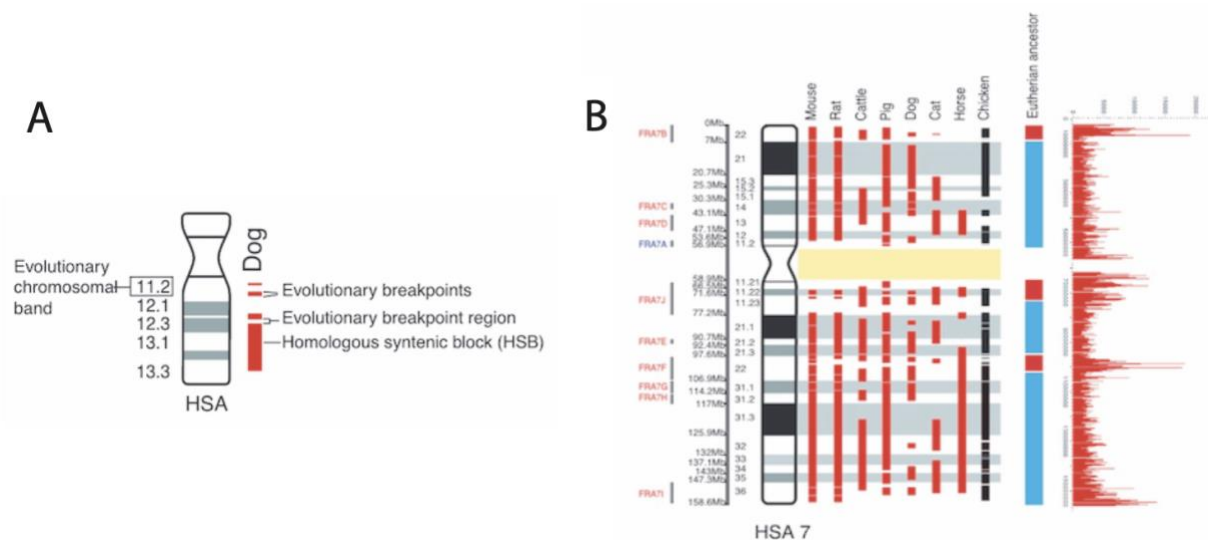
The mechanisms responsible for genome restructuring/reshuffling are still poorly understood. We know that there is an association between these changes and genome instability, but the causes of this instability remain yet unknown. The combination of cytogenetics with comparative genomics in the last decades, has been extremely useful in studies that have tried to explain genome structure and evolution, especially when trying to reconstruct phylogenetic relationships between closely and distantly related mammals (Murphy et al., 2005; Ruiz-Herrera et al., 2006; Larkin et al., 2009; Capilla et al., 2016; Kim et al., 2017; M. Farré et al., 2019; Deakin et al., 2019).

The first model aiming to explain how CRs took place and the effects of genome reshuffling was the “Random Breakage Model”, which tried to understand the dynamics of mammalian genome organization (Nadeau & Taylor, 1984). This model relied on two assumptions: (i) many chromosome segments were expected to be conserved between species, and (ii) CRs were randomly distributed within genomes. Their work was based on comparisons of human and mouse linkage maps based on the correspondence of over 80 orthologous genes. In their experiments they also described the presence of approximately 180 regions that were conserved among both species’ genomes. While the first premise was confirmed by other experiments, the second was questioned by a new model several years later. In the meantime, their proposal of uniform distribution of CRs among genomes was generally accepted for many years.

Years later, the “Fragile Breakage Model” was proposed (Pevzner & Tesler, 2003). This new model agreed with the “Random Breakage Model” in its first premise, the conservation of some large genomic regions in mammalian species, while it disagreed with the second proposal. This new model was against the random distribution of CRs within genomes, on the contrary, they proposed that there were some regions in the genome, “hotspots”, which were

more prone to break and reorganise. They also described that, for multichromosomal genomes, the most common rearrangements were inversions, translocations, fusions, and fissions, and furthermore, they named the number of such rearrangements in a most parsimonious scenario as the genomic distance. In their experiments, they identified 281 conserved regions between human and mouse genomes.

Other studies also obtained similar conclusions to the ones proposed by the “Fragile Breakage Model”. In one of these studies, they proposed the name “homologous synteny blocks” (HSBs) to designate those regions of the genome that were conserved among species (Murphy et al., 2005). In other study, they designated the “hotspots”, these regions described as susceptible to suffer breakages on them, as evolutionary breakpoint regions (EBRs), and also described how most of them overlapped with fragile sites of genomes. These fragile sites were defined as heritable loci located in specific regions of chromosomes that were expressed as gaps or breaks when cells were exposed to specific conditions such as chemical agents that act as inhibitors of DNA replication or repair (Ruiz-Herrera et al., 2006). On **Figure 1.5** we can see how HSBs and EBRs are described as a result of a multispecies comparison analysis.



Subsequent studies proposed the “Intergenic Breakage Model”, which considered that evolutionary rearrangement breakages happened with a uniform propensity along the genome but were deleterious in genes, their regulatory regions and in-between (Peng et al., 2006; Becker & Lenhard, 2007). By simulating CRs in human and mouse genomes randomly and analysing the genomic distribution of both intergenic and regulatory regions, they proposed that genomic distribution of EBRs could be also affected by the combination of long regulatory regions and the distribution of intergenic region sizes. Furthermore, they also suggested selection as a mechanism that would act against EBRs located in genes, and that could cause changes in their regular expression. In fact, this model proposed EBRs as regions of genomic instability, not only because of their intrinsic DNA but also because their gene composition. As mentioned before, for the “Intergenic Breakage Model”, EBRs were located in regions of the genome where negative selection is minimal. According to this premise, the DSBs causing these EBRs not only would not have deleterious effects but also they would be fixed within populations.

In addition to all the premises proposed by the “Intergenic Breakage Model”, other studies showed that the susceptibility to breakage is highly heterogeneous along human chromosomes and proposed the susceptibility of “open” chromatin regions to breakage as a third force to add to previous theories trying to explain the role of rearrangements in evolution (Lemaitre et al., 2009). The “open” state of the chromatin not only would facilitate the appearance of DSBs and posterior CRs, but also would have consequences in gene expression.

Despite of the study of both models, the “Fragile Breakage Model” and the “Intergenic Breakage Model” as independent models with different premises, they are not mutually exclusive. This reason was the trigger that impulsed the suggestion of a new model which tried to put together all their common factors to give a better explanation to the role that CRs play in genomes, and its consequences in adaptation and speciation.

The new model proposed to explain the importance of CRs was the “Integrative Breakage Model”, which brought together several factors as key elements in determining the genomic distribution of EBRs: (i) DNA sequence composition, (ii) the analysis of the high-level structural organization of genomes, and (iii) the effect on gene expression (M. Farré et al., 2015). The basic premise which this model proposed was that certain properties of DNA sequences (such as the presence of repetitive sequences in their content), together with the

epigenetic state of the chromatin, could promote the change of chromatin (specially if this chromatin reached an “open” conformation), and this way it would contribute to the origin of CRs and their posterior consequences. The “Integrative Breakage Model” explained the mechanisms that would facilitate genome reshuffling and the conditions that make CRs more prone to occur within a genome. According to this model, CRs and posterior genome reshuffling would occur in regions of the genome that: (i) physically interact inside the 3D nuclear space (this is related to the 3D organization of the nucleus in chromosomal territories and A and B compartments within them); (ii) whose chromatin have an accessible state promoting genome instability; and (iii) do not disturb essential genes.

However, the possibility that genome reshuffling could disrupt genes essential for survival, and subsequently be subject to purifying selection has been reported (Becker & Lenhard, 2007; Lemaitre et al., 2009; Ruiz-Herrera et al., 2010), and in these cases, EBRs can represent opportunities for the development of novel functions that may promote the adaptation of species. This would be consistent with the idea of a connection between mammalian EBRs and the development of new adaptive gene functions, such as in the immune system or olfactory receptors (Larkin et al., 2009; Groenen et al., 2012; Capilla et al., 2016).

Overall, nowadays we know that during evolution, some regions of the genome are highly conserved among several species, the so called HSBs. On the contrary, other regions are more prone to suffer breakages and re-joins, originating genome reshuffling. These regions are the EBRs, which have been found to be associated to fragile sites of the genomes, which proves that EBRs are not randomly distributed across genomes. In fact, these EBRs need specific conditions to be originated, and these conditions are not only related to specific DNA sequences but also are related to location within the 3D organization of the nucleus, chromatin state, and gene expression. Moreover, the characterisation of how chromatin conformation and DNA-protein interactions have evolved during speciation is providing a new interpretative hypothesis on the mechanism(s) responsible for the origin of genome architecture and plasticity (Álvarez-González, Burden, et al., 2022).

1.3.3.2.4 Consequences of CRs

CRs and the subsequent genome reshuffling can have several effects on genomes. One of the characteristics that can be affected by CRs is karyotype organization. When karyotypes

alterations occur in the germ line, they have the potential to be fixed during evolution. This is one of the mechanisms responsible for the distinct karyotype organisation between species, and sometimes, even within individuals of the same species. That is one of the main reasons to reconstruct the evolutionary history of chromosome relationships between species, to be able to understand how CRs play a role as evolutionary forces with the potential of shaping genomes.

These karyotype alterations have the potential to modify chromosome number but also, they have the potential to modify chromosome size. One of the processes that regulate the size of the chromosomes is chromatin compaction, and as it has been said before, CRs are produced in specific regions of the genome that are more prone to break and re-join (EBRs) and which also have been linked to the 3D structure of the chromatin.

As described before, there is an association between EBRs and fragile sites of the genome. These fragile sites are associated, at the same time, with the amplification of repeat motifs (such as CCG repeats and A-T rich regions), which are features that may affect DNA replication and chromatin condensation (Ruiz-Herrera et al., 2006). Moreover, posterior studies established that the distribution of the EBRs would be associated with an “open” (also considered decondensed) state of the chromatin. Furthermore, they related this “open” chromatin state with specific characteristics such as high transcriptional activity, early replication, and subject to DNA hypomethylation (Lemaitre et al., 2009). This correlation between EBRs and hypomethylated regions of the genomes, has been considered one of the causes that would make chromatin in these regions more accessible to transposons being integrated into these regions (Buenrostro et al., 2013).

Not only TEs have been reported to be more abundant in regions near EBRs, there also have been reported higher numbers of SDs and copy number variants (CNVs). This fact remarks the relationship of EBRs and the increase of genomic variation, which probably act together as a major reservoir for producing adaptive phenotypes by evolutionary CRs (Larkin et al., 2009).

But CRs not only have effects in karyotype shape or number, they also have effects on genomes at the molecular level. One of these effects is the suppression of recombination within the rearranged region (Navarro & Barton, 2003; Joron et al., 2011; M. Farré et al., 2013). Direct and indirect evidence of this suppression of recombination have been reported. Within the

direct evidence it is included the analysis of recombination in the gametes and/or the offspring when reorganization (inversions and/or translocations) occurs (Faria & Navarro, 2010). In fact, it is well documented that in humans, somatic and germline genomic rearrangements involving EBRs may cause major disease phenotypes, including a variety of cancers and developmental disorders (Murphy et al., 2005). On the other hand, indirect evidence of this suppression of recombination included the analysis of rates of genetic divergence between rearrangements and collinear chromosomes (Navarro & Barton, 2003). High rates of sequence divergence detected in genes located at, or near, CRs have been interpreted as indirect evidence of chromosome speciation through suppressed recombination (Marques-Bonet et al., 2007; M. Farré et al., 2013). The main cause of this suppression of recombination would be an incomplete pairing during meiosis, which could lead to the accumulation of genetic incompatibilities, and in some cases, speciation (M. Farré et al., 2019).

Some studies have been able to detect this reduction of recombination rates within the EBRs in mouse genomes (Capilla et al., 2016). They postulated that low recombination rates within EBRs could lead to a high genomic differentiation and the fixation of new mutation in genes related to the species-specific phenotypes (such as genes involved in mating and individual recognition, reproductive isolation, and oxidative stress). Their theory would reinforce the adaptative value of genome reshuffling caused by CRs.

But not only gene location is affected by CRs, some studies have proved the relationship of genome reshuffling and gene expression. The “Integrative Breakage Model” explained before, postulated how DSBs were one of the main causes of CRs, and only the DSBs which did not affect or modify the expression of genes related to developmental or basic cellular maintenance (i.e. housekeeping genes), would have the potential to become fixed, remarking their evolutionary significance (M. Farré et al., 2015). In fact, they saw that EBRs were more prone to co-localize with genes that facilitated adaptative responses. Previous studies of Robertsonian fusions in mice already established that breakpoints occurring within a genomic region with a high concentration of housekeeping genes and/or tissue-specific genes would impact on the cell’s ability to maintain its basic central functions, and its persistence in meiosis would be unlikely, making the rearrangements caused by these breakages unable to pass to the offspring (Ruiz-Herrera et al., 2010). Additionally, the “Integrative Breakage Model” also showed that conserved HSBs between multiple species, were enriched for gene networks which controlled embryonic and tissue development (basic functions necessary for the organisms) (M. Farré et

al., 2015). However, these findings would still let some unresolved questions, like whether these large HSBs have remained intact for long evolutionary time periods by chance or due to selection.

Furthermore, EBRs have the potential to disrupting coding sequences or altering gene expression of adjacent genes by separating them from their regulatory elements, bringing new regulatory sequences, moving the genes to different regulatory domains, or reconfiguring chromatin interactions by rearranging TADs (Puig et al., 2015a; Krefting et al., 2018; M. Farré et al., 2019). All these relationships make the search for non-random chromatin assemblies, the mechanism(s) responsible for their formation, and their functional implications, one of the major goals of nuclear architecture research (Cremer & Cremer, 2010).

The relationship between CRs, chromatin configuration, and changes in gene expression has been demonstrated in some species (Marques-Bonet et al., 2007; Fuller et al., 2016; Capilla et al., 2016), without attempting any explicit ancestral reconstruction or inferring the origin of the genomic or epigenomic features. However, reconstructing ancestral genomes and stablishing evolutionary trajectories of karyotypes between species are key to explain this relationship. Tracing the chromosome organization from a common ancestor allows to detect and classify CRs that would not be possible to detect only by comparing extant species, and subsequently define EBRs and HSBs within the lineages involved in the ancestral reconstruction. Moreover, comparing current species with an ancestral form, not only increases the number of inter- and intrachromosomal rearrangements that could be detected, but also facilitates the timing when these CRs appeared in evolution. Additionally, ancestral karyotype reconstructions enable the estimation of the rearrangement rates leading the evolution of the different lineages, as well as the classification of the CRs, EBRs and HSBs.

Likewise, ancestral reconstructions are key to address unresolved questions such as: Why do chromosome numbers vary greatly within some species of the same group, but are largely the same in others? Why are genomes so extensively changed among closely related species and others strongly conserved? Why do some CRs appear to lead to speciation? Which are the mechanisms behind speciation? Is the underlying mechanism responsible for chromosomal speciation the same as that leading to CRs in a disease context (i.e., cancer)?

1.4 Studying chromosome evolution: From cytogenetics to genomics

1.4.1 The origin of the first cytogenetic techniques

The study of genome organization, the mechanisms responsible for CRs, and their connection to adaptation and speciation, are key points to better understand the evolutionary dynamics of genomes.

Relevant technological advances have been achieved for the study of CRs. Early cytogenetic studies in the early 1970s were focusing on banding patterns comparisons which permitted to identify and differentiate chromosomes among different species, specifically differences in karyotypes between great apes (Caspersson et al., 1970; De Grouchy et al., 1972). As mentioned before, karyotyping started with the staining of chromosomes in metaphase during meiosis, followed by their observation by microscopy. The basis of this study was the differential staining levels of chromatin, which produced different patterns of dark and light stripes along chromosomes. This technique allowed the recognition of specific chromosomes by their unique banding pattern. For instance, one of the most common banding techniques, G-banding, shows chromosomes as a configuration of dark and light bands depending on the type of chromatin, heterochromatin, or euchromatin, respectively (Montazerinezhad et al., 2020). Thanks to these banding techniques, breakpoints involving translocations, inversions, insertions, and deletions have been detected over time.

In the mid 90s, a new high-resolution technique was developed, fluorescence in situ hybridization (FISH). It consisted in the use of DNA probes corresponding to whole chromosomes, known as chromosome paintings, that can be hybridized on chromosome preparations of target species (Scherthan et al., 1994). In the early 2000s, these probes were often obtained from bacterial artificial chromosomes (BACs), which are DNA molecules artificially created to clone DNA sequences in bacterial cells. The BAC clones possess a known DNA sequence from the reference species, they are then fluorescently labelled and hybridised to the target species, allowing the observation of their location within chromosomes using a fluorescence microscope (Speicher & Carter, 2005). In this way, it was possible to detect conserved chromosome regions between two given species based on DNA sequence homology and the use of one of the variation of FISH. Additionally, FISH allowed not only the detection

of homologous regions between different genomes, but also the detection of karyotypical aberrations associated with diseases (**Figure 1.6**).

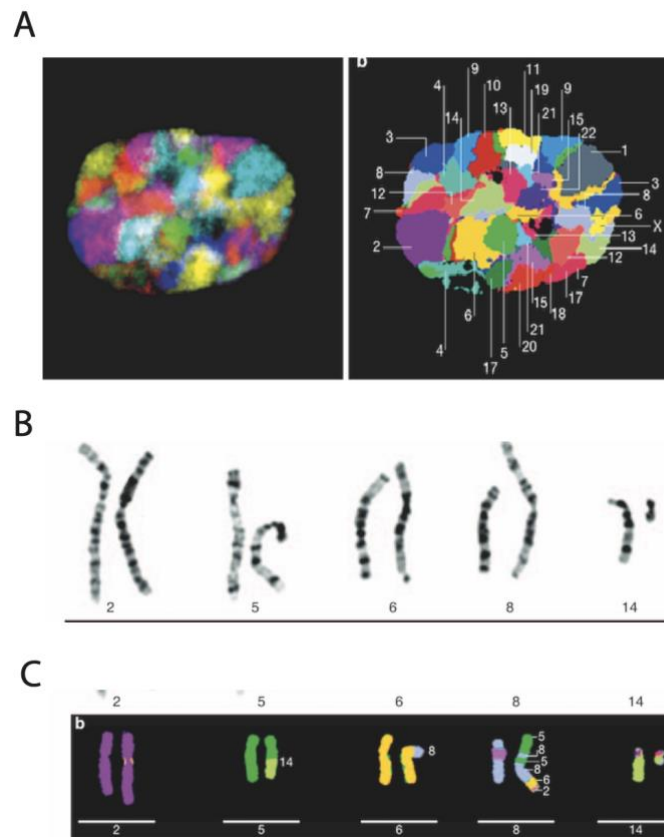


Figure 1.6. FISH performed in human chromosomes to detect complex CRs related to disease. A. All the chromosome territories in the human genome can be visualised simultaneously in intact interphase nuclei, each in a different colour. Every number corresponds with human chromosomes. B. Chromosome banding analysis that reveals aberrant banding patterns related to disease. C. Multiplex FISH identifying that the aberrant patterns in panel B are due to CRs that affect several chromosomes. Figure adapted from (Speicher & Carter, 2005).

Chromosome painting has allowed the comparison between many species representative of the major phylogenetic groups, including birds (Guttenbach et al., 2003; Griffin et al., 2008), primates (Stanyon et al., 2000; Ruiz-Herrera et al., 2002), rodents (Veyrunes et al., 2006; Graphodatsky et al., 2008), ruminants (F. Yang et al., 1997; Fröncke & Wienberg, 2001; Balmus et al., 2007; Kulemzina et al., 2011, 2014), marsupials (Toder et al., 1997, 1998; O'Neill et al., 1999; Glas, 1999; Rens et al., 2001), and afrotherians (F. Yang et al., 2003; Fröncke et al., 2003; Pardini et al., 2007). Moreover, using chromosome painting, also known as Zoo-FISH, enables the detection of conserved regions within genomes. This technique can detect conserved chromosome fragments that now are present in independent chromosomes

but originally came from a single chromosome in the common ancestor, or whole chromosomes that remain intact when comparing with the ancestral karyotype.

1.4.2 Comparative genomics: from past to current techniques

Despite of all the methodological advances of molecular cytogenetic techniques, FISH probes can only hybridize with closely related species. In fact, for distantly related species, FISH probes showed a lack of resolution that is the main cause for not allowing a detailed characterization of chromosome breakpoints in distant related species.

1.4.2.1 Genetic maps

In order to fill this gap, other methods have been developed. One of these methods is genetic mapping, which has been used to identify homology regions between chromosomes of different species. What makes this approach more accurate is that not only the chromosomal regions are identified, but also the gene order within chromosomes is obtained. On one hand, high-density molecular linkage maps have been used to characterise the relative position of two loci based on the frequency they are observed together in gametes. On the other hand, radiation hybrid maps estimate the distance of two loci after breaking chromosomes into fragments that can be studied in hybrid cells (Pierce, 2020). Both mapping methods are based on the assumption that proximity influences inheritance, that is, those loci physically close on chromosomes tend to be inherited together, those loci far from each other on the same chromosome tend to be inherited together less often, and loci from different chromosomes are inherited independently. A third different type of genetic mapping is the cytogenetic map, which consist in the hybridization of a fluorescently labelled DNA probe containing a marker of interest (usually a BAC clone) that maps to a chromosome preparation (Pierce, 2020). The main weakness of cytogenetic mapping is that it grossly localizes a marker of interest within a chromosome of one species but without a high-resolution order.

Even with their limitations, genetic maps have been an essential tool for the assembly of new genomes. They contribute to the correction, merging and ordering of the scaffolds obtained by sequencing methods to generate chromosome-level genome assemblies (Murphy et al., 2005; Griffin et al., 2008; Graphodatsky et al., 2008; Damas et al., 2017).

1.4.2.2 Genome sequencing

Whole-genome sequencing has allowed researchers to obtain a wide catalogue of genomic information that can be used to elucidate questions related to how species have evolved and how their genomes have been modified, and how adaptation and speciation occurred. The initiation of the Human Genome Project, and after the publication of the first draft of the human genome (Lander et al., 2001), new advances in genome sequencing together with the appearance of new mathematical algorithms, transformed the traditional way of studying genome structure and evolution.

In this regard, genome sequencing technologies, especially next-generation sequencing (NGS), have transformed the way evolution has been studied in biology. The appearance of both, small and large-scale sequencing projects, have created molecular sequence databases which include genomes complete or partially sequenced of thousands of organisms. Over the past decade, NGS technologies have continued to evolve in order to provide genome reads as long as some entire genomes (Goodwin et al., 2016). At the same time, these new technologies have been improved in order to reduce both, the time and the cost of their use, and because of that, they have become routinary techniques to perform in biology laboratories.

These techniques can be divided into short- or long-reads NGS techniques, depending on the length of the DNA molecule sequenced. On one hand, short-reads sequencing approaches, such as Sanger sequencing, Illumina, or Ion Torrent, generally started with DNA fragmentation, DNA end-repair, adapter ligation, surface attachment, and in situ amplification (Hu et al., 2021). These short-reads sequencing technologies involve the massively parallel sequencing of reads up to 800 bp. These then need to be assembled to form long stretches of continuous DNA. Because of the short size of molecules, the assembly process can present challenges in regions containing repetitive elements. On the other hand, long-reads sequencing technologies, also called third-generation sequencing technologies, are able to generate sequences over 10,000 bp long directly from native DNA (Goodwin et al., 2016). Two of the primary long-read sequencing technologies are Pacific Biosciences (PacBio) and Oxford Nanopore Technology (ONT). PacBio long-reads can generate molecules up to 300 Kb, and have been reported to have an accuracy up to 99.8%, with a precision rate of 99.91% to detect single-nucleotide variants (SNVs), 95.98% to detect insertions and deletions, and 95.99% to detect structural variants (Hu et al., 2021). ONT can generate long-reads greater than 3 Mb, but

the accuracy of this methodology to detect SNVs is slightly lower. The accuracy of ONT long-reads is dependent on the base-calling (translation of the electrical signal to DNA sequence) algorithm that is used (Hu et al., 2021).

Overall, the general advantages of long-read sequencing technologies compared to short-read sequencing technologies include better phasing of polymorphic genes, detection and proper characterization of structural rearrangements, data collection in real-time, and faster turnaround times. Additionally, the potential errors that could be introduced during the DNA amplification process of short-reads are completely eliminated when using long-reads technologies. Furthermore, when comparing PacBio with ONT, the second has been proved to require instruments with lower cost, portable, and of significantly smaller size, which can be very useful for low-income settings or field applications (Hu et al., 2021).

Two of the pioneer projects whose aims are to sequence a great number of genomes were: (i) the 1,000 Genomes Project, whose purpose was to obtain whole-genome sequences from 1,000 humans in order to discover and understand human genetic variation (Siva, 2008); and (ii) the Genome 10K Project, whose main objective was to obtain 10,000 whole-genome sequences of different vertebrate species (Genome 10K Community of Scientists, 2009). Both projects are still ongoing. The 1,000 Genomes Project has already completed its phase 3 with the sequencing of 3,202 individuals and the calling of 111,048,944 SNVs (Byrska-Bishop et al., 2022). The Genome 10K Project is organized in several specialised projects such as the Bat 1K, the Bird 10K or the Vertebrates Genome Project (VGP). VGP aims to generate high-quality, complete reference genomes for all of the roughly 70,000 extant vertebrate species and to help to enable a new era of discovery across the life sciences (Rhie et al., 2021). Furthermore, other projects, such as The Earth BioGenome Project (EBP), have proposed that all known eukaryotic species on our planet should be sequenced as a resource for future bioscience and environmental stewardship (The Darwin Tree of Life Project Consortium, 2021). EBP is a global collaboration of groups generating data from species selected by taxonomy or location, such as the Darwin Tree of Life Project (DToL), whose aim is to sequence high quality genomes of all known eukaryotic species in a defined geographical region: the islands of Britain and Ireland. In May 2023, DToL have successfully sequence 2,749 individuals (Darwin Tree of Life Data Portal, 2023).

1.4.2.3 Genome assembly and annotation

However, despite the availability of many sequencing techniques with a high accuracy, the acquisition of a complete chromosome-level assembly of a genome is not an easy task. The ultimate aim of assembling a genome *de novo* is to obtain a set of contigs, each of them representing an entire chromosome end-to-end. Depending on the sequencing methodology used, the first step of genome assembly generally produces hundreds or thousands of contigs (defined as contiguous gapless stretches of DNA sequence assembled from smaller overlapping sequenced fragments (raw reads) (Lewin et al., 2019)), whose order and orientation relative to one another cannot be further described, usually due to regions in the genome that are too repetitive to determine how to extend them (Rice & Green, 2019). This type of assembly is still very useful for many purposes such as comparing protein sequences between related species or performing gene expression studies, but chromosome-level assemblies are needed to study chromosome evolution.

Genome annotation is directly influenced by the quality of genome assemblies (Armstrong et al., 2019), because of this, reaching a chromosome-level assembly would facilitate the detection of repetitive elements or other copy-number or structural variants (Rice & Green, 2019), as well as the correct identification of genes.

1.4.2.4 Computational approaches: The origin of bioinformatics

The foundations of bioinformatics were laid in the early 1960s with the application of computational methods to protein sequence analysis, as reviewed in (Gauthier et al., 2019). Computers and specialized software have become essential for biologists, either for routine DNA or protein sequence analysis, or for the management of massive gigabyte-sized biological data sets. But with the development of NGS methods, it was introduced the necessity of creating and improving new algorithms for analysing data.

The American physical chemist Margaret Dayhoff was pioneer on applying computational methods to the field of biochemistry. Together with the physicist Robert S. Ledley, developed “COMPROTEIN”, which was a complete computer program designed to determine protein primary structure using Edman peptide sequencing data (Dayhoff & Ledley, 1962).

Subsequently, other scientists started to wonder if it was possible for proteins to evolve and if they would be able to reconstruct the evolutionary history of those proteins at the same time that reconstruct the sequence of their ancestors. This was the beginning of the “Paleogenetics”, term coined by Emile Zuckerkandl and Linus Pauling in 1963 (Pauling & Zuckerkandl, 1963). Both of them observed that orthologous proteins from vertebrate organisms, such as haemoglobin, showed a degree of similarity too high over long evolutionary time to be the result of either chance or convergent evolution. Furthermore, they observed that the number of differences in orthologs from different species seemed proportional to the evolutionary divergence between those species. For instance, they observed that human haemoglobin showed higher conservation with chimpanzee (*Pan troglodytes*) haemoglobin than with mouse (*Mus musculus*) haemoglobin, showing a correspondence between these results and divergence estimations derived from the fossil record. Zuckerkandl and Pauling hypothesized that orthologous proteins evolved through divergence from a common ancestor, and that by comparing the sequence of haemoglobin in currently extant organisms, it could be possible to predict the ‘ancestral sequences’ of haemoglobin and, in the process, its evolutionary history (Hagen, 2000; Gauthier et al., 2019).

Since that moment, many mathematical algorithms, tools, software, etc., have been developed to trace back genome histories. The study of current species genomes, the comparison between them, and their association with their phylogeny, have become fundamental in order to further study evolution.

1.4.3 Ancestral karyotype reconstructions

As mentioned previously, CRs not only originate changes in the structure of chromosomes but also have the potential to originate new phenotypes that can be fixed in following generations and sometimes, even lead to speciation. That is why tracing how chromosome organization is in a common ancestor is paramount to describe the role of CRs in evolution and speciation. When this idea of reconstructing common genome ancestors to study different characteristics between species was first proposed (Pauling & Zuckerkandl, 1963) the methodology available for this purpose was very poor. Nowadays, especially with the increase of computational techniques, it is possible to reconstruct ancestral genomes with a high accuracy.

The first attempts to reconstruct ancestral genomes were performed using cytogenetics and genetic mapping. These studies used chromosome banding, FISH probes and/or genetic maps, and they worked with the same premise: when comparing different species with a common ancestor, all regions shared between their genomes should be considered a representation of their common ancestral relative. One of these first studies which tried to reconstruct an ancestral karyotype was performed by comparing chromosome banding data from several platyrrhine monkeys (Dutrillaux & Couturier, 1981). They were able to detect several intraspecific variations, such as inversions, translocations, or variation in the amount of heterochromatin which could affect chromosome morphology. They initiated their ancestral karyotype reconstruction following one rule: “when a chromosome is identical in several species, no mutation has occurred, and the identity of chromosomes reflects the persistence of an ancestral chromosome in two or more independent species”. Using those criteria and the phylogenetic relationships between the species, they managed to reconstruct the common ancestral karyotype of platyrrhine monkeys. Posterior studies have proposed a mammalian ancestral karyotype after comparing chromosome painting probes of human, African and Asian elephants (*Loxodonta Africana* and *Elephas maximus*, respectively) and armadillo (*Oryzomys azeri*) (F. Yang et al., 2003). They chose these species because elephants and armadillos are mammalian species belonging to the Afrotheria superorder, whose radiation from other mammals is known to be rooted in Africa and is strongly supported by DNA sequence. With their analysis they were able to establish HSBs conserved between human and afrotherians, both as single chromosomes or as fragments, and they also established their possible karyotype evolution, originating in this way, one of the first ancestral karyotypes proposed for a mammalian ancestor. Other studies attempted to reconstruct the ancestral karyotype of eutherian mammals by comparing chromosome banding performed by zoo-FISH, in this case they identified HSBs conserved between 29 species belonging to nine different mammalian orders and they used human as reference species (F. Yang et al., 2003). Some other studies have used a combination of both, FISH probes and genome mapping, to try to reconstruct the ancestral eutherian genome organization (Frönicke et al., 2003). In this specific case, they painted human chromosomes with zoo-FISH probes derived from elephant and vice versa and then compared their results with previous mapping of the African elephant, armadillo and human genomes (F. Yang et al., 2003). This combination of both techniques allowed them to obtain more details regarding smaller conserved segments. Furthermore, these reciprocal painting allowed them to identify the segment-to-segment homologies between elephant and human karyotypes and, also confirm and amplify the list of CRs that had been proposed for a

common ancestor of eutherian species and provided first insights of genome organization of the ancestor of extant placental mammals, the proto-afrotherian ancestor.

These pioneer reconstructions clearly showed that genome structures in different taxa evolved in different ways and, were also able to relate mechanisms such as CRs or chromatin conformation to chromosome evolution and speciation. However, these ancestral reconstructions based only in cytogenetic methods presented two main limitations. The first one was the lack of resolution, which resulted in many CRs, especially intrachromosomal rearrangements, not being detected. FISH probes have a resolution of 100 Kb (Huber et al., 2018) meaning that smaller CRs cannot be detected. The second limitation was the lack of hybridization of chromosome paintings between distantly related species, making unfeasible the reconstruction of ancestral karyotypes at evolutionary deep nodes.

The availability of new genome assemblies, combined with the development of new algorithms, allow, nowadays, the reconstruction of ancestral genomes at further evolutionary depth with higher resolution. The first attempt to reconstruct ancestral karyotypes from sequence data was a study that used genomes from different rodents, human, and chicken (as an outgroup) to reconstruct the mammalian ancestor (Bourque et al., 2005). They performed an analysis using an algorithm called GRIMM-synteny (Pevzner & Tesler, 2003), which detects HSBs between the species. Once the HSBs were identified, they used them as input for GRIM (Tesler, 2002) and MRG (Bourque & Pevzner, 2002), both algorithms reconstruct the likely rearrangement evolutionary scenarios, considering inversions, translocations, fusions, and fissions. In this process, MGR also estimates the number of rearrangement events on the evolutionary tree and suggests the reconstructed genomes of different ancestors, the one for rodent species and the one for all mammal species. So, in this first attempt, by comparing their genome sequences, they were able to identify the synteny blocks with high accuracy and they reconstructed different ancestors at the same time.

Posterior studies combined both, sequencing and genetic mapping, to reconstruct the ancestor of several mammalian species: human, horse (*Equus caballus*), cat (*Felis catus*), dog (*Canis lupus*), pig (*Sus scrofa*), cow (*Bos taurus*), rat (*Rattus norvegicus*) and mouse (*Mus musculus*) (Murphy et al., 2005). This study detected lineage-specific EBRs, and the authors were able to associate some EBRs with high gene density regions and the location of SDs. Posterior studies also showed the improvement of genetic mapping, with a new methodology

which allowed to reconstruct ancestral karyotypes by using physical and virtual mapping (Deakin et al., 2013). The virtual comparative maps allow to compare gene order between species and were performed with the tool AutoGRAPH (Derrien et al., 2007). In this study, they reconstructed the marsupial ancestral karyotype comparing three marsupial species, tammar wallaby (*Macropus eugenii*), common opossum (*Didelphis marsupialis*) and fat-tailed dunnart (*Sminthopsis crassicaudata*), with human and chicken. They were able to detect intrachromosomal rearrangements occurred after the divergences of every specific species, as well as common HSBs and EBRs. Furthermore, they also reconstructed the ancestor of all mammalian species with a higher accuracy than previous studies.

Subsequently, the algorithms used for ancestral reconstructions based on genome sequence data have been improved and can take two main approaches: (i) global parsimony methods that rely on the minimum number of genome rearrangements required to obtain the synteny of extant genomes, and (ii) the local parsimony methods that infer the most parsimonious scenario for each individual adjacency. Some examples of software used in ancestral karyotype reconstructions are InferCARS (Ma et al., 2006), AGNES (Jones et al., 2012), AGORA (Muffato, 2010; Muffato et al., 2022), and DESCHRAMBLER (Kim et al., 2017).

Among all these recent developed algorithms, only DESCHRAMBLER can reconstruct complex genome histories, including duplications and deletions (Kim et al., 2017). Moreover, DESCHRAMBLER is the only algorithm able to work with fragmented genomes and obtain high-resolution results. What DESCHRAMBLER generates are reconstructed ancestral chromosome fragments (RACFs) using synteny blocks (also called synteny fragments (SFs)) previously identified from whole-genome alignments. DESCHRAMBLER can process a great number of species, and can work with both, chromosome-level or scaffold-level genomes. The first time this program was used, was to reconstruct the chromosome structure of seven different mammalian ancestors and was applied to the genomes of 21 species (14 with chromosome-level assemblies and seven with scaffold-level assemblies), and when comparing this analysis with previous studies, not only many rearrangements were confirmed but also new rearrangements were detected (Kim et al., 2017).

Posterior studies have used DESCHRAMBLER to reconstruct different ancestral mammalian karyotypes, using cattle as reference genome, and were able to associate that genes near EBRs detected in the ruminant lineage exhibited more divergent expression profiles

among species, which was consistent with the phylogenetic origin of those EBRs (M. Farré et al., 2019). They also observed that the liver cell putative enhancers in the ruminant EBRs were highly enriched for DNA sequences under selective constraint acting on lineage-specific TEs, proving that EBRs may contribute to changes in regulation of the nearby genes during evolution by providing high-density of TE sequences.

One of the most recent studies using DESCHRAMBLER, included eight scaffold- and 26 chromosome-scale genome assemblies representing 23 of the 26 mammalian orders to computationally reconstruct ancestral karyotypes and syntenic relationships at 16 nodes along the mammalian phylogeny, and as a new approach, they used three different reference genomes (human, sloth (*Choloepus didactylus*), and cattle) representing phylogenetically distinct mammalian superorders (Damas et al., 2022). In this study, they proposed prezygotic purifying selection as the evolutionary mechanism preventing meiotic CRs that disrupt HSBs. They also observed that EBRs are enriched for SDs and TEs, while the TAD structure of HSBs is maintained during evolution. Moreover, they established trends in the rates of CRs which supported the association between CRs and speciation.

Overall, DESCHRAMBLER algorithm has created new opportunities for studying genome evolution due to its ability to handle both, chromosome-level and scaffold-level assemblies, and after using different reference species it has been showed an increase in the resolution and robustness of the ancestral karyotype reconstructions. Nonetheless, the accuracy of DESCHRAMBLER to detect EBRs, is directly dependant of the quality of the assemblies used as input species.

1.5 Mammals: our target species

According to the Mammal Diversity Database (MDD), an open database originally released in 2018 and which provides a continuously updated listing of the world's mammal species and higher taxa (Burgin et al., 2018; Mammal Diversity Database, 2022), there are a total of 6,495 living mammalian species, arranged in 1,342 genera, 167 families, and 27 orders. Within mammals, placental mammals are the crown clade and comprise the vast majority of living mammals, with more than 6,100 species (Murphy et al., 2021). Placental mammals show a great diversity of morphologies and sizes, for instance, species can range from 190 tones in deep sea whales to 2 grams in many bat species, and species colonize different environments

due to their different capabilities, such as the ability to flight, to dive, to swim, to walk, and even to build subterranean tunnels (Murphy et al., 2021). Mammals have been extensively studied and in the next sections I will be exploring what is known about them and which questions remain still unresolved.

1.5.1 Phylogeny

Mammalian genome evolution needs to be interpreted in a phylogenetic context. Among vertebrates, phylogenomic investigations have focused principally on mammalian genome evolution, increasing the availability of chromosomal and genomic information for this clade substantially.

Genome-scale data have helped to resolve many phylogenetic problems across the tree of life, including the deep relationships among placental mammals (Meredith et al., 2010; Song et al., 2012). Resolving the phylogeny of eutherian mammals has been challenging, but in the past decade, the division of eutherian mammals into four superorders (Euarchontoglires, Laurasiatheria, Afrotheria, and Xenarthra) has been well supported (**Figure 1.7**) (Song et al., 2012). Other mammalian groups, such as marsupials, have presented many controversies in their location within the mammalian phylogenetic tree due to their rapid diversification (Duchêne et al., 2018), and some studies have proved the importance of reconstructing ancestral genomes in order to better resolve phylogeny controversies (Meredith et al., 2010). Among these controversies, the rate of evolution of mammalian species is still under debate. Some studies agree in describing a pulse of early radiation about 100 - 85 Mya, followed by low rates until about 10 Mya after the Cretaceous - Paleogene boundary, when diversification again sped up (Purvis et al., 2011). Other studies described no major diversification rate shifts for mammals or mammalian subgroups before 33 Mya (Stadler, 2011).

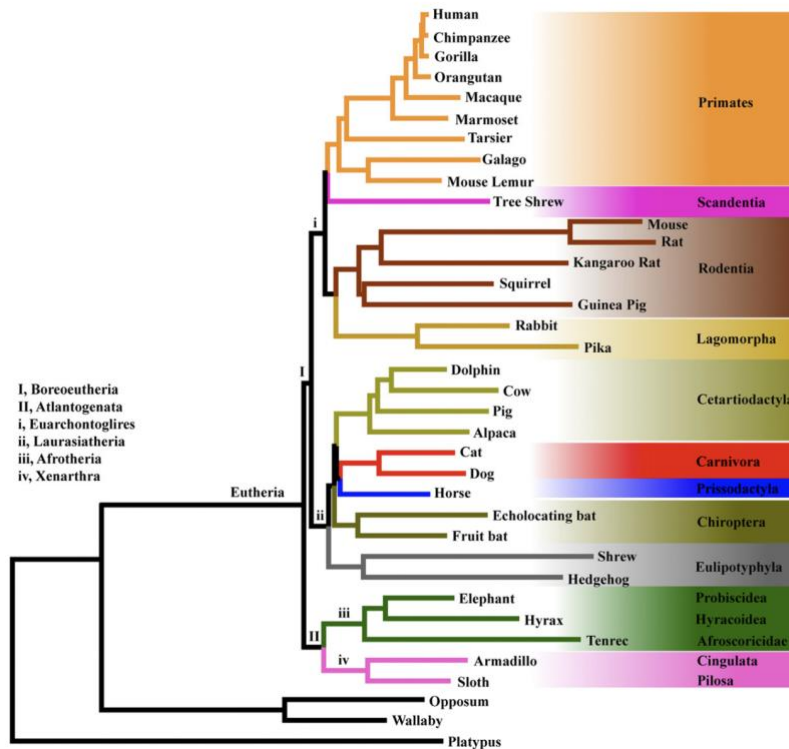


Figure 1.7. Evolutionary relationships of eutherian mammals. Figure adapted from (Song et al., 2012).

Just only recently, alignments of hundreds, even thousands, of genomes spanning the rich living and extinct diversity of species offer unparalleled power to resolve these phylogenetic controversies, identify genomic innovations of adaptation, and dissect the genetic architecture of reproductive isolation (Murphy et al., 2021).

Scientists have long struggled to untangle the deep phylogenetic relationships among the morphologically diverse ordinal clades of mammals, as well as the relationships within recent adaptive radiations. These questions remain not so much because of a lack of data but because of variation in genome quality and gene annotation, accurate ascertainment of orthologs, and other systematic and analytical errors that produce gene tree errors (Murphy et al., 2021). However, new phylogenomic approaches based on whole genome sequence alignments have demonstrated being more accurate to resolve phylogenetic relationships within the recently diverged species with histories of gene flow.

The pursuit of a resolved mammalian phylogeny at the species level is also motivated by a desire to understand how genomic changes have enabled mammals to occupy the most diverse habitats across the world. One of the main reasons that causes many controversies in the

mammalian phylogeny was to rely on morphological characters. Within mammals, morphological characters are often limited in their application because of pervasive ecomorphological convergence and correlated character evolution across deep spans of time. Indeed, mammals show striking examples of ecomorphological convergence evolution, where phenotypically similar species evolved independently in distantly related clades (Murphy et al., 2021).

However, not only studying the phylogenetic relationships among species is important to describe mammal evolution. Comparing the chromosomes of mammalian species, much can be learned not only about karyotype evolution, but also about the mechanisms involved and their significance for speciation and, indeed, about some of the genetic factors that distinguish closely related organisms (Ferguson-Smith & Trifonov, 2007).

1.5.2 Genome size

Studying genome size and its variation is key because it constitutes the link between analytical and molecular studies of the genome and the holistic and synthetic view of the genome (Redi & Capanna, 2012). Furthermore, it is very interesting that the genome size of several mammalian taxa shows a wide range, and it does not necessary correlate with the complexity of the organisms, although, genome size in mammals shows remarkably little interspecific variation compared with other taxa.

In mammals, genome size ranges from 1.6 to 6.3 Gb (Kapusta et al., 2017). One of groups that shows more differences in genome size are marsupials, with a variation of more than 1.2 Gb between species (Kasai et al., 2018). One of the hypotheses trying to explain these differences in genome variation point to the different levels of GC content among genomes. GC level content has been correlated with increasing levels of recombination in several species and it is also being reported that the total level of GC content is very variable between species (Kasai et al., 2018). And not only the relationship of genome GC content and higher recombination rates have been described, it also has been reported that chromosomal regions of high GC content are unusually susceptible to breakage and consequent rearrangement. For instance, in the human karyotype, GC poor chromosomes tend to be large and experience less recombination (Goodstadt et al., 2007).

Other factor that has been associated with differences in genome size is the genome content of repetitive sequences. It has been hypothesized that DNA gain and loss during the eutherian and avian evolution have played an important role in how their genomes present differences in size and composition. In fact, it has been proven that the dynamic interplay between TE-mediated DNA acquisition and subsequent DNA loss has played an important role in driving phenotypic evolution (Kapusta et al., 2017).

1.5.3 Chromosome number

Not only chromosome size could be very variable among mammalian species. Mammals show a wide range of chromosome numbers, ranging from $2n = 6/7$ in females and males individuals of Indian muntjac (*Muntiacus muntjak*) respectively, to $2n = 102$ in the red viscacha rat (*Tympanoctomys barrerae*) (**Figure 1.8**) (Ruiz-Herrera et al., 2012).

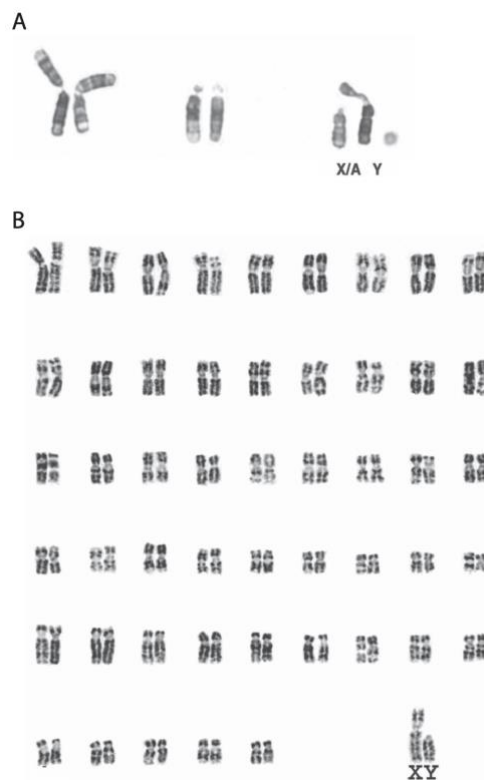


Figure 1.8. Karyotypes corresponding to a male individual of Indian muntjac (A) and a male individual of Red viscacha rat (B). Figures adapted from (O'Brien et al., 2005).

Chromosome number variation has traditionally been considered a proxy for the structural modification of karyotypes, especially so in groups of organisms where detailed information such as the differential staining of chromosomes, the extent and location of heterochromatin,

and the location and number of nucleolar organizers is lacking (Ruiz-Herrera et al., 2012). Subsequently, the biggest advances in determining chromosome numbers in deep branches of the vertebrate tree of life have resulted from the *in silico* scans of sequenced genomes.

In general, mammals are characterized by markedly different pathways of chromosome number evolution, and the variation among orders (**Figure 1.9**) reflects the complex dynamics of mammalian chromosomal evolution.

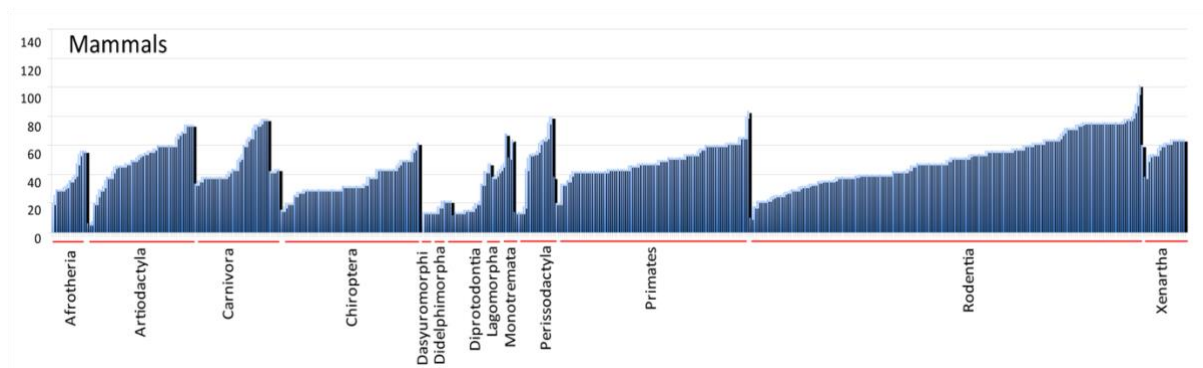


Figure 1.9. Chromosomal number variation among mammals. The x axis indicates the diploid chromosomal number (2n). The y axis groups species in different orders. From (Ruiz-Herrera et al., 2012).

These differences in karyotype number are mainly produced by CRs that alter the configuration of the chromosomes and have the potential to reduce or increase the number of chromosomes in the species. The identification of these rearrangements that have played an important role along mammalian evolution is key to understand the mechanisms leading adaptation and speciation. The detection and timing of these CRs have been the premise under several studies which attempted to reconstruct different mammalian ancestors (Kim et al., 2017; M. Farré et al., 2019; Deakin & O'Neill, 2020; Damas et al., 2022; Álvarez-González, Arias-Sardá, et al., 2022).

1.5.4 Chromosome structure and evolution

One important source of karyotype variability in mammals is related to heterochromatin. Some studies have proved that once the amount of heterochromatin is subtracted from the total genome content of all mammals, the species have very similar genome sizes (Graphodatsky et

al., 2011). It is well documented that heterochromatin may consist of different types of repetitive DNA, and it also can vary greatly between karyotypes of even closely related species.

However, despite the differences in chromosome number and size in the mammalian lineage, genome sequencing of an increasing number of organisms have revealed that the transcribed sequences of genomes of all species are highly conserved. This conservation has been seen even in large regions of chromosomes. Among mammals, large blocks of DNA (syntenic blocks) revealed by cytogenetic methods, often amounting to whole chromosomes or chromosome arms, can be shared by distantly related species. Although species might differ in chromosome number and chromosome morphology, the differences are due to these syntenic blocks being assembled in different combinations. Blocks that are fused together in one species can be separated on different chromosomes in another (Ferguson-Smith & Trifonov, 2007).

Furthermore, it has become clear that the interplay between the organisation of the genome and nuclear architecture is central to genome function. Some studies have remarked that genome reshuffling permissiveness is influenced by: (i) the physical interaction of genomic regions inside the nucleus, (ii) the accessibility of chromatin states, and (iii) the maintenance of essential genes and/or their association with long-range cis-regulatory elements (Deakin et al., 2019). Evolutionary studies have detected EBRs within regions enriched for repetitive sequences and genes, besides being located in regions with an open, actively transcribed chromatin.

A variety of methods have been used for resolving the evolutionary histories of mammalian chromosomes. For example, for closely related species with good map-anchored assemblies, such as human, chimpanzee, and rhesus, it was possible to determine the boreoeutherian ancestral karyotype by simple observational comparisons. This reconstruction infers a karyotype of 24 pairs of chromosomes and showed most of inversions, translocations, fusions, and fissions that occurred during evolution (Murphy et al., 2005). Posterior reconstructions using different approaches, have been extremely useful for the definition of models that explain genome structure and evolution. For instance, the eutherian ancestor was reconstructed using DESCHRAMBLER for the first time, and then curated comparing to previous FISH-based ancestral reconstructions using 19 eutherian species (Kim et al., 2017). This ancestor showed 32 reconstructed ancestral chromosome fragments (RACFs) and detected 162 EBRs that occurred during the 105 My of mammalian evolution, from the eutherian ancestor's genome to

the human genome. Posterior studies, have determined ancestral karyotypes of different mammalian lineages such as ruminants (M. Farré et al., 2019) and marsupials (Deakin et al., 2019), with diploid numbers of $2n = 48$ and $2n = 14$, respectively. Latest studies have used new approaches to determine ancestral karyotypes in different mammalian lineages using three different assemblies (human, cattle, and Southern two-toed sloth) as reference genomes (Damas et al., 2022). This study identified 323, 262, and 257 EBRs that occurred along the lineage from the mammalian ancestor to the human, sloth, and cattle genomes, respectively, showing a great variability of breakpoint rates during ~180 My of mammalian evolution.

In order to fully understand the mechanisms shaping mammalian genomes and driving speciation, it will be necessary to further study not only the functional constraints that would accompany genome reshuffling, but also the analysis of the structural organization of genomes.

1.5.5 Looking at the extremes: ruminants, marsupials and afrotherians

How genomes are organized and which types of CRs are implicated in speciation and macroevolutionary events are fundamental questions to understand the dynamics of chromosomal evolution. By studying the most extreme clades in terms of genome size and diploid numbers, I aim to shed some light in the mechanisms of chromosome evolution in mammals.

For this purpose, I selected three different mammalian lineages characterised by their broad range of chromosome sizes and number: the suborder Ruminantia, the infraclass Marsupialia, and the superorder Afrotheria, with a total of 234, 407, and 90 living species, respectively (last reviewed in September 2022).

Ruminants show a wide variation in diploid chromosome number, with karyotypes ranging from $2n = 6$ in females and $2n = 7$ in males of the Indian muntjac (*Muntiacus muntjak*) to $2n = 70$ in the white-tailed deer (*Odocoileus virginianus*). Marsupials, on the contrary, possess a more stable diploid number, with many of their species having either a $2n = 14$ (like Tasmanian devil, *Sarcophilus harrisii*) or a $2n = 22$ karyotypes (like many American marsupials belonging to the *Didelphis* genera). Afrotherians, similarly to ruminants, present more variability in terms of chromosome number, with diploid numbers ranging from $2n = 20$ in armadillo (*Oryzomys*) to $2n = 56$ in the African elephant (*Loxodonta africana*). These differences in

chromosome number are also reflected in the chromosome size of species. Species with small diploid numbers, such as the Indian muntjac, many marsupials, or aardvark, have extremely big chromosomes. On the other side, species with bigger number of chromosomes are related to smaller chromosome sizes. Their hypothesised ancestral diploid number is $2n = 48$, $2n = 14$ and $2n = 46$ for the ruminant (M. Farré et al., 2019), marsupial (Deakin & O'Neill, 2020) and afrotherian (Frönicke et al., 2003) ancestors, respectively. From these ancestral diploid numbers, many species had an extreme reduction in the number of chromosomes to reach their current chromosome configurations, leading to an increase in their chromosome sizes. The scenario that would better explain this reduction of diploid number is a succession of several fusions in their genomes.

This makes us wonder how these features evolved in such diverged/far groups of species and whether this reduction of chromosome number can be considered a convergent trait. Convergent evolution consists in the independent evolution of similar features in different evolutionary lineages (Losos, 2011). Whereas some scientists see in convergent evolution the unfettered ability of natural selection to produce optimal solutions to problems repeatedly posed by the environment, others see in it constraints on the evolutionary process resulting from biases in the production of phenotypic variation (Losos, 2011). Furthermore, convergent evolution can occur for reasons unrelated to adaptation and natural selection. For instance, random evolutionary changes can cause species to become more similar to each other than were their ancestors.

Because convergent evolution is widespread across the mammal tree of life, comparative morphology often proved misleading in determinations of mammal relationships, leading to arguments about relationships between major groups of mammals (Helgen, 2011). Nowadays, the study of signatures of convergence in genomes is possible due to the great availability of genomes already sequenced. Some studies that have reconstructed ancestral karyotypes across the mammalian tree and that have delimited CRs among species, also have identified orthologous regions defined by their correspondence with human chromosomes (Murphy et al., 2005; Ruiz-Herrera et al., 2012). And given an alignment of orthologous sequences, it can be tested whether the numbers of convergent or parallel substitutions along the branches in their phylogeny exceed the neutral expectation (Storz, 2016).

One of the most studied cases of convergent evolution is the case of marine mammals, such as manatees (within the order Sirenia), whales (within the infraorder Cetacea), or seals (within the superfamily Pinnipedia). Mammals have evolved to inhabit the marine environment on multiple independent occasions. Cetaceans and sirenians emerged during the Eocene epoch through diversification from the Cetartiodactyla and Afrotheria, respectively. Pinnipeds emerged approximately 20 My later during the Miocene from within the Carnivora. Despite their independent evolutionary origins, pinnipeds, sirenians, and cetaceans share a number of phenotypic adaptations to the pathogenic, locomotory, thermal, sensory, communication, and anaerobic challenges of an aquatic existence, including limbs adapted for swimming, bone density adapted to manage buoyancy, and a large total oxygen store relative to body size (Foote et al., 2015). Therefore, marine mammals are a classic example of convergent evolution, and many studies have identified genes evolving under positive selection in marine mammals by studying genes related to a marine phenotype.

But the adaptation to the marine world is not the only example of convergent evolution in mammals. Other examples of convergent evolution are the echolocating genes that can be found in diverse mammals (such as bats or marine mammals) (Parker et al., 2013), or the first genomic evidence of a trend to increased lifespan in mammals, probably driven by positive selection throughout the mammal clade (X. Farré et al., 2021).

Additionally, other studies have seen that among all the evolutionary alterations that could affect chromosome number and/or chromosome size (including reciprocal translocations, deletions, insertions, unequal crossover, dispersion of repetitive sequences, duplications, fusions, and fissions), reciprocal translocations constitute one of the major forces driving these changes (X. Li et al., 2011). This study has established a connection between features such as chromosome size, repetitive content, and genes that probably have evolved under positive selection (X. Li et al., 2011). Further studies in this topic are needed to potentially explain the role of these genes in chromatin condensation, conservation of repetitive elements and genome reshuffling.

1.5.5.1 Methods to detect convergent evolution

To study whether different lineages have undergone convergent evolution, and therefore can be considered convergent lineages, scientists have pointed to the identification of substitutions

that delimitate their convergent phenotypes. These substitutions can be adaptative, when they are fixed through positive selection, or non-adaptive, when they are fixed through a relaxation of selection. Nowadays, there are many methods searching for substitutions that are correlated with a phenotype, especially for convergent (or adaptative) substitutions. These methods, reviewed in (Rey et al., 2019), vary in how the similarity between substitutions is quantified, resulting in different sets of candidate substitutions.

Ideally, the best approach to study convergent evolution would need to consider the entire genome at the same time, along with “the fitness landscape,” to take into account all the dependencies between sites. And “the fitness landscape” would describe the mapping between genotypes and fitness in a species, for a given set of constraints on the phenotype (Khan et al., 2011). However, for computational reasons and due to the lack of experimental studies on fitness landscapes, such approach is out of reach. It is a known fact that adjacent nucleotide sites share some general parameters ruling their evolution (e.g. branch lengths, some parameters of the substitution matrices) but, conditionally on those parameters, each site is typically modelled independently of the others. The methodology used nowadays to detect convergent evolution works with the study of site-independent models that simplify assumptions, for instance, fitness landscapes only depend on the phenotype, and not on the lineage under consideration, as reviewed in (Rey et al., 2019).

The search for convergent traits in different species, such as the study of echolocating genes in both bats and marine mammals (Y. Li et al., 2010), has shown that positive selection repeatedly targets the same genes in several known cases of phenotypic convergence. The approach used in this study was the identification of convergent or parallel amino acid substitutions, that occurred between the branches of different species in a Maximum Likelihood (ML) phylogenetic tree, by comparing the inferred ancestral sequences in all interior nodes of the species tree and the extant sequences. The fact that positive selection repeatedly targeted the same genes leads to the prediction that these genes may show convergent shifts in evolutionary rates (i.e. number of nucleotide or amino acid substitutions per unit time), due to reduced or increased selective constraints on them (Kowalczyk et al., 2019). Another example is the study of both genes and non-coding elements involved in the loss of eye function as a selective constraint in subterranean mammals (Partha et al., 2017). This study showed that looking at shifts of evolutionary rates at a genome-wide scale, concomitantly with a well characterised convergent phenotype, can be a good approach to infer targets of convergent

selection. However, it has to be mentioned that their methods detect convergent changes in the rates of evolution of genes and hence are not designed to detect species-specific changes that might contribute to the subterranean adaptation.

In summary, the comparison of coding sequences across species is a good way to define convergent evolution, as they offer a window into where the mutation process and the selective process meet (Rey et al., 2019).

Overall, in this project, we will try to explain how the differences in chromosome number and size found in mammals can be a consequence of convergent evolution. To address this, we identified traces of positive selection on specific genes within the Ruminantia, Marsupialia and Afrotheria lineages. And together with the characterization of CRs, EBRs and HSBs, we will try to shed some light in the understanding of the role that genome reshuffling plays in adaptation and speciation.

1.6 Project aims

This thesis main goal is to study chromosome rearrangements (CRs) and their role during speciation in mammals. To achieve it, I focused on three main objectives:

1. To reconstruct eight ancestral genomes of the three selected mammalian lineages. For this purpose, I followed these steps:
 - (i) First, pairwise alignments of the selected species with the reference genomes of each lineage were performed. Cattle (*Bos taurus*), Tasmanian devil (*Sarcophilus harrisii*) and African elephant (*Loxodonta africana*) were selected as reference species of ruminants, marsupials and afrotherians, respectively, due to the quality of their genomes.
 - (ii) Then the reconstructions were performed using DESCHRAMBLER.
 - (iii) Finally, the reconstructed ancestral chromosome fragments (RACFs) obtained with DESCHRAMBLER, were curated by comparing with previous reconstructions and cytogenetic studies.

2. To detect evolutionary breakpoints (EBRs) and homologous synteny blocks (HSBs) in every lineage. For this purpose:
 - (i) EBRs and HSBs were detected using the chromosome coordinates obtained with the ancestral genome reconstructions.
 - (ii) These EBRs and HSBs were also classified in ancestor specific or species specific.
 - (iii) Gene ontology enrichment analysis (GOEA) analysis was performed to study the functions of the genes within or near EBRs, within conserved HSBs, and within rearranged HSBs.
 - (iv) Finally, statistical permutation tests were performed to find possible associations between EBRs and HSBs with TE-rich and gene-rich regions.

3. To identify signatures of convergent evolution in mammals with an extreme range of diploid numbers and chromosome size. For this purpose:
 - (i) Representative species of the three lineages were selected. All three lineages contained species divided in two groups. Species with high number of chromosomes which have lower chromosome size, and species with lower number of chromosomes which have higher chromosome size.
 - (ii) Orthologous genes were identified in all species.
 - (iii) Convergent evolution was assessed in the orthologous genes. Only those genes showing patterns of convergent evolution were selected for further analysis.
 - (iv) Finally positive selection was studied in the selected genes of species belonging to the group of lower number of chromosomes with higher chromosome size.



2. Materials and methods

2.1 Genomes used for the ancestral genome reconstructions

There are a total of 6,495 living mammalian species, and more specifically, there are 234 ruminant, 407 marsupial and 90 afrotherian species (Burgin et al., 2018; Mammal Diversity Database, 2022). Not all mammalian genomes have been sequenced and completely assembled and hence, there are currently 72 publicly available ruminant, 16 marsupial, and 12 afrotherian genome assemblies. Their assembly level is variable ranging from contigs, which are contiguous gapless stretches of DNA sequence assembled from smaller overlapping sequenced fragments; scaffolds, which are computationally ordered and oriented arrays of contigs that have sequence gaps along their length; super-scaffolds, when the scaffold order is determined by methods such as optical mapping and chromosome conformation capture technologies; c-scaffolds (chromosome-scale scaffolds), when a scaffold or super-scaffold appears to span all of a chromosome arm or a complete chromosome; and chromosomes, which are widely recognized as DNA-protein complexes housing a single linear DNA molecule that bears a linear array of genes (Lewin et al., 2019).

The use of high-quality genome assemblies is crucial to address questions related to the overall genome evolution and speciation. Specifically, chromosome-level reference genomes underpin the study of functional, comparative, and population genomics within and across species (Rhie et al., 2021). Because of the assembly quality of mammalian genomes is so variable, in this project we devised two main criteria to select the highest quality genomes for inclusion in the study: (i) span the maximum number of families or subfamilies in a given clade; (ii) high continuity of the assembly, such as genomes assembled at either chromosome-level or with a scaffold N50 > 1Mb (the scaffold N50, defined by the NCBI, is length such that scaffolds of this length or longer include half the bases of the assembly).

From all publicly available genomes in NCBI (National Center for Biotechnology Information) (Sayers et al., 2022) and DNAZOO (Dudchenko et al., 2017), a total of 47 genomes passed the quality filter (**Table 2.1**). Of these genomes, 22 belong to ruminants, five to marsupials, seven to afrotherian and 12 other mammalian and no mammalian species used as outgroups in the different ancestral genome reconstructions.

Table 2.1. Genomes used for the ancestral genome reconstructions. Size (Gb) refers to the total genome size. A.L. refers to the assembly level, and the species can be assembled in chromosomes (Chr), HiC-scaffolds (HiC-Sca), super-scaffolds (S-Sca), C-scaffolds (C-Sca) and scaffolds (Sca). The N50 refers to the scaffold N50.

	Species	Latin name	Family	Size (Gb)	A.L.	N50 (Mb)	Used in *
Ruminantia	Pronghorn	<i>Antilocapra americana</i>	Antilocapridae	2.96	Sca	18.84	1ab
	Cattle	<i>Bos taurus</i>	Bovidae	2.72	Chr	103.31	1ab,3
	Wild yak	<i>Bos mutus</i>	Bovidae	2.76	Sca	16.59	1a
	Water buffalo	<i>Bubalus bubalis</i>	Bovidae	2.66	Chr	117.22	1ab
	Goat	<i>Capra hircus</i>	Bovidae	2.92	Chr	102.34	1ab
	Blue wildebeest	<i>Connochaetes taurinus</i>	Bovidae	2.65	S-sca	98.1	1ab
	Sable antelope	<i>Hippotragus niger</i>	Bovidae	2.60	S-sca	100.24	1ab
	Gerenuk	<i>Litocranius walleri</i>	Bovidae	2.98	S-sca	100.17	1ab
	Gemsbok	<i>Oryx gazella</i>	Bovidae	2.74	C-sca	86.25	1ab
	Sheep	<i>Ovis aries</i>	Bovidae	2.87	Chr	107.69	1b
	Przewalski's gazelle	<i>Procapra przewalskii</i>	Bovidae	2.69	Sca	5.52	1a
	Hog deer	<i>Axis porcinus</i>	Cervidae	2.68	Chr	77.08	1ab
	Red deer	<i>Cervus elaphus</i>	Cervidae	3.44	Chr	107.36	1a
	Chinese water deer	<i>Hydropotes inermis</i>	Cervidae	2.53	Sca	13.82	1ab
	Chinese muntjac	<i>Muntiacus reevesi</i>	Cervidae	2.60	Chr	114.9	1ab
	White-tailed deer	<i>Odocoileus virginianus</i>	Cervidae	2.38	Chr	69.88	1a
	Reindeer	<i>Rangifer tarandus</i>	Cervidae	NA	Chr	NA	1b
	Giraffe	<i>Giraffa camelopardalis</i>	Giraffidae	2.44	Chr	177.94	1ab
	Okapi	<i>Okapia johnstoni</i>	Giraffidae	2.89	Chr	97.43	1ab
	Forest musk deer	<i>Moschus berezovskii</i>	Moschidae	2.73	Chr	102.12	1ab
	Siberian musk deer	<i>Moschus moschiferus</i>	Moschidae	2.11	Sca	11.73	1a
	Java mouse-deer	<i>Tragulus javanicus</i>	Tragulidae	2.59	Chr	14.08	1a
	Java mouse-deer	<i>Tragulus javanicus</i>	Tragulidae	NA	Chr	NA	1b
Marsupialia	Tasmanian Devil	<i>Sarcophilus harrisii</i>	Dasyuridae	3.09	Chr	611.35	2,4
	Gray short-tailed opossum	<i>Monodelphis domestica</i>	Didelphidae	3.57	HiC-sca	524.29	2
	Tammar wallaby	<i>Macropus eugenii</i>	Macropoidae	3.45	HiC-sca	387.07	2,4
	Red kangaroo	<i>Macropus rufus</i>	Macropoidae	3.60	HiC-sca	238.28	2
	Wombat	<i>Vombatus ursinus</i>	Vombatidae	3.34	HiC-sca	576.11	2
Afrotheria	Cape golden mole	<i>Chrysochloris asiatica</i>	Chrysochloridae	4.21	Sca	13.47	3
	African elephant	<i>Loxodonta africana</i>	Elephantidae	3.21	HiC-sca	120.06	3,4
	Asian elephant	<i>Elephas maximus</i>	Elephantidae	3.21	HiC-sca	95.96	3
	Aardvark	<i>Orycteropus afer</i>	Orycteropodidae	4.42	HiC-sca	644	3,4
	Cape rock hyrax	<i>Procavia capensis</i>	Procaviidae	3.61	HiC-sca	133.72	3
	Tenrec	<i>Echinops telfairi</i>	Tenrecidae	2.95	Sca	54.42	3
	Indian manatee	<i>Trichechus manatus</i>	Trichechidae	3.10	HiC-sca	143.72	3

Vertebrates	Camel	<i>Camelus dromedarius</i>	Camelidae	2.15	Chr	70.37	1ab
	Dog	<i>Canis lupus</i>	Canidae	2.34	Chr	64.2	1ab,4
	Macaque	<i>Macaca mulatta</i>	Cercopithecidae	2.97	Chr	82.35	1b,4
	Horse	<i>Equus caballus</i>	Equidae	2.51	Chr	87.23	1b
	Human	<i>Homo sapiens</i>	Hominidae	3.10	Chr	67.79	All
	Chimpazee	<i>Pan troglodytes</i>	Hominidae	3.05	Chr	53.1	1b
	Mouse	<i>Mus musculus</i>	Muridae	2.73	Chr	54.52	1ab,4
	Rat	<i>Rattus norvegicus</i>	Muridae	2.87	Chr	14.99	1b
	Platypus	<i>Ornitorhynchus anatinus</i>	Ornithorhynchidae	1.82	Chr	83.34	2,4
	Chicken	<i>Gallus gallus</i>	Phasianidae	1.07	Chr	20.79	2,4
	Sperm Whale	<i>Physeter catodon</i>	Physeteridae	2.51	Chr	122.18	1ab
	Pig	<i>Sus scrofa</i>	Suidae	2.50	Chr	88.23	1ab,3

1a - Aligned to cattle as reference, part of Ruminantia ancestor reconstruction (first version)

1b - Aligned to cattle as reference, part of Ruminantia ancestor reconstruction (second version)

2 - Aligned to Tasmanian devil as reference, part of Marsupialia ancestor reconstruction

3 - Aligned to African elephant as reference, part of Afrotheria ancestor reconstruction

4 - Aligned to human genome as reference

In ruminants, we selected representatives of all six ruminant families (Antilocapridae, Bovidae, Cervidae, Giraffidae, Moschidae, and Tragulidae), including: cattle, wild yak, water buffalo, gerenuk, Przewalskii gazelle, gemsbok, sable antelope, blue wildebeest, goat, sheep, reindeer, forest musk deer, Siberian musk deer, white-tailed deer, Chinese water deer, hog deer, red deer, Chinese muntjac, giraffe, okapi, pronghorn, and Java mouse deer. In marsupials, representatives of four out of the 21 marsupial families (Dasyuridae, Didelphidae, Macropodidae and Vombatidae) were included: Tasmanian devil, gray short-tailed opossum, red kangaroo, tammar wallaby, and common wombat. While in the Afrotheria lineage, representatives of six out of the nine afrotherian families (Elephantidae, Orycteropodidae, Chrysochloridae, Tenrecidae, Procaviidae and Trichechidae) were selected: Asian elephant, African elephant, armadillo, cape golden mole, lesser hedgehog tenrec, cape rock hyrax, and West Indian manatee. For the ancestral genome reconstructions, we used 12 species as outgroups: (i) 11 more mammalian species not belonging to any of the previously described lineages: platypus (Monotremata order), Linnaeus's two-toed sloth (Pilosa order), sperm whale (Artiodactyla order), dromedary camel (Artiodactyla order), horse (Perissodactyla order), wild pig (Artiodactyla order), dog (Carnivora order), house mouse (Rodentia order), rat (Rodentia order), chimpanzee (Primates order), and human (Primates order); (ii) chicken, as a representative of avian species.

2.2 Filtering and masking genomes

Even though the genomes we selected showed high quality assemblies, some unplaced or short scaffolds were still present. To avoid spurious alignments in downstream analysis, we filtered the genomes using faFilter tool (with the parameter *--minSize*) from UCSC (University of California, Santa Cruz) Kent Utilities (Kent et al., 2010; Kuhn et al., 2013). All unplaced scaffolds in genomes assembled at chromosome or c-scaffold level were removed using the size of the smallest chromosome in each species. Similarly, scaffolds smaller than 10 Kb in genomes assembled at super-scaffold level were removed. This filtering step was applied in all the species used in the ancestral genome reconstructions.

An additional curation of the genomes was performed to mask their repetitive sequences (such as transposable elements and satellites) and low-complexity DNA sequences. The genomes of all marsupials, afrotherians and outgroup species, as well as all ruminants downloaded from DNAZOO, were masked using RepeatMasker version 4.0.9 (Tempel, 2012). All ruminant species were masked using *-pa -64 -xsmall -species cattle*, while the rest of mammalian species were masked using *-pa -72 -xsmall -species mammal*. Only the chicken genome was downloaded directly masked (galGal6.fa.masked) from the UCSC Genome Browser webpage.

Both filtering and masking steps are crucial for the ancestral genome reconstructions, as they are key to increasing the resolution of the pairwise alignments and subsequently reduce the manual curation of the reconstructed ancestral genomes. Additionally, reducing the number of chromosomes or scaffolds and removing the repetitive elements of the genomes are important factors to reduce the time of analysis.

2.3 Whole-genome pairwise alignments

Once all the genomes were filtered and masked, the next step was performing the pairwise alignments using LASTZ version 1.04.00 (Harris, 2007). For every lineage whose ancestors we aimed to reconstruct, we selected a reference species that followed these criteria: (i) a high quality, chromosome-level genome assembly; (ii) previous literature describing their karyotypes and comparing them among other species (including painting probes and

fluorescence in situ hybridization (FISH) data); (iii) being a descendant species of the ancestor we aimed to reconstruct.

Considering all these requirements, the genomes selected as reference species to perform the pairwise alignments were cattle, Tasmanian devil, and African elephant for the Ruminantia, Marsupialia, and Afrotheria reconstructions, respectively. The species aligned to these reference genomes will be referred as a query species during the thesis. A fourth set of pairwise alignments was done using human as reference species to establish homologous syntenic blocks (HSBs) of human and study their conservation in the three lineages. Information of the alignments is shown in **Table 2.1**.

Before using LASTZ, all genomes were converted from their fasta format to a 2bit format by using faToTwoBit tool from UCSC Kent Utilities. Then, LASTZ was used with two different set of parameters.

The first set of parameters was used in the totality of the Ruminantia and Afrotheria lineages, in some of the alignments of the Marsupialia lineage, and in the alignments using human as reference genome. These parameters were LASTZ default parameters: -minScore = 1000, -linearGap = medium, C = 0, E = 30, H = 2000, K = 3000, L = 3000, O = 400 and Q matrix.

$$Q = \begin{bmatrix} & A & C & G & T \\ 91 & -114 & -31 & -123 \\ -114 & 100 & -125 & -31 \\ -31 & -125 & 100 & -114 \\ -123 & -31 & -114 & 91 \end{bmatrix}$$

For the Marsupialia lineage and the alignments using human as reference genome, the default parameters were used except for the alignments involving the sloth, platypus, and chicken genomes. In these specific cases, the parameters of LASTZ were changed into -minScore = 5000, -linearGap = loose, C = 0, E = 30, H = 2000, K = 2200, L = 6000, O = 400, M = 50, Y = 3400 and a new Q matrix.

$$Q = \begin{bmatrix} & A & C & G & T \\ 91 & -90 & -25 & -100 \\ -90 & 100 & -100 & -25 \\ -25 & -100 & 100 & -90 \\ -100 & -25 & -90 & 91 \end{bmatrix}$$

After running LASTZ, the obtained files in psl format were converted to chain and net formatted files. All tools used are part of the UCSC Kent Utilities and all of them were used with default parameters (Kuhn et al., 2013): (i) `axtChain -psl -verbose = 0 -minScore = 1000 -linearGap = medium`; (ii) `chainAntiRepeat`; (iii) `chainSort`; (iv) `chainPreNet`; (v) `chainNet`; and (vi) `netSyntenic`. In the Marsupialia lineage, due to the chromosome sizes of Tasmanian devil and to bypass the hardcoded limits of UCSC Kent tools, LASTZ alignment files in psl format which were bigger than 50 Gb were split into smaller files with a maximum file size of 25 Gb. After this split and to concatenate them into chain and net files, the same approach was used, only changing the parameters: `axtChain -psl -verbose = 0 -minScore = 2000 -linearGap = medium`.

HSBs were then constructed from chain and net alignment files using `maf2Synteny` (Kolmogorov et al., 2018) at three resolutions of detection (50 Kb, 100 Kb and 300 Kb). They were visualised using the function `draw.eh` from the R package `syntenyPlotter` (M. Farré et al., 2019). This function produces images in a format called Evolution Highway (EH) images. After comparing the EH images of all species at the three different resolutions, the 300 Kb resolution was chosen for further analysis because it achieved a balance between higher coverage of the reference genome and number of rearrangements detected. In all cases, Y chromosomes were omitted due to their low-quality assembly.

2.4 Ancestral karyotypes reconstruction

To reconstruct the ancestral karyotypes, we need the pairwise alignments and a phylogenetic tree. All the trees made for this study were written in Newick format and visualized using `FigTree` (Rambaut, 2009). `TimeTree` (Kumar et al., 2017) was used to establish the divergence time among all species. Each of the phylogenetic trees not only contained species of their lineages but also outgroup species to facilitate the ancestral genome reconstructions. The phylogenetic tree for the Ruminantia lineage was created with 20 ruminants and six outgroups;

the one for the Marsupialia lineage was created with five marsupials and five outgroups; and the tree for the Afrotheria lineage was created with seven afrotherians and three outgroups.

To reconstruct ancestral genomes we used DESCHRAMBLER (Kim et al., 2017). The method used by DESCHRAMBLER is an extension of the algorithm for reference-assisted chromosome assembly (RACA) (Kim et al., 2013) and uses syntenic fragments (SFs) produced by whole-genome comparisons of both high-quality chromosome scale and fragmented assemblies. The improvements regarding to RACA are: (i) implementation of a probabilistic framework to predict adjacencies of SFs in a target species; (ii) flexibility to handle both chromosome and scaffold level assemblies; and (iii) possibility to accommodate a large number of descendant species.

DESCHRAMBLER was run with a resolution of 300 Kb and a minimum adjacency score of 0.0001. The hardcoded constriction of 500,000,000 bp chromosome size was modified to 1 Gb in DESCHRAMBLER to allow for Tasmanian devil chromosome sizes. This modification was directly made in DESCHRAMBLER code folder, specifically in the file `species.h` (`#define MAXNUM 1000000000`). All species with a genome assembly in c-scaffold were labelled as chromosome level assembly. A total of five ancestors were reconstructed in the Ruminantia lineage, two in the Marsupialia lineage, and one in the Afrotheria lineage.

After running DESCHRAMBLER, the ancestral genomes obtained were organised in RACFs. Previous studies have shown that RACFs are more fragmented than what was expected, in part because of scaffold assemblies of descendant species where the exact tips of chromosomes are not known, and in part because of ambiguous cases resulting from insufficient evidence of adjacency (Kim et al., 2017). Due to this additional fragmentation of the ancestral genomes, we manually curated them to increase the level of accuracy of the reconstructions. To do so, we first reorganized and/or merged the fragments: (i) by referring to FISH-based reconstruction results in the case of the ruminant ancestors (M. Farré et al., 2019); (ii) by referring to FISH-based comparisons between the reference species and human in the case of the afrotherian ancestor (Fröncke et al., 2003); or (iii) by combining both of the approaches in the case of the marsupial ancestor (Deakin & O'Neill, 2020). Finally, the orientation of the merged RACFs in the ancestral chromosomes was established by minimizing the number of rearrangements.

To visualise the ancestral genomes and to continue with further analysis, plots of the RACFs were produced in three different ways using the functions *draw.eh*, *draw.ideogram*, and *draw.pairwise* of the R package *syntenyPlotter*.

2.5 Detection of Evolutionary Breakpoint Regions (EBRs) and Homologous Syntenic Blocks (HSBs)

After RACFs have been curated into the final version of the ancestral genomes, we compared the ancestors and the reference species to infer the evolutionary rearrangements between them. This process allows us to identify the different types of chromosome rearrangements that have occurred during their evolution such as fusions, fissions, inversions, and complex rearrangements. Not only the chromosome rearrangements are identified, but also EBRs and HSBs.

To define EBRs and HSBs, both the RACFs images and coordinate files were used. In the cases where there was more than one ancestor (ruminants and marsupials) when one EBR was shared by some of the ancestors but each of them presented a different size, the coordinates chosen to define that EBR were the ones corresponding to the minimum size. In **Figure 2.1** a EH image show this EBRs and HSBs classification.

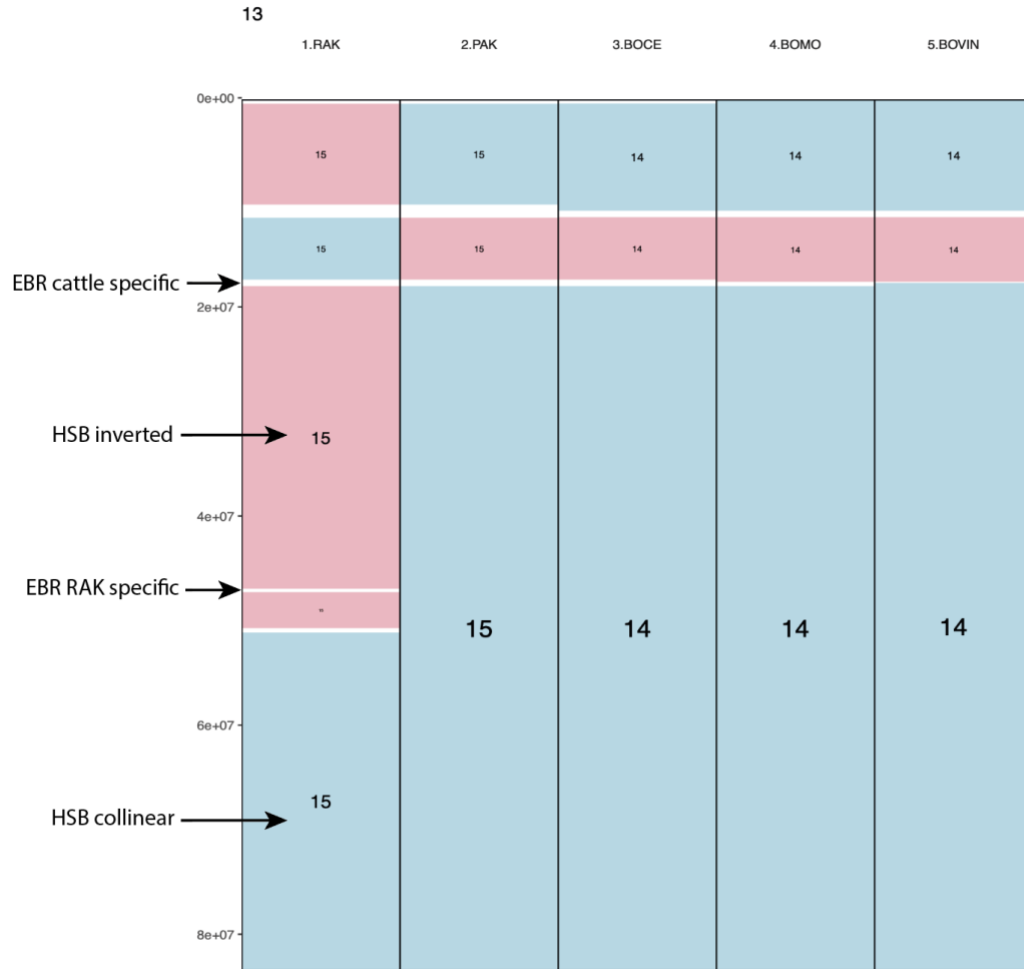


Figure 2.1. EH image of cattle chromosome 13 aligned with the five different ruminant ancestors: RAK, PAK, BOCE, BOMO, and BOVIN. The different colours indicate same (blue) or reverse (pink) orientation of the HSBs. Numbers within chromosomes of every ancestor indicate their chromosome number. When one EBR is shared by all ancestors is considered specific of the reference species, in this case, cattle, and the coordinates taken are the smaller ones, in this case the BOVIN ones. When a EBR is present in the oldest ancestor, in this case RAK, is classified as ancestor specific.

EBRs were classified by size into well-defined EBRs (ranging from 0 bp to 50 Kb) and not well-defined EBRs (ranging from 50 Kb to 300 Kb). In all cases where the EBRs size was higher than 300 Kb, those EBRs were considered alignment gaps and discarded from our analysis. The EBRs were then phylogenetically classified depending on the ancestral lineage in which they occurred into ancestor specific or reference species specific. Finally, EBRs were further separated by the type of rearrangement they delimited into inversion EBRs or inter-chromosome EBRs, if they demarcated inversions or were the result of fusion or fission events, respectively.

In the case of the HSBs classification, they were classified as collinear, inversion, fusion, or fission HSBs depending on whether they have been conserved during evolution (collinear HSBs) or whether they are the result of any type of chromosome rearrangement that change their position during evolution. In the specific cases of the ruminant and the afrotherian evolution, one HSB can be a product of two rearrangements events (**Figure 2.2**).

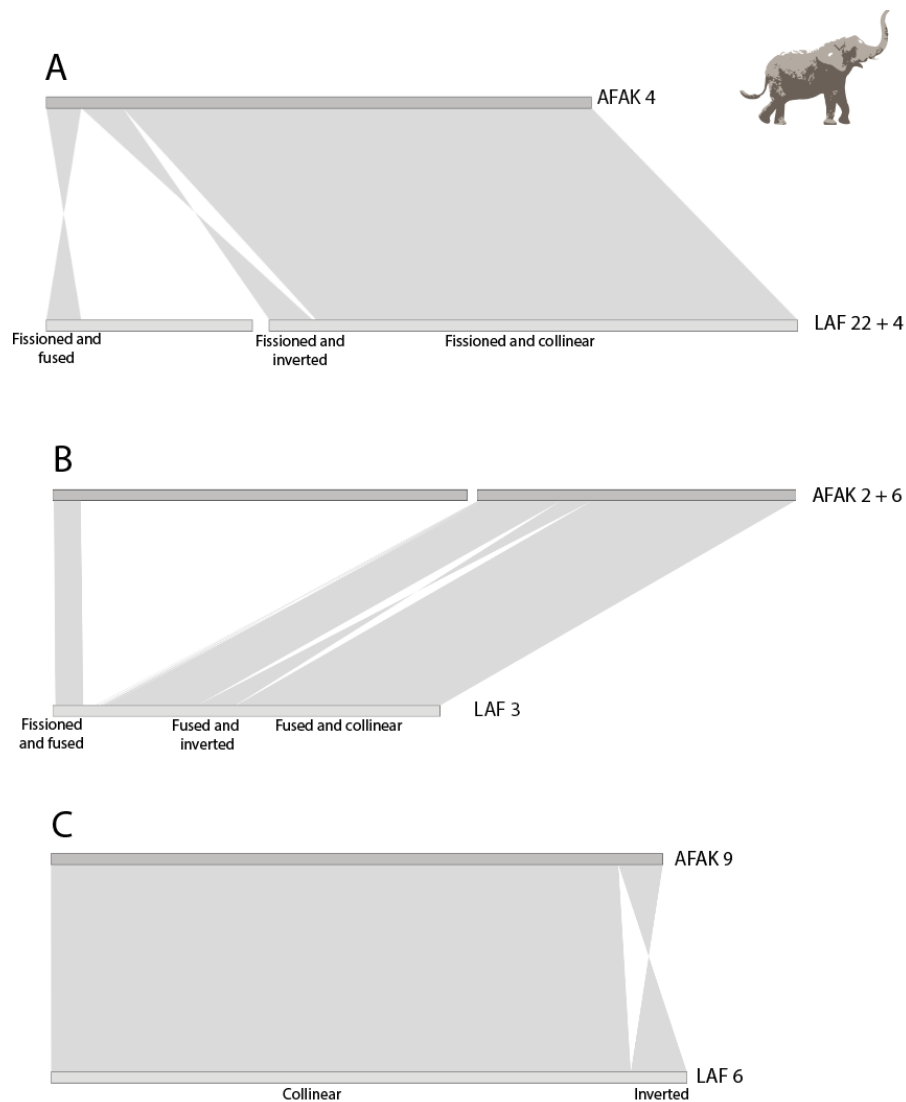


Figure 2.2. African elephant HSBs classification comparing to the afrotherian ancestor. A. Classification of HSBs of African elephant (LAF) chromosomes 4 and 22 after their evolution from the afrotherian ancestral karyotype (AFAK) chromosome 4. B. Classification of HSBs of LAF chromosome 3 after heir evolution from AFAK chromosomes 2 and 6. C. Classification of HSBs of LAF chromosome 6 after their evolution from AFAK chromosome 9.

2.6 African elephant genome annotation

After the EBRs and HSBs were identified, statistical analysis correlating these specific regions with gene content was performed, but in the case of the African elephant, a better annotation was needed. Previous studies have shown that genome annotations are directly influenced by the quality of genome assemblies, as genes might be missing or split between multiple scaffolds (Dessimoz et al., 2011), and that annotations based on RNA-seq data might miss lowly or very specifically expressed transcripts (Keilwagen et al., 2016). To avoid these issues, the method we selected uses a comparative approach between different species and does not need RNA-Seq data.

Gene Model Mapper (GeMoMa) (Keilwagen et al., 2016) was the software used for the annotation of the African elephant genome. Because GeMoMa is a homology-based gene prediction tool, two sets of genome annotations were needed as reference annotations. For this purpose, we selected the previous annotation of the African elephant (GCF_000001905.1_Loxafr3.0) and the current annotation of the small Madagascar hedgehog tenrec (GCF_000313985.2_ASM31398v2). Both genome annotations were downloaded from the NCBI webpage. These two species belong to two different families of the Afrotheria lineage: Elephantidae and Tenrecidae, and they diverged 78 Mya approximately.

GeMoMa (version 1.7.1) was run using default parameters without RNA-Seq data (Keilwagen et al., 2019) and following the next steps:

- (i) Starting Extractor: `java -jar $jar CLI Extractor a=${ref_annotation} g=${ref_genome} Ambiguity=AMBIGUOUS outdir=${out} r=true sefc=false d=false.`
- (ii) Starting tblastn: `makeblastdb -out ${out}/blastdb -hash_index -in ${target_genome} -title "target" -dbtype nucl; and tblastn -query ${out}/cds-parts.fasta -db ${out}/blastdb -eval 100.0 -out ${out}/search.txt -outfmt "6 std sallseqid score nident positive gaps ppos qframe sframe qseq sseq qlen slen salltitles" -db_gencode 1 -matrix BLOSUM62 -seg no -word_size 3 -comp_based_stats F -gapopen 11 -gapextend 1 -num_threads $threads.`
- (iii) Starting GeMoMa: `java -jar -Xms5g -Xmx50g $jar CLI GeMoMa s=${out}/search.txt c=${out}/cds-parts.fasta a=${out}/assignment.tabular t=${target_genome} sort=false Score="Trust" outdir=${out}.`

- (iv) Starting GAF: `java -jar $jar CLI GAF p=TEN g=${out1}/predicted_annotation.gff w=10 p=LOX g=${out2}/predicted_annotation.gff w=20 f='sumWeight>=20' outdir=${out}.`
- (v) Starting AnnotationFinalizer: `java -jar $jar CLI AnnotationFinalizer g=${target_genome} a=${out}/filtered_predictions.gff outdir=${out} rename=NO.`

The two independent annotations of the African elephant and tenrec were combined using the GAF command from GeMoMa (step iv) using different parameters. These parameters, called “weights” established the importance of every reference genome in the annotation. In our case, as we were annotating the African elephant genome, we gave more importance (or “weight”) to the previous African elephant annotation over the tenrec one. For the selection of the best set of parameters, gffCompare (version 0.12.6) (Pertea & Pertea, 2020) was used with the options `-chr-stats -r` to compare by pairs all the annotations generated with the different parameters in GAF. Finally, the selected annotation was the one with a weighting of 10 and 20 for tenrec (TEN) and elephant (LOX) genomes, respectively and a combined `sumWeight >=20`. This annotation had a total of 23,204 genes.

GffCompare can also be used to evaluate and compare the accuracy of transcript assemblers - in terms of their structural correctness (exon/intron coordinates). When gffCompare is run with the `-r` option (i.e. comparing with a reference annotation), tracking rows will contain a “class code” value showing the relationship between a query transcript and the closest reference transcript (where applicable). One of these “class code” is “=”, which is produced when there is a complete match of the new gene and the reference gene. All the genes tagged as “=” were maintained in the new annotated genome. For the rest of the different “class code”, we visualised all the genes using the Integrative Genomics Viewer (IGV) (J. T. Robinson et al., 2011; Thorvaldsdottir et al., 2013) and compared them with the annotation of both reference genomes. Only two more categories of “class code” were maintained: (i) “case j”: when a gene contained multiple exons and at least one of them matches to the reference gene; and (ii) “case k”: when there was a containment of reference (meaning that all the reference gene is contained in the new gene). The rest of genes, not matching these criteria, were removed from the final annotation. After this manual curation, a total of 20,641 genes were annotated in the final version of the African elephant genome annotation.

2.7 Multi-associations, statistical analysis and gene ontology enrichment analysis (GOEA)

Once we defined the EBRs and HSBs in the reference species of every lineage and we annotated the African elephant genome, we then downloaded the unique protein coding genes of cattle genome ARS-UCD1.2 and Tasmanian devil genome mSarHar1.11 using Ensembl BioMart (Smedley et al., 2015). To assess the relationship between EBRs, HSBs, and other genomic features, such as transposable elements or genes, we used *RegioneR* R package version 1.26. All the permutations, 1,000 per each dataset, were performed using *randomizeRegions* and *NumOverlaps* respectively as randomization and evaluation function. The value of the z-score for association with an adjusted P-value greater than 0.05, was considered as 0.

An additional analysis was performed to study the correlation of the EBRs regions and the TEs content of the reference species. For this analysis we used the Genomic Association Test (GAT) version 1.3.4 (Heger et al., 2013), which is a tool for computing the significance of overlap between multiple sets of genomic intervals. In our case, we compared the overlap between two different intervals (EBRs and TEs coordinates) within each reference genome, and GAT reported the fold enrichment as the ratio of observed and expected overlap and associated an empirical P-value with it. The parameters we used for running GAT were: *--ignore-segment-tracks* and *--num-samples=1000*.

Finally, we performed a gene ontology enrichment analysis (GOEA) with the EBRs and the HSBs. The gene annotations of cattle, Tasmanian devil, and African elephant were used, together with the coordinates of the EBRs and HSBs, and the basic version of the GO annotations (go-basic.obo), downloaded from the Gene Ontology Project website (Ashburner et al., 2000, The Gene Ontology Consortium et al., 2021) in November 2021. The GOEA was performed using a modified version of GO::TermFinder Perl module (Boyle et al., 2004). We considered significantly enriched only those terms with a corrected P-value ≤ 0.05 and a calculated false discovery rate (FDR) $\leq 1\%$. Plots were created with *ggplot2* (version 3.3.5) in R (version 4.1.0). The multiple hypothesis correction method used by GO::TermFinder Perl module is the Bonferroni correction. Additionally, to avoid the false negatives that can be originated by this type of correction, GO::TermFinder also calculates the FDR, which is the

expected proportion of true null hypotheses rejected out of the total number of null hypotheses rejected. GO::TermFinder calculates the FDR for each hypothesis from the real data as the average number of nodes per simulation that have a P-value as good or better than the real node's P-value, divided by the number of nodes in the real data that have a P-value as good or better than that P-value (Boyle et al., 2004).

2.8 Detection of orthologs

Additional analyses were performed to detect orthologous genes within the three lineages. For this purpose, representative species of each lineage were selected: gray short-tailed opossum, Tasmanian devil, brushtail possum, and wallaby as representatives of Marsupialia, cattle, white-tailed deer, red deer, Chinese water deer, black muntjac, Chinese muntjac, and Indian muntjac as representatives of Ruminantia, African elephant and aardvark as representatives of Afrotheria, and platypus as an outgroup within mammalian species. It was not a random selection, we tried to select species accomplishing one of the following requirements: (i) species with long chromosomes and low chromosome number; or (ii) species with small chromosomes and high chromosome number. All gene annotations (coding sequence – CDS) were downloaded from NCBI, ENSEMBL (Cunningham et al., 2022), and the Ruminant Genome Database (RGD) (L. Chen et al., 2019) (**Table 2.2**).

Table 2.2. Genomes used for the detection of orthologs. Size (Gb) refers to the total genome size. Average size refers to the average chromosome size. A.L. refers to the assembly level, and the species can be assembled in chromosomes (Chr) and scaffolds (Sca). The 2n refers to the diploid number, and the muntjacs have different diploid numbers depending on their sex (female/male). All gene annotations were the latest version found in NCBI, except for Black muntjac and Chinese water deer (found in RGD), and tammar wallaby (found in Ensembl). Colours indicate the lineage, green for Ruminantia, red for Marsupialia, blue for Afrotheria, and grey for other vertebrates.

	Species	Latin name	2n	Size (Gb)	Average size (Mb)	A.L.
R	Cattle	<i>Bos taurus</i>	60	2.72	87.6	Chr
	Red deer	<i>Cervus elaphus</i>	68	3.44	81.5	Chr
	Chinese water deer	<i>Hydropotes inermis</i>	70	2.53	70.5	Sca
	White-tailed deer	<i>Odocoileus virginianus</i>	70	2.38	66.8	Sca
	Chinese muntjac	<i>Muntiacus reevesi</i>	46	2.60	108.7	Chr
	Black muntjac	<i>Muntiacus crinifrons</i>	8/9	2.46	615.0	Sca
	Indian muntjac	<i>Muntiacus muntjak</i>	6/7	2.51	838.5	Chr

M	Tasmanian Devil	<i>Sarcophilus harrisii</i>	14	3.09	440.2	Chr
	Gray short-tailed opossum	<i>Monodelphis domestica</i>	18	3.57	389.1	Chr
	Tammar wallaby	<i>Macropus eugenii</i>	16	3.45	346.2	Sca
	Brushtail possum	<i>Trichosurus vulpecula</i>	20	3.33	332.7	Chr
A	African elephant	<i>Loxodonta africana</i>	56	3.21	112.0	Sca
	Aardvark	<i>Orycteropus afer</i>	20	4.42	425.5	Sca
V	Platypus	<i>Ornitorhynchus anatinus</i>	52	1.82	69.7	Chr

Then, OrthoFinder version 2.5.4 (Emms & Kelly, 2015, 2019) was run with the parameter *-f* in these species to find orthologous genes shared for all of them. All gene sequences appearing in the result folder “Single_Copy_Orthologue_Sequences”, all phylogenetic trees produced for each single-copy orthologous gene in the result folder “Resolved_Gene_Trees”, and the master phylogenetic tree produced in the folder “Species_Tree” were the files selected for further analysis.

2.9 Detection of convergence rates

Once all orthologs were identified, the next step was to search for convergent rates of evolution. For this purpose we used RERConverge (Chikina et al., 2016; Kowalczyk et al., 2019), which is a R package to identify genes whose evolutionary rates shift in association with a change in a trait.

Before using RERConverge, we re-formatted all files from OrthoFinder. The master tree branches' length was directly formatted following the instructions from RERConverge, and all branches belonging to ingroup species were modified to have a length of 1, while all branches belonging to outgroup species were modified to have a length of 0. For every set of single-copy orthologous genes, each of their amino acid sequences was aligned using Muscle (Edgar, 2004) with the parameters *muscle -align*. And finally, the trees of each single-copy orthologous genes were modified using the R package Phangorn, which estimates Maximum Likelihood (ML) trees from a given topology for a list of alignments, and writes a text file compatible with the function *readTrees()* used by RERConverge. For this purpose, we ran separately the script “estimateTreeFuncs.R” found within the folder “R” of RERConverge, and we modified the parameters *type = “AA”, format = “clustal”,* and we eliminated the parameter *writealn=TRUE*.

Once we obtained all the files, RERConverge was run to estimate the relative evolutionary rates (or RERs) of every gene according to a specific classification of the species. Accordingly, species were classified into two different groups: (i) all species with long chromosomes (genome sizes higher than 300 Mb) and low diploid number ($2n \leq 20$), including all marsupials, black and Indian muntjacs, and aardvark; and (ii) all species with small chromosomes (genome sizes lower than 120 Mb) and high diploid number ($2n > 40$), including platypus, cattle, white-tailed deer, Chinese water deer, red deer, Chinese muntjac, and African elephant.

The functions used for running RERConverge were the following:

- readTrees(), without any specification in the parameter *max.read*.
- getAllResiduals(), with the parameters: *transform = "sqrt"*, *weighted = T*, *scale = T*.
- readRDS().
- plotTreeHighlightBranches().
- foreground2Tree() with the parameters: *clade="ancestral"* for ingroup species, *clade="terminal"* for outgroup species, or *clade="all"*, *weighted=TRUE* for all species.
- foreground2Paths() with the parameter *clade="all"*.
- tree2Paths().
- correlateWithBinaryPhenotype(), with the parameter *weighed="auto"*.
- plotRers().
- returnRersAsTree(), with the parameter *plot=TRUE*.
- returnRersAsTreesAll().
- treePlotRers(), with the parameter *type= "c"*.

Using the function “correlateWithBinaryPhenotype()” we could associate variation in these RERs with variation in a binary trait across the tree.

2.10 Detection of positive selection

Only genes showing convergence in Group 1 (all species with long chromosomes and low diploid number) were further studied. To identify the most informative genes, we first studied their interactions using the online tool GeneMANIA (Warde-Farley et al., 2010). This software was run using human as reference species and the networks selected to study were: InterPro,

Physical Interactions, Co-expression, Genetic interactions, miRNA-target-predictions, Transcriptional-factor-targets, Reactome, Predicted, Pathway, Co-localization, and Shared protein domains. All genes associated by these networks were further studied, and three networks related with DNA repair and chromatin organization were selected for further analysis (**Figure 2.3**).

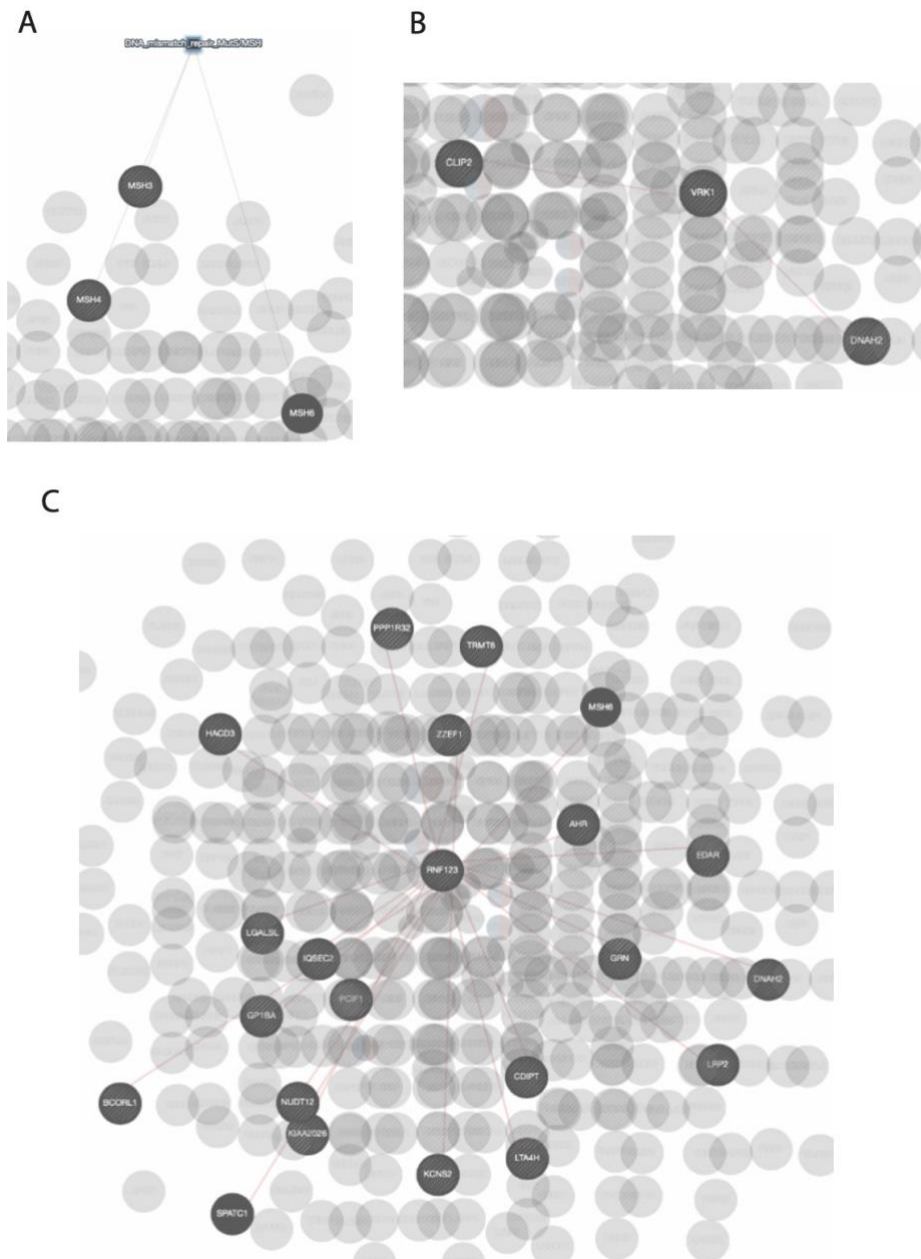


Figure 2.3. The three selected networks from GeneMANIA analysis: DNA mismatch repair (A), physical interaction (B) and another physical interaction (C). The light grey circles indicate all genes analysed in GENEMANIA. Black circles with gene names indicate genes interacting in the three selected networks. Red lines connect genes interacting in the three selected networks.

Then, genes belonging to the informative networks (**Figure 2.3**) were used to detect positive selection. For this purpose, we used CodeML implemented within EasyCodeML (Gao et al., 2019), a software that allows to detect genomic signatures of positive selection and evolutionary constraints in their coding sequences. EasyCodeML was run using the Branch Site Model setup from Nested Models (default configuration). This model allowed us to detect the genes and specific amino acids under positive selection in each phylogenetic branch studied. We ran EasyCodeML for each branch under investigation (**Figure 2.4**), including species of Group 1 and their ancestral branches.

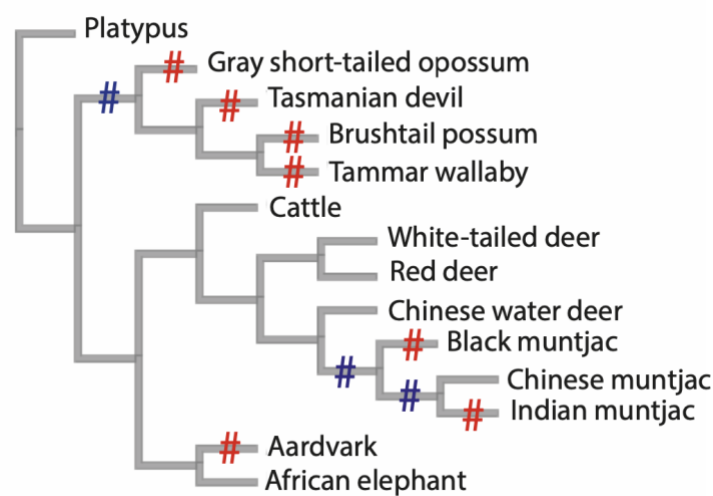


Figure 2.4. Phylogenetic tree used as input tree in EasyCodeML. Specific branches for the species of group 1 selected for the analysis appear with a red #. Ancestral branches selected for the analysis appear with a blue #.



3. The Ruminantia lineage

3.1 Introduction

Ruminantia, the largest suborder within the Cetartiodactyla order, comprises 234 species, classified in 76 genera, within six families (Antilocapridae, Bovidae, Cervidae, Giraffidae, Moschidae, and Tragulidae) and divided in two infraorders, Tragulina, with only three genera, and Pecora, including all the other 73 genera (Burgin et al., 2018; Mammal Diversity Database, 2022).

Molecular and fossil evidence suggests that Ruminantia evolution has been far from constant, and also proves that some of the major speciation and lineage turnover events have occurred within relatively short periods (Zachos et al., 2001). The two infraorders, Tragulina and Pecora, split around 50 Mya, a period of time concurring within the Eocene climatic optimum, while the pecoran radiation (around 33.2 Mya) coincided broadly with a strong glacial event at the onset of the Oligocene and was ongoing until approximately 28.1 Mya (Zachos et al., 2001; Fernández & Vrba, 2005). Despite of this rapid radiation of Pecora, subfamilies within Cervidae and Bovidae began to differentiate in the early Miocene (between 25.4 Mya and 13.5 Mya). Overall, ruminant evolution coincided with periods of climatic and vegetation change all around the globe. Although ruminants lack a fully resolved phylogeny, current studies (L. Chen et al., 2019) have proposed a more accurate timing of the ruminant radiations within the Geological Time Scale (GTM) of Earth, by studying the phylogenetic relationship of particular groups, such as the Antilocapridae and Moschidae families (**Figure 3.1**).

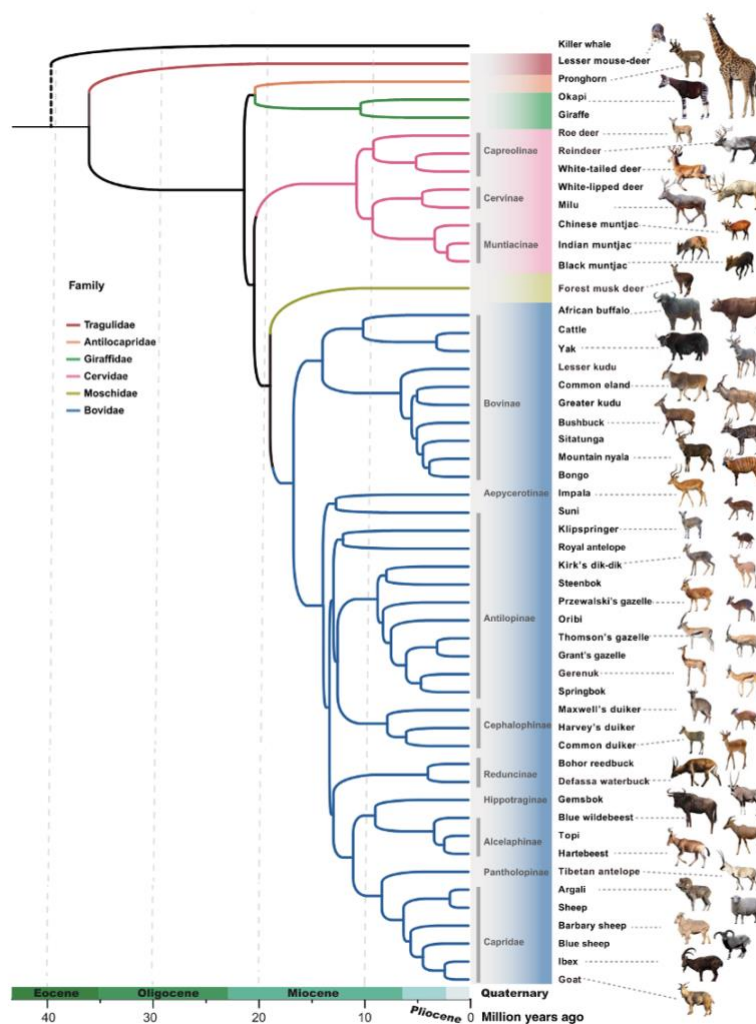


Figure 3.1. Phylogeny of ruminants among the Geological Time Scale (GTM) of Earth. Figure adapted from (L. Chen et al., 2019).

Ruminants have played an important role in human civilization, containing several key livestock species such as cattle (*Bos taurus*), yak (*Bos mutus*), sheep (*Ovis aries*) and goat (*Capra hircus*), widely farmed across the globe (L. Chen et al., 2019). Among the six Ruminantia families, Bovidae is the most species-rich family with a high number of livestock species. Ruminants are among the most successful herbivores, however, around 28% and 16% of ruminant species are considered endangered or critically endangered and near threatened, respectively, by the International Union for Conservation of Nature (IUCN) Red List (IUCN, 2022). Additionally, more than 66% of ruminant species show decreasing numbers in populations (**Table S1**).

Moreover, ruminants show a wide range of extreme morphological and ecological characteristics:

- (i) Pecoran ruminants have a four chambered stomach composed of the rumen, reticulum, omasum, and abomasum, while ruminants from the Traguliadae family lack the omasum, reviewed in (L. Chen et al., 2019). The appearance of the rumen and its close interaction with microorganisms allows ruminants to gain energy efficiently, thus providing unique evolutionary advantages and promoting the prosperity and diversity of ruminant groups (B. Wang et al., 2019).
- (ii) Of the six extant families of ruminants, four possess cranial appendages (denominated headgear): cervids have antlers, bovids have horns, giraffids have ossicones, and the antilocaprid has pronghorns (Davis et al., 2011). Moreover, the most remarkable characteristic of these different cranial appendages is the rapid regeneration ability of cervid antlers, being the only completely regenerable organ found in mammals with an extremely rapid growth rates (~1.7 cm/day in red deer), with rates of cell proliferation that surpass even cancerous tissue growth (Y. Wang et al., 2019).
- (iii) Ruminants also have specialized dentition patterns, lacking upper incisors and having a high prevalence of high-crowned or hypsodont teeth, reviewed in (L. Chen et al., 2019). It is believed that the distinct dentition of ruminants have contributed to their evolutionary success of colonisation very diverse habitats by expanding their food sources (Cantalapiedra et al., 2014).
- (iv) Adaptations to a highly cursorial locomotion have allowed many lineages to open habitats by adopting more cursorial body plans designed for rapid and/or sustained movement in an open habitat, reviewed in (L. Chen et al., 2019)
- (v) Ruminants possess a wide body size and weight ranges that span from 2 kg (found in the lesser mouse deer, *Tragulus kanchil*) to 1,200 kg (found in the giraffe, *Giraffa camelopardalis* and the African buffalo, *Syncerus caffer*) (B. Wang et al., 2019). Changes in body size are postulated as a mechanism for differentiating the feeding requirements of herbivores as large herbivores can extract more energy from plant material than smaller herbivores (Demment & Van Soest, 1985).

Concomitantly with this morphological and ecological diversity, ruminants are distributed across extensive habitats, including different altitudes, latitudes, and ecological environments (B. Wang et al., 2019). Different species from the same family can be found indistinctly in tropical and Arctic regions, such as hog deer (*Axis porcinus*) and reindeer (*Rangifer tarandus*), respectively, both belonging to the Cervidae family.

More interestingly, ruminant species also show a wide variation in diploid chromosome number, with karyotypes ranging from $2n = 6$ in females and $2n = 7$ in males of the Indian muntjac (*Muntiacus muntjak*) to $2n = 70$ in the white-tailed deer (*Odocoileus virginianus*). Nevertheless, most of the species, especially in the Bovidae family, have karyotypes with a diploid chromosome number of $2n = 60$ chromosomes (**Figure 3.2**).



Figure 3.2. Diploid chromosome number in ruminants. A. Representation of diploid chromosome number in all ruminant species individually. Red line indicates the most repeated diploid number ($2n = 60$). B. Range of diploid chromosome number in all ruminant species classified in subfamilies.

As a result of all this variability within the Ruminantia lineage, several studies on morphology and cellular biology have been done and yet, tracing their phylogeny and establishing the chromosome rearrangements leading to their evolution is still ongoing.

First studies used molecular cytogenetic methods based on comparisons of chromosome banding in close-related species to determine the evolution of their karyotypes (Slate et al., 2002). Posterior studies used genetic mapping comparisons (Murphy et al., 2005) and FISH probes to improve the resolution of these close-related species comparisons (Kulemzina et al., 2009, 2011). Such information is key to understand their karyotypic phylogeny and to establish a reconstructed karyotype of their common ancestor although the methodology used produced results in a low resolution. Nowadays, thanks to the improvement of the genome assembly techniques, the great number of genomes already sequenced and the development of software tools which can work with many species at the same time, the establishment of phylogenetic relationships, and the prediction of ancestral karyotypes have increased their resolution and reliability. Some studies have worked on the reconstruction of the ruminant evolution in a more accurate way, by using DESCHRAMBLER and constructing syntenic fragments from whole-genome comparisons of both high-quality chromosome scale and fragmented assemblies, and posteriorly resolving any possible discrepancy using FISH data. These studies showed that ruminant karyotypes are populated with multiple inter- and intrachromosomal rearrangements enriched for lineage-specific transposable elements (TEs), in contrast to other mammalian groups, such as primates, that had a higher percentage of intrachromosomal rearrangements in their recent evolutionary history (Kim et al., 2017; M. Farré et al., 2019).

To assess the functional contribution of chromosomal rearrangements to the evolution of ruminants, we reconstructed five ruminant ancestral genomes. We used two different sets of species to compare how the reconstructions could be affected by a different selection of species. For the reconstructions we used 22 ruminant genomes: cattle, wild yak, water buffalo, gerenuk, Przewalskii gazelle, gemsbok, sable antelope, blue wildebeest, sheep, goat, forest musk deer, Siberian musk deer, white-tailed deer, Chinese water deer, hog deer, red deer, reindeer, Chinese muntjac, giraffe, okapi, pronghorn, and Java mouse deer (see in **Methods Table 2.1**). The ruminant ancestral karyotypes reconstructed are: RAK, the most primitive ruminant ancestral karyotype; PAK, the pecoran ancestral karyotype; BOCE, the ancestral karyotype that included Cervidae, Moschidae, and Bovidae families; BOMO, the ancestral karyotype that included Moschidae and Bovidae families; and BOVIN, the ancestral karyotype of the Bovidae family.

3.2 Results

The aims of this chapter are: (i) to reconstruct the ancestral karyotype of all ruminant species; (ii) to identify the chromosome rearrangements (CRs) that have played a role in the evolution from this ancestral karyotype to current ruminant species; (iii) to identify the evolutionary breakpoints (EBRs) and the homologous synteny blocks (HSBs) (conserved and/or rearranged) in cattle genome; (iv) to perform a gene ontology enrichment analysis (GOEA) of the genes near and within EBRs; (v) to perform a GOEA of the genes within rearranged and/or conserved HSBs; and (vi) to study the presence of TEs near and within EBRs.

3.2.1 Pairwise alignments

The first step was to perform pairwise alignments of the 21 ruminant species and the six non-ruminant (outgroup) species against cattle. We selected cattle as reference genome not only because of the high-quality of its assembly but also because of the great amount of information about its genome, such as gene annotation, TEs annotation, and many comparative studies with other species like human.

Table 3.1. Coverage of net files and HSBs in the alignments of ruminant species against cattle. % nets refers to the coverage of all sequences of non-overlapping gapless blocks. % HSBs refers to the coverage of defined HSBs.

Species	Latin name	Size (Gb)	* % nets	* % HSBs
Pronghorn	<i>Antilocapra americana</i>	2.96	99.71	94.63
Wild yak	<i>Bos mutus</i>	2.76	99.71	94.44
Water buffalo	<i>Bubalus bubalis</i>	2.66	99.66	91.84
Goat	<i>Capra hircus</i>	2.92	99.62	93.88
Blue wildebeest	<i>Connochaetes taurinus</i>	2.65	99.85	95.91
Sable antelope	<i>Hippotragus niger</i>	2.6	99.84	95.99
Gerenuk	<i>Litocranius walleri</i>	2.98	99.29	94.99
Gemsbok	<i>Oryx gazella</i>	2.74	99.84	84.57
Przewalski's gazelle	<i>Procapra przewalskii</i>	2.69	98.47	89.3
Hog deer	<i>Axis porcinus</i>	2.68	99.83	95.87
Red deer	<i>Cervus elaphus</i>	3.44	99.74	95.24
Chinese water deer	<i>Hydropotes inermis</i>	2.53	99.59	94.25
Chinese muntjac	<i>Muntiacus reevesi</i>	2.6	99.72	95.74
White-tailed deer	<i>Odocoileus virginianus</i>	2.38	99.82	95.67

Giraffe	<i>Giraffa camelopardalis</i>	2.44	99.75	95.81
Okapi	<i>Okapia johnstoni</i>	2.89	99.79	95.59
Forest musk deer	<i>Moschus berezovskii</i>	2.73	99.84	95.84
Siberian musk deer	<i>Moschus moschiferus</i>	2.11	99.83	90.87
Java mouse-deer	<i>Tragulus javanicus</i>	2.59	98.96	92.95

*Coverage against cattle genome (size=2.71Gb)

Thanks to the good quality of the assemblies used in these alignments (see **Methods Table 2.1**), the coverage reached in the alignments of the ruminant species against cattle genome was very high (**Table 3.1**). The net alignment coverage ranges from 98% to 99% in all species. Moreover, the coverage of the HSBs produced is also high, with values ranging from 84.57% in gemsbok to 95.99% in sable antelope and an average coverage of 93.86%. Overall, these data indicate that the ancestral reconstruction using DESCHRAMBLER will have high accuracy levels.

These alignments were visualised through Evolution Highway (EH) images (**Figure 3.3**) to detect CRs specific of cattle. Many chromosomes have been conserved during evolution (**Figure 3.3A**), while others have been affected by CRs (**Figure 3.3B**). When one chromosome has been maintained during evolution, all species are seen as one only HSB, except for species with scaffold-level assemblies (**Figure 3.3A**). On the contrary, when a chromosome has been rearranged during evolution it will be reflected in the EH images, and if all species present the same rearrangement pattern it will mean that this CR is specific from the reference species (**Figure 3.3B**).

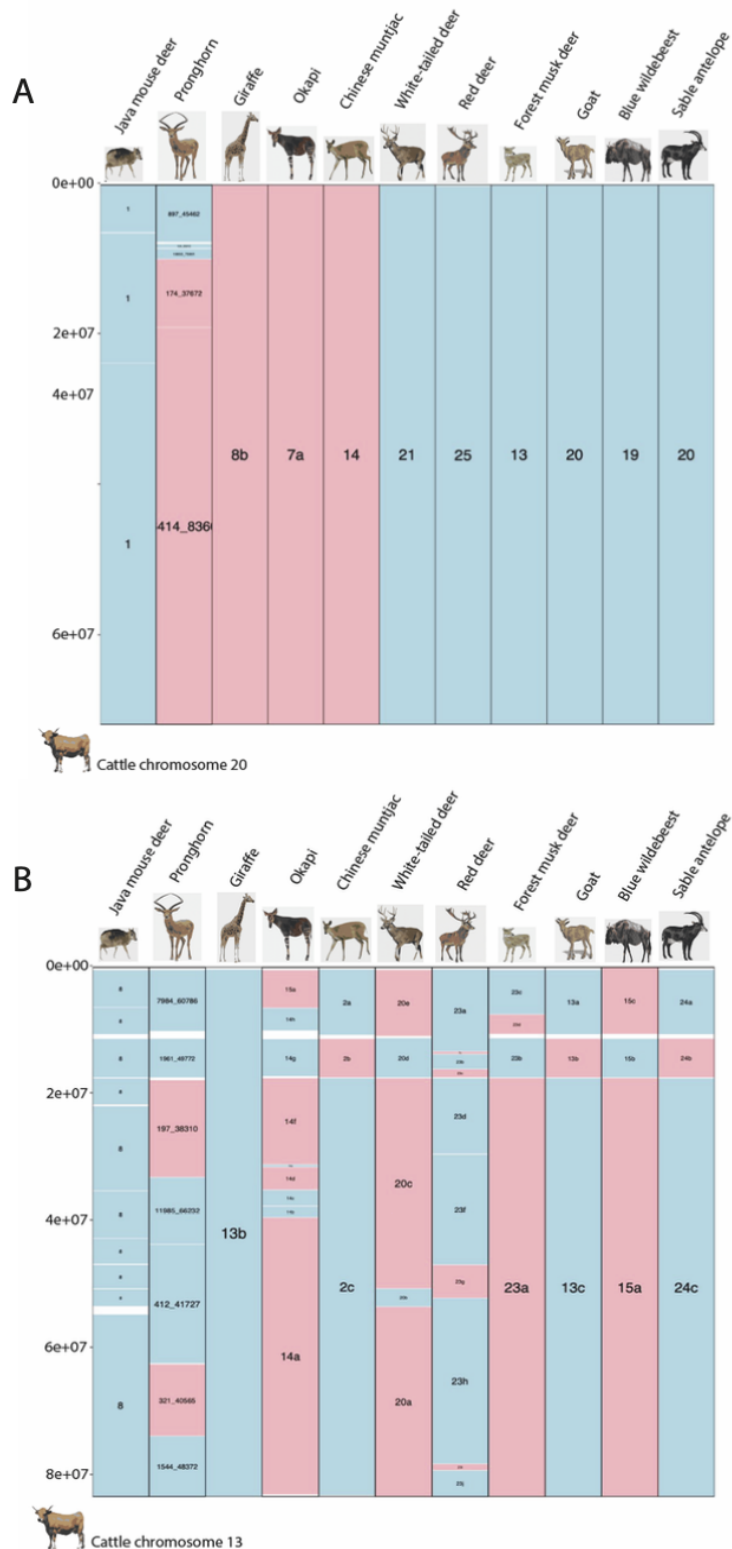


Figure 3.3. EH image of chromosomes 20 (A) and 13 (B) of cattle aligned with 11 of the ruminant species. Blue and pink colours indicate same or reverse orientation of the query chromosomes or fragments, respectively. Numbers within each column indicate either chromosome or scaffold IDs of the 11 species. The 11 query species and cattle are representatives of the six ruminant families: Tragulidae (Java mouse deer), Antilocapridae (pronghorn), Giraffidae (giraffe and okapi), Cervidae (Chinese muntjac, white-tailed deer, and red deer), Moschidae (forest musk deer) and Bovidae (goat, blue wildebeest, sable antelope, and cattle).

Overall, looking to the EH images chromosome by chromosome, we can detect cattle specific rearrangements, species- or lineage-specific rearrangements, and conserved HSBs among species. However, the best approach to study these CRs that have occurred during ruminant evolution, is to reconstruct ancestral genomes and compare them with current ruminant species.

3.2.2 Ancestral karyotype reconstructions

Thanks to the availability of several ruminant genomes, we were able to reconstruct five different ancestral karyotypes: RAK, PAK, BOCE, BOMO, and BOVIN (Figure 3.4).

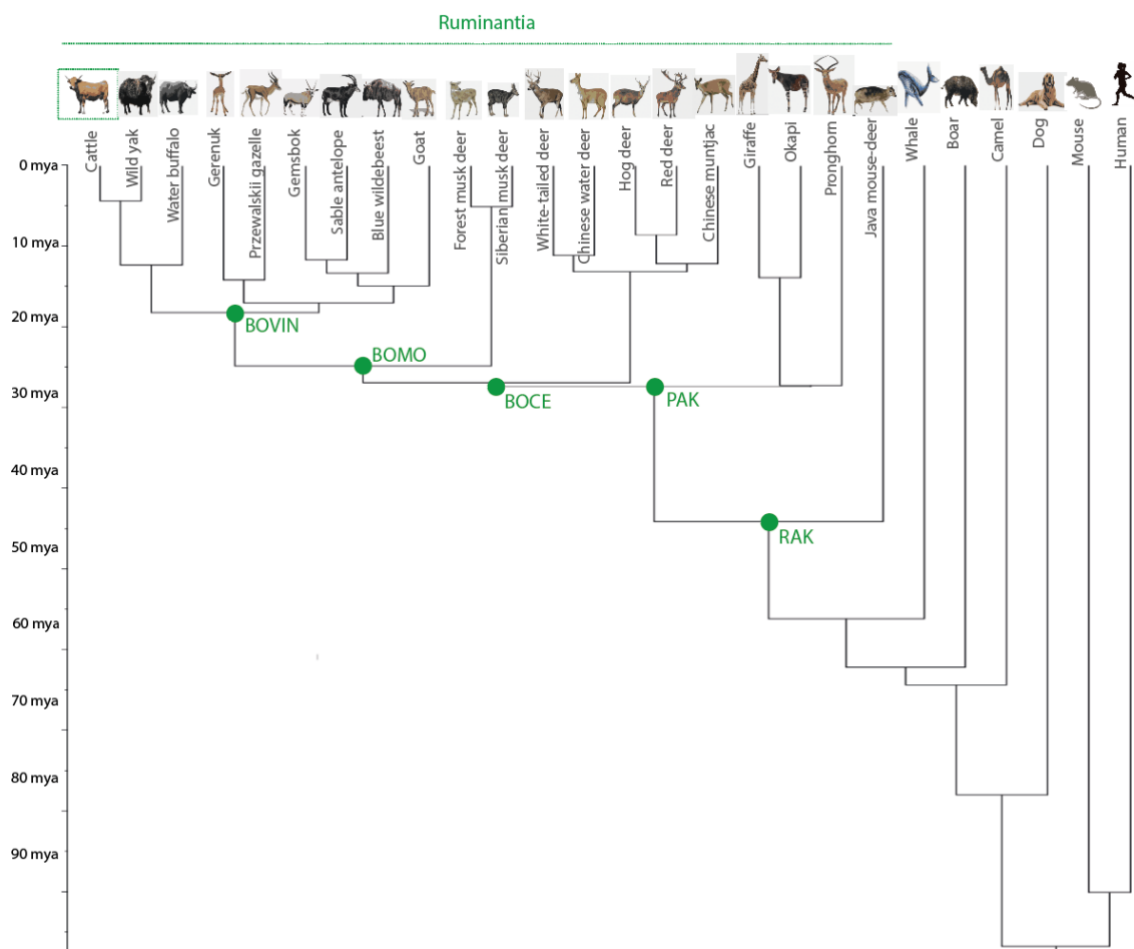


Figure 3.4. Phylogenetic tree of all species included in the ruminant ancestral reconstruction. In green the ancestral nodes reconstructed.

We used DESCHRAMBLER to reconstruct the five ancestors with a resolution for the alignment blocks of 300 Kb and an adjacency score of 0.01. For all ancestors we obtained high genome coverages comparing to cattle genome, with values ranging from 88.6% in RAK to 93.83% in BOVIN. Those ancestors more closely related to cattle showed higher coverage

values. And when looking at the initial RAFCs processed by DESCHRAMBLER, RAK showed the highest number, 40, followed by PAK with 32 and, BOCE, BOMO and BOVIN with 30 each. Comparing these numbers with previous reported chromosome numbers for the same ancestors (M. Farré et al., 2019), and when looking at the smaller reconstructed ancestral chromosome fragments (RAFCs) sizes obtained in some ancestors, it was noticeable the necessity to perform a curation of this data.

We then manually curated the RACFs to increase coverage and minimise overestimation of CRs. First, we compared RAK and PAK with previously published ancestral reconstructions (M. Farré et al., 2019). RACFs were then manually merged comparing different ancestors and alignments using the most parsimonious scenario (see **Methods section 2.4**). In general, small RACFs correspond to unplaced fragments due to the fragmented nature of some of the genome assemblies. For instance, if a species assembled at scaffold-level is key for the node whose ancestor is going to be reconstructed, DESCHRAMBLER is unable to find the exact breakpoint shared by other species and produce an extra RACF. These unplaced RACFs were then manually merged when the query coordinates belong to the same chromosome and are consecutive one from another. Finally, in the cases where one of the RACFs was in a reverse orientation, coordinates and the orientation were reversed before proceeding with its merging with another chromosome.

After the curation, the statistics of the reconstructed ancestors improved (**Table 3.2**). DESCHRAMBLER results showed quite similar chromosome numbers than previous reconstructions (M. Farré et al., 2019). The number of RACFs was reduced to 27 in RAK and 30 in the rest of ancestors, with RAK and PAK as the only ancestors which needed curation. On one hand, RAK was reduced in 13 RACFs and increased its minimum RACFs size to 1,348 Kb, while on the other hand, PAK was reduced in two RACFs and neither its maximum nor minimum RACFs sizes were modified.

Table 3.2. DESCHRAMBLER results. “FISH number” is the haploid number of chromosome reported by a previous reconstructions using FISH and computational analysis (M. Farré et al., 2019). The “No. ingroups” indicates the number of ruminant species comprised by each ancestor.

Ancestor	No. RACFs	FISH number	Max. Size (Kb)	Min. Size (Kb)	Total Size (Kb)	% *	No. ingroups
RAK	40	23+X	192,591	543	2,406,225	88.6	20
RAK (curated)	27	23+X	192,591	1,398	2,406,225	88.6	20
PAK	32	28+X	151,319	855	2,453,880	90.35	19
PAK (curated)	30	28+X	151,319	855	2,453,880	90.35	19
BOCE	30	-	151,973	1,579	2,472,791	91.05	16
BOMO	30	-	154,317	7,360	2,493,967	91.18	11
BOVIN	30	29+X	184,490	365	2,530,756	93.83	9

*Coverage against cattle genome (size=2.71Gb)

Overall, RAK contained 27 RACFs, with RAK10 and RAKX fragmented in two and three pieces, respectively. Taking this into consideration, our final reconstruction, same that previously reported (M. Farré et al., 2019), shows a haploid number of 23+X chromosomes, ranging from 192,591 to 1,398 Kb (**Table 3.1**). Instead, PAK consists of 30 RACFs, with PAKX fragmented in two pieces, showing the same haploid number (28+X) than the pecoran ancestor reported by (M. Farré et al., 2019). BOCE and BOMO are being reported for the first time in this study, and both of them contain 30 RAFCs with a haploid number of 28+X, due to BOCEX and BOMO11 being fragmented in two pieces each. Finally, BOVIN, which is the closest ancestor to cattle, present 30 RACFs, with BOVIN24 fragmented in two pieces. For BOVIN, our haploid number (28+X) lacks one chromosome after comparing to the previously reported ancestor (29+X) (M. Farré et al., 2019).

To identify CRs shaping ruminant genome evolution we cross-compared the reconstructed ancestral karyotypes. We focused on determining the CRs between RAK and PAK (**Figure 3.5**), and PAK and BOVIN (**Figure 3.6**). A total of 59 CRs differentiated PAK from RAK, with 12 fissions, three fusions, and 44 inversions (**Figure 3.5**). These results clearly show the importance of interchromosomal rearrangements between the evolution from RAK to PAK, as well as the main role that inversions played during evolution, being the main type of CRs. Finally, only six chromosomes were maintained as HSBs, RAK chromosomes 11, 12, 13, 16, 21, and 22.

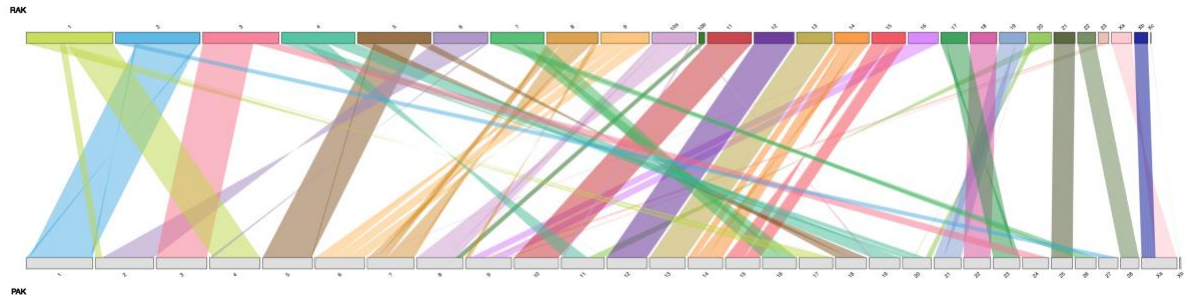


Figure 3.5. Pairwise comparison between RAK and PAK.

Comparing PAK to BOVIN the number of CRs is clearly lower than in the comparison between RAK and PAK. A total of eight CRs differentiated PAK from BOVIN, with one fusion and seven inversions (**Figure 3.6**). These results are evidence of the main role that intrachromosomal rearrangements have played during the evolution from PAK to BOVIN. Between these two ancestors there are 21 chromosomes that have been conserved as HSBs, PAK chromosomes 2, 3, 4, 5, 7, 8, 11, 13, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, and 28.

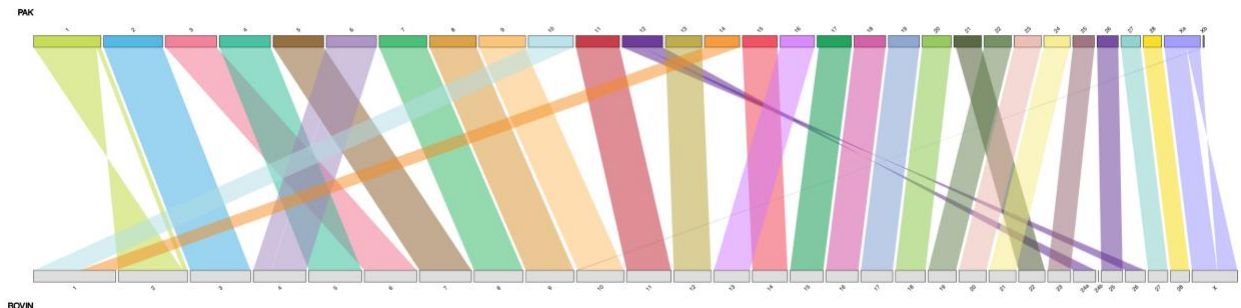


Figure 3.6. Pairwise comparison between PAK and BOVIN.

Only comparing both ancestors, it is clear that ruminant karyotypes have been shaped mainly in an interval of time between 44 Mya and 28 Mya, when RAK and PAK karyotypes appeared, respectively. During this period of time, the chromosome rearrangements rates between these ancestors were 0.94 interchromosomal rearrangements per million years, 2.75 inversions per million years, and a total of 3.69 CRs per million years. However, looking at the evolution from PAK to BOVIN, we can see a clear reduction of chromosome rearrangements rates: 0.1 interchromosomal rearrangements per million years, 0.7 inversions per million years and a total of 0.8 CRs per million years.

As we used cattle as reference species, we can also compare the number of CRs that separate cattle from all the ancestors (**Figures 3.7, S1 and S2 and Table 3.3**).

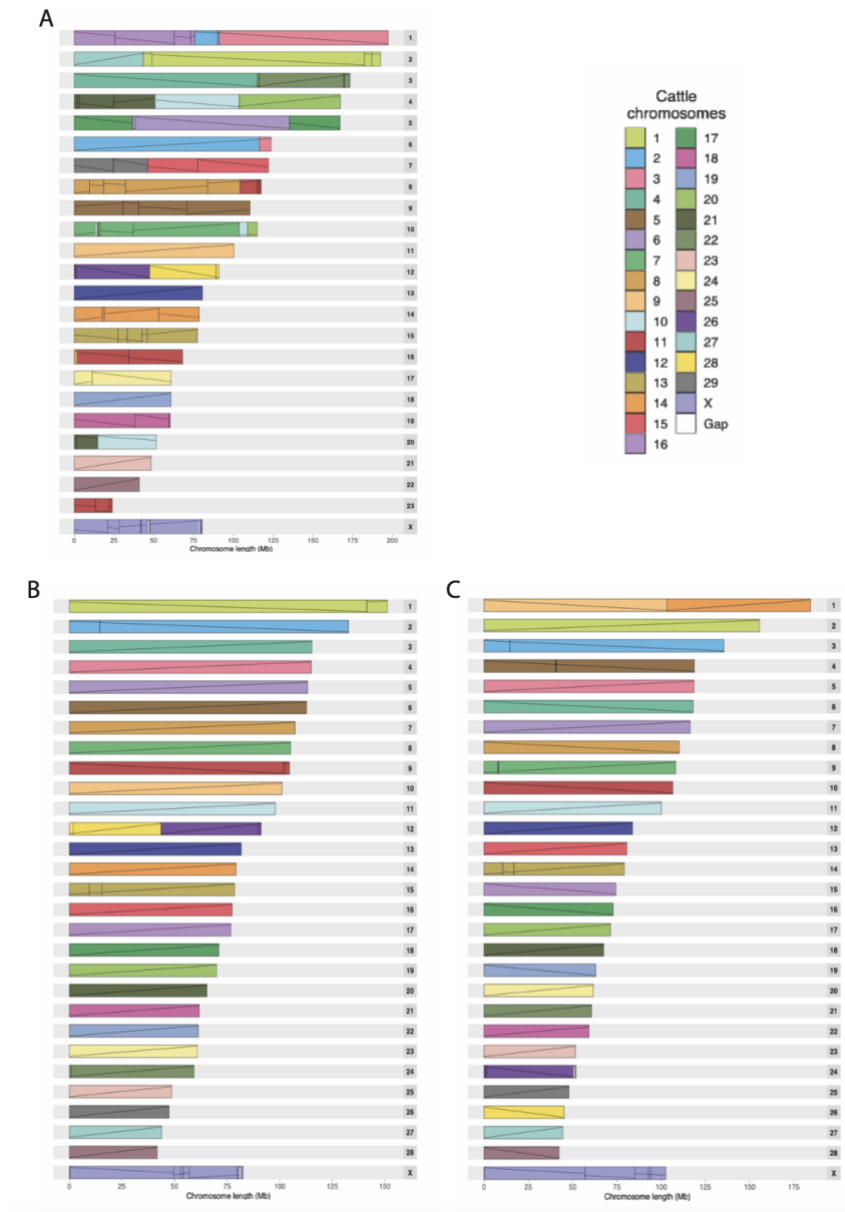


Figure 3.7. Ideograms comparing RAK (A), PAK (B), and BOVIN (C) with cattle chromosomes. Ancestral chromosomes are colour-coded using cattle chromosomes.

As mentioned before, the main episode of rearrangements occurred between RAK and PAK ancestors, which explains why RAK presented a more rearranged configuration compared to cattle than the rest of the ancestors. RAK (**Figure 3.7A**) differs from cattle karyotype in six fusions, ten fissions, and 54 inversions, showing interchromosomal, inversion, and total rearrangement rates per million years of 0.36, 1.22, and 1.59, respectively. PAK (**Figure 3.7B**) differs from cattle in two fusions and 15 inversions, with interchromosomal, inversion, and total rearrangement rates per million years of 0.07, 0.5, and 0.6, respectively. BOCE (**Figure**

S1) presents one fusion and nine inversions while **BOMO (Figure S2)** presents one fusion, three fissions, and eight inversions compared to cattle karyotype. Their interchromosomal, inversion, and total rearrangement rates per million years are 0.04, 0.33, and 0.37 for **BOCE** and 0.16, 0.32, and 0.48 for **BOMO**. Finally, the youngest ancestral karyotype reconstructed, **BOVIN (Figure 3.7C)**, differs from cattle in two fusions and 13 inversions, with 0.11, 0.72, and 0.83 of interchromosomal, inversion, and total rearrangement rates per million years of evolution, respectively.

CRs between ancestors can revert back, meaning that these when looked using cattle as reference are not detectable, this explains why the numbers don't always add up (Table 3.3).

Table 3.3. Summary table of type of CR differentiating every ancestor from cattle.

No ruminant species used	22				
No non-ruminant species used	6				
No ancestral genomes reconstructed	5				
	RAK-Cattle	PAK-Cattle	BOCE-Cattle	BOMO-Cattle	BOVIN-Cattle
No of CRs	54 inversions	15 inversions	9 inversions	8 inversions	13 inversions
	6 fusions	2 fusions	1 fusion	1 fusion	2 fusions
	10 fissions	0 fissions	0 fissions	3 fissions	0 fissions
Rearrangement rate (per my)	3.69	0.8	0.37	0.48	0.83

3.2.3 Assessing the reconstructions

To assess the robustness of our reconstructions, we reconstructed the same set of ancestors using different genome assemblies (see in **Methods Table 2.1**). We used the same parameters and reference genome as previously: 300 Kb resolution of HSBs detection, cattle as reference species, and five ruminant ancestral karyotypes reconstructed (RAK2, PAK2, BOCE2, BOMO2, and BOVIN2). Many of the species used in the first reconstruction were maintained in this analysis, with the exception of five species that were excluded (wild yak, Przewalski's gazelle, red deer, white-tailed deer, and Siberian musk deer), and five species that were included (sheep, reindeer, and three outgroup species: horse, chimpanzee, and rat). Additionally, the genome of Java mouse deer was changed for a new version not published yet as part of a collaboration.

Table 3.4. DESCHRAMBLER results with second set of species. “FISH number” is the haploid number of chromosome reported by a previous reconstructions using FISH and computational analysis (M. Farré et al., 2019). The “No. ingroups” indicates the number of ruminant species comprised by each ancestor.

Ancestor	No. RACFs	FISH number	Max. Size (Kb)	Min. Size (Kb)	Total Size (Kb)	% *	No. ingroups
RAK2	46	23+X	194,787	539	2,459,972	90.6	17
RAK2 (curated)	27	23+X	203,471	1,085	2,459,972	90.6	17
PAK2	34	28+X	155,723	313	2,490,374	91.7	16
PAK2 (curated)	30	28+X	155,723	1,618	2,490,374	91.7	16
BOCE2	32	-	155,975	365	2,529,422	93.14	13
BOMO2	31	-	156,462	365	2,553,715	94.03	9
BOVIN2	32	29+X	156,462	714	2,569,283	94.6	8

*Coverage against cattle genome (size=2.71Gb)

Although eleven species were different, final numbers of RACFs obtained after the curation were very similar, with same numbers for RAK2 and PAK2, two more RAFCs in BOCE2 and BOVIN2, and one more in BOMO2. In terms of coverage compared to cattle genome, the second reconstruction have slightly bigger percentages, ranging from 90.6 in RAK2 to 94.6 in BOVIN2 (**Table 3.4**).

Comparing RAK (**Figure 3.7A**) and RAK2 (**Figure 3.8A**) we can see 16 RACFs with the same structure, representing 71% of all RAK2 ancestral chromosomes. Of all the differences that were observed, some were due to five unplaced RACFs in RAK2 that appear together in RAK chromosome 7, while others were due to CRs disagreements between both reconstructions, including RAK 2 chromosomes 1, 3, 11, and 13. Between PAK (**Figure 3.7B**) and PAK2 (**Figure 3.8B**), nine inversions were detected in PAK2 chromosomes 6, 7, 14, and 16, which were not present in PAK, and two inversions were detected in PAK not present in PAK2. BOCE and BOMO ancestors (**Figures S1 and S2**) present the same structure than BOCE2 and BOMO2, respectively. Finally, BOVIN (**Figure 3.7C**) and BOVIN2 (**Figure 3.8C**) only differ in BOVIN chromosome 1, which in BOVIN2 is split in BOVIN2 chromosomes 10 and 14, and in BOVIN2 chromosome X, which reports the fusion of a small fragment of cattle chromosome 2 with chromosome X, not reported in BOVIN.

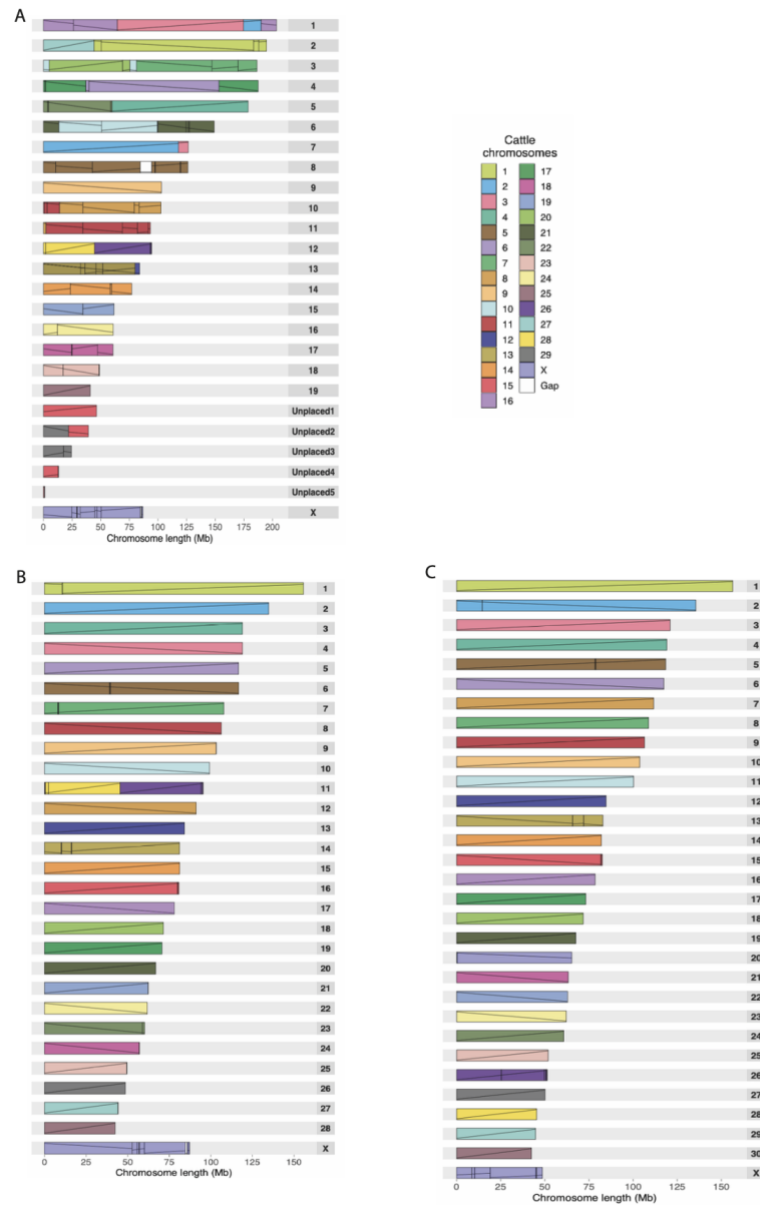


Figure 3.8. Ideograms comparing RAK2 (A), PAK2 (B), and BOVIN2 (C) with cattle chromosomes. Ancestral chromosomes are colour-coded using cattle chromosomes.

Overall, both reconstructions agreed in most RACFs and CRs, clearly indicating that our reconstructions are robust. But reconstructions performed with the first set of species showed a better resolution and agreement with FISH data, probably due to a couple of species in the second set that were assembled at scaffold-level (see in **Methods Table 2.1**). This led to the fragmentation of some of the RACFs produced by DESCHRAMBLER. This fact clearly shows that the selection of species to reconstruct ancestors using DESCHRAMBLER is key. As such, to achieve the best possible results, a high-quality assembly is needed, using species assembled in chromosome and HiC-scaffold if possible. For the following sections of this chapter, we used the reconstruction of the first set of species.

3.2.4 Detection and study of evolutionary breakpoint regions (EBRs)

Using the alignment files of cattle with all the ancestors, EBRs and HSBs were detected and classified. We found a total of 56 EBRs, divided in 32 well-defined EBRs (when their size was between 0 and 50 Kb) and 24 not well-defined EBRs (when their size was between 50 and 300 Kb), while 34 were considered gaps. The EBRs were also classified depending on whether they were cattle or ancestor specific. Within the 32 well-defined EBRs, three were cattle-specific, three were BOVIN-specific, two were RAK+PAK-specific, and 24 were RAK specific. Within the 24 not well-defined EBRs, five were cattle-specific, one was RAK+PAK-specific, and 18 were RAK specific. This classification (**Table 3.5**) reports more EBRs and therefore, more CRs specific of the evolution from RAK to the current cattle genome.

Table 3.5. Classification of cattle EBRs.

Chromosome	Start	End	Classification	Type
1	151,472,023	151,487,795	RAK specific	Well-defined EBRs
2	121,738,699	121,738,700	cattle specific	Well-defined EBRs
3	112,356,219	112,356,219	RAK specific	Well-defined EBRs
5	79,097,362	79,115,205	cattle specific	Well-defined EBRs
5	79,520,453	79,541,481	cattle specific	Well-defined EBRs
5	109,396,514	109,396,514	RAK specific	Well-defined EBRs
7	3,601,581	3,633,466	RAK specific	Well-defined EBRs
7	21,527,243	21,527,243	RAK specific	Well-defined EBRs
7	42,959,455	42,959,455	RAK specific	Well-defined EBRs
8	74,712,035	74,753,191	RAK specific	Well-defined EBRs
8	110,810,351	110,810,351	RAK specific	Well-defined EBRs
10	65,451,273	65,451,273	RAK specific	Well-defined EBRs
10	102,765,562	102,765,562	RAK specific	Well-defined EBRs
11	46,168,548	46,168,548	RAK specific	Well-defined EBRs
11	59,816,851	59,830,207	RAK specific	Well-defined EBRs
11	92,296,003	92,296,003	RAK+PAK specific	Well-defined EBRs
11	103,195,815	103,221,935	RAK+PAK specific	Well-defined EBRs
11	104,148,952	104,154,937	RAK specific	Well-defined EBRs
14	44,825,761	44,825,761	RAK specific	Well-defined EBRs
14	80,922,608	80,922,608	RAK specific	Well-defined EBRs
15	34,354,323	34,378,969	RAK specific	Well-defined EBRs
16	41,016,027	41,016,027	RAK specific	Well-defined EBRs

16	51,587,821	51,587,821	RAK specific	Well-defined EBRs
16	54,789,351	54,789,351	RAK specific	Well-defined EBRs
17	35,809,155	35,809,155	RAK specific	Well-defined EBRs
18	40,211,040	40,211,040	RAK specific	Well-defined EBRs
21	29,271,954	29,271,954	RAK specific	Well-defined EBRs
22	1,469,989	1,470,157	BOVIN specific	Well-defined EBRs
24	12,545,173	12,545,173	RAK specific	Well-defined EBRs
26	381,823	386,612	BOVIN specific	Well-defined EBRs
28	2,842,626	2,842,626	BOVIN specific	Well-defined EBRs
29	26,488,536	26,497,630	RAK specific	Well-defined EBRs
1	139,172,168	139,423,787	RAK+PAK specific	Not well-defined EBRs
1	146,726,363	146,840,384	RAK specific	Not well-defined EBRs
2	120,088,667	120,269,146	RAK specific	Not well-defined EBRs
2	121,336,349	121,431,663	cattle specific	Not well-defined EBRs
5	42,289,087	42,524,587	RAK specific	Not well-defined EBRs
5	75,986,990	76,065,898	RAK specific	Not well-defined EBRs
8	11,061,887	11,204,254	RAK specific	Not well-defined EBRs
8	90,088,231	90,178,399	RAK specific	Not well-defined EBRs
10	6,369,544	6,647,026	RAK specific	Not well-defined EBRs
11	9,370,122	9,451,181	RAK specific	Not well-defined EBRs
13	17,606,514	17,672,822	cattle specific	Not well-defined EBRs
13	46,973,602	47,268,723	RAK specific	Not well-defined EBRs
14	18,894,568	19,017,836	RAK specific	Not well-defined EBRs
18	1,098,245	1,371,975	RAK specific	Not well-defined EBRs
20	6,607,444	6,859,219	RAK specific	Not well-defined EBRs
21	23,269,496	23,562,461	RAK specific	Not well-defined EBRs
21	25,052,205	25,163,007	RAK specific	Not well-defined EBRs
21	55,911,356	56,186,157	RAK specific	Not well-defined EBRs
22	55,854,666	56,092,341	RAK specific	Not well-defined EBRs
26	51,097,954	51,234,471	cattle specific	Not well-defined EBRs
X	38,569,267	38,853,992	RAK+PAK specific	Not well-defined EBRs
X	54,821,241	55,096,761	All but RAK specific	Not well-defined EBRs
X	84,757,216	84,852,751	RAK specific	Not well-defined EBRs
X	94,116,978	94,401,267	RAK specific	Not well-defined EBRs

Well-defined EBRs are not uniformly distributed along cattle chromosomes, there is not a correlation between number of EBRs and chromosome size (**Figure 3.9A**). In fact, some of the

chromosomes with a medium size (average cattle chromosome size is 87 Mb) present the highest number of well-defined EBRs. Chromosome 11 is the one containing more well-defined EBRs (5) while cattle chromosomes 4, 6, 9, 12, 13, 19, 20, 23, 25, 27 and X have not well-defined EBRs. However, when we expand the EBR selection and study both, well-defined and not well-defined EBRs together (**Figure 3.9B**), there is a correlation between number of EBRs and chromosome size (P-value = 0.015). In this scenario, chromosome 11 is still the one containing more EBRs (6), however, cattle chromosomes 4, 6, 9, 12, 19, 23, 25 and 27 have no EBRs at all. The lack of EBRs in these chromosomes could be due to: (i) lack of resolution, some of the CRs would be delimited by EBRs higher than 300 Kb, and in these cases, we have considered them as gaps instead of EBRs, one example is in chromosome X; or (ii) cases where no CRs happened during evolution, such as cattle chromosomes 4 and 9.

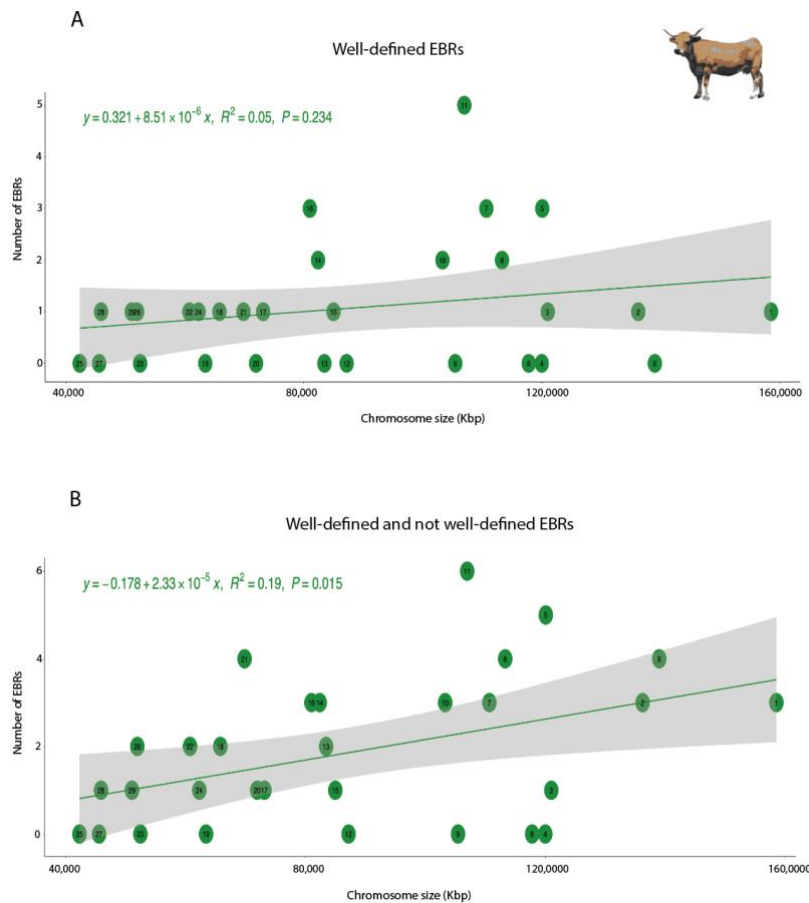


Figure 3.9. Correlation between the number of well-defined (A) and well-defined and not well-defined (B) EBRs and cattle chromosome sizes in Kb.

To determine whether EBRs occur in non-random locations of the genome, we first studied their co-location within gene-rich regions. For this purpose, several permutation tests were

performed over three different associations: (i) regions within well-defined EBRs; (ii) regions within not well-defined EBRs; and (iii) regions near EBRs (**Figure 3.10**). A clear negative association (Z-score = -2.191 and P-value = 0.0195) between well-defined EBRs and genes was found (**Figure 3.10A**) and, furthermore, regions near these well-defined EBRs showed higher number of genes than the EBRs as such (P-value in the 10 Kb downstream region from well-defined EBRs = 0.013 and P-value in the 17 kb upstream region from well-defined EBRs = 0.011) (**Figure 3.10C**), following previous publications (M. Farré et al., 2019). Because many of the well-defined EBRs are very narrow, sometimes even with sizes of 0 or 1 bp, it is highly unlikely that genes are found within these EBRs, instead well-defined EBRs are surrounded by gene-rich regions, as pointed previously (M. Farré et al., 2019). For not well-defined EBRs no statistically significant association was found (**Figure 3.10B**).

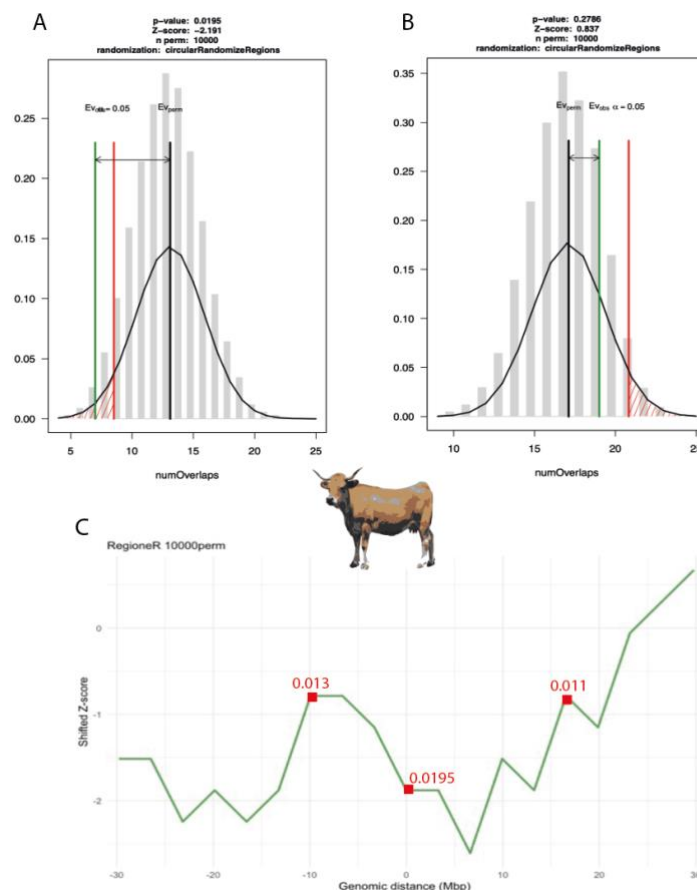


Figure 3.10. Permutation tests performed to establish the association between gene content and well-defined EBRs (A), not well-defined EBRs (B), and regions near well-defined EBRs (C) in cattle. In C, red number show P-values in the regions marked with a red square.

We then determined the type of genes located within or near EBRs by performing GOEA (**Figure 3.11**). For the well-defined EBRs, there were eight genes present (*PIK3R4*, *APTX*,

AADACL3, an open reading frame (*C11H2orf74*), and four novel genes (ENSBTAG000000039186, ENSBTAG00000000893, ENSBTAG000000052161, ENSBTAG00000015027)), leading to no significant GO terms. We then expanded the well-defined EBRs, up- and downstream until all of them reached 50 Kb in size. After the expansion, 29 genes were included in the analysis, and only lipid glycositation GO term was significant (**Figure 3.11**). On the other hand, when we performed GOEA in the not well-defined EBRs, 48 genes were included, leading to an enrichment of genes related to enzymatic activity (N-acetyl-beta-D-galactosaminidase, hexosaminidase and alkaline phosphatase activity), copper ion binding, or with components of the membrane, among others (**Figure 3.11**).

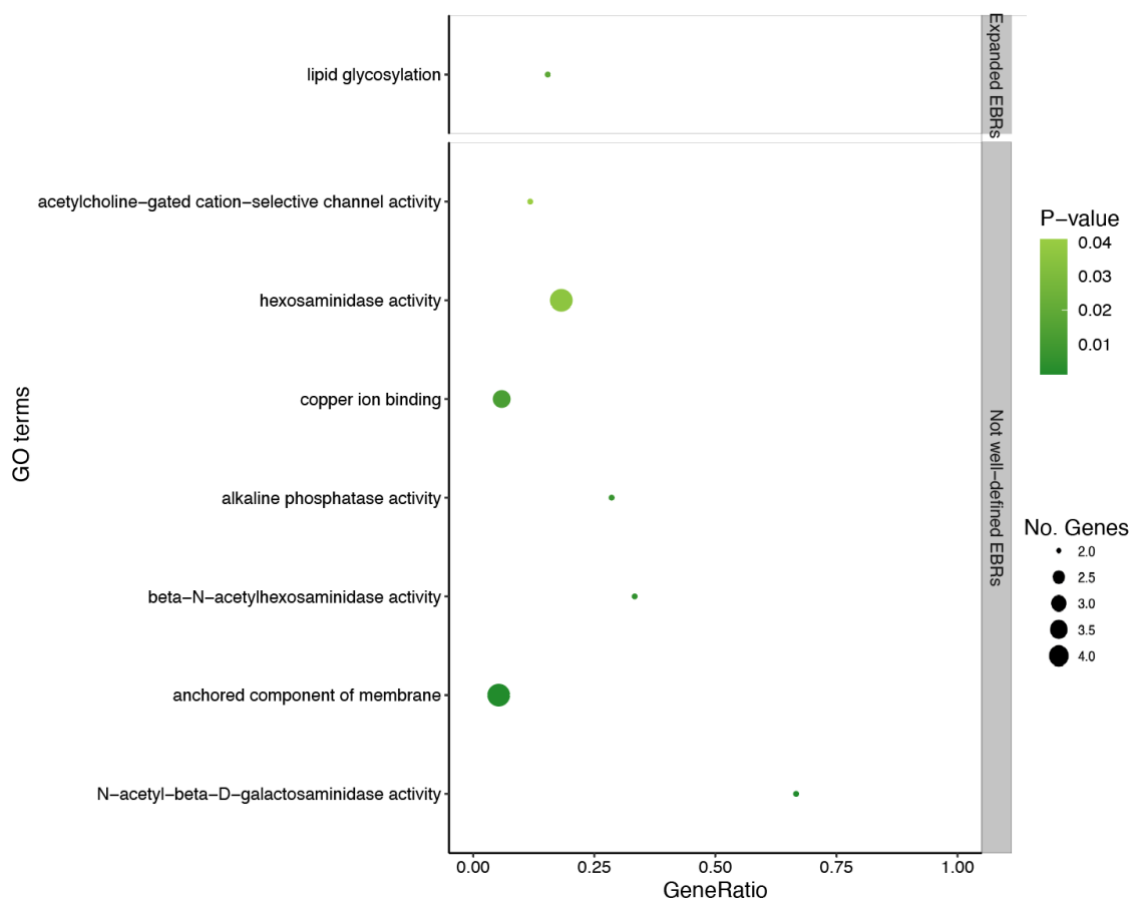


Figure 3.11. GOEA of cattle well-defined EBRs expanded to 50 Kb (top) and not well-defined EBRs (bottom). X axis represents the gene ratio (number of genes/total number of genes). Y axis represents the GO terms. Bubbles size varies depending on their number of genes in every GO term (No. Genes). Bubbles colour varies depending on their P-value.

Finally, we wanted to know if the EBRs were in regions of the genome with high density of repetitive elements. For this purpose, we performed an enrichment analysis to look for genomic association between repetitive elements and well-defined EBRs and not well-defined EBRs. In

both cases we found an enrichment in repetitive elements (**Figure 3.12**). One of the most remarkable differences is that well-defined EBRs presented a very diverse repetitive elements content (58 different types) while not well-defined EBRs presented less variability of repetitive elements (19 different types). Among these classes of repeats, some of them are ruminant specific, such as BovB and Bov-tA3. For well-defined EBRs, Bov-tA3 showed a negative log₂ of fold change (12fold), which is indicative of less presence of this element than it was expected, although in the not well-defined EBRs, Bov-tA3 has a positive 12fold. The most common type of repeats are long interspersed nuclear elements (LINEs; L1), and long terminal repeats (LTRs), medium reiterated repeats (MER), and endogenous retroviruses (ERVs).

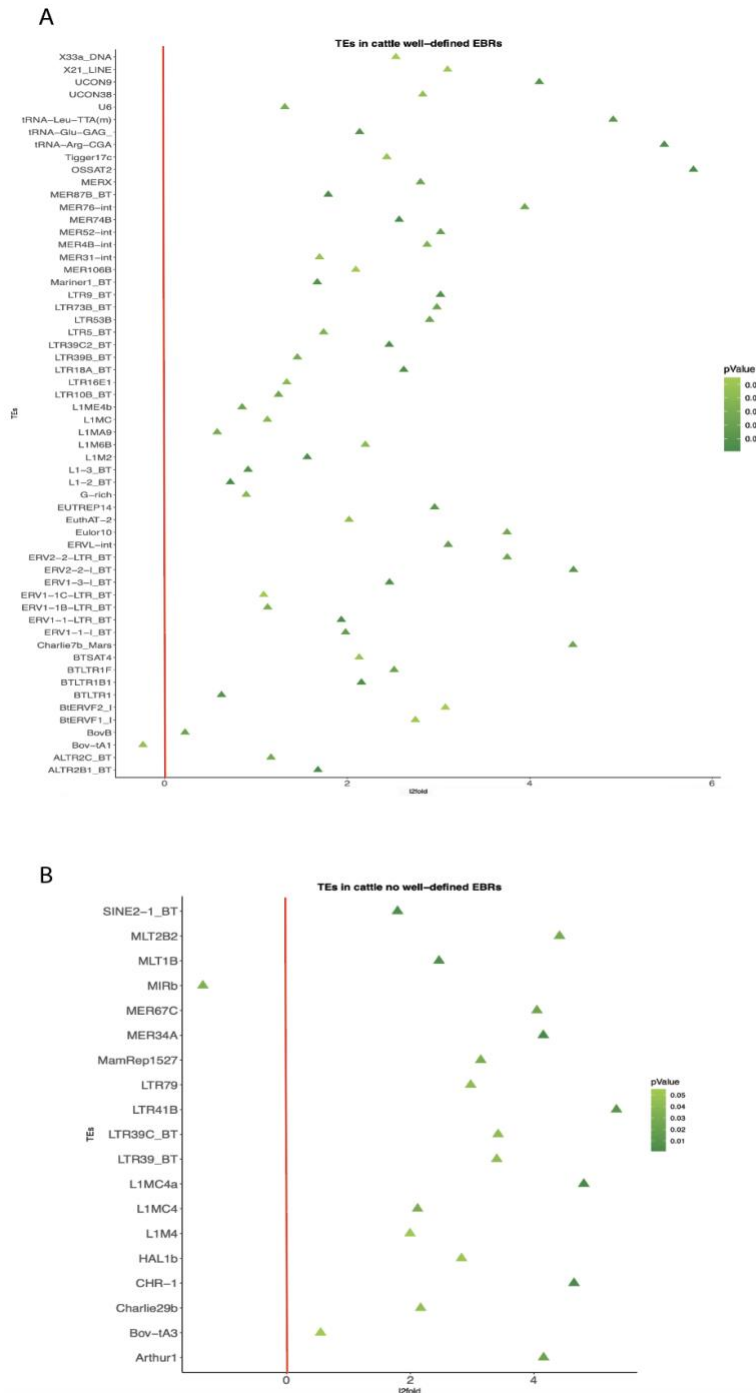


Figure 3.12. TEs associated with well-defined (A) and not well-defined EBRs (B). X axis represents the log2fold, which indicates less TEs observed than expected in case of negative numbers and more TEs observed than expected in case of positive numbers. The red line highlights the division of negative and positive log2fold values. Y axis represents the different TEs. Triangles colour varies depending on their P-value.

3.2.5 Detection and study of homologous synteny blocks (HSBs)

After locating the EBRs, by definition, the fragments comprised between them would be the so called homologous synteny blocks (HSBs). These HSBs can be classified depending on the

type of rearrangement that have generated them in collinear (when they have been conserved between species), inversion, fission, fusion, or a combination of some of the previous cases (to see an explanation of these special cases see **Methods Figure 2.2**).

In cattle, we have classified these HSBs according to the rearrangements occurring during the evolution from our ancestral karyotype reconstructions to the current cattle genome (**Table 3.6**). Within this classification, 48.84% of cattle genome corresponds to collinear HSBs, 8.77% to inverted HSBs, and 40.48% to HSBs related to interchromosomal rearrangements (specifically, 17.74% to fissioned and collinear HSBs, 21.37% to fissioned and fused HSBs, and 1.67% to fissioned and inverted HSBs). These results together with our previous results, remark the important role of interchromosomal rearrangements in genome reshuffling during Ruminantia species evolution.

Table 3.6. Classification of HSBs in cattle.

Chromosome	Start	End	Classification
1	613,557	139,172,168	Collinear
1	139,423,787	146,594,703	Inversion
1	146,840,384	151,472,023	Inversion
1	151,487,795	157,720,637	Collinear
2	84,594	120,088,667	Fission+Fusion
2	120,269,146	121,156,470	Fission+Fusion
2	121,431,070	121,738,698	Fission+Fusion
2	121,738,699	136,231,102	Fission+Fusion
3	145,957	112,356,219	Fission+Fusion
3	112,356,219	121,005,158	Fission+Fusion
4	169,304	119,917,700	Collinear
5	551,555	42,289,087	Collinear
5	42,524,587	75,986,990	Inversion
5	76,065,898	109,396,514	Collinear
5	109,396,514	120,081,431	Inversion
6	1,091,372	2,982,940	Fission+Collinear
6	3,363,160	117,795,612	Fission+Collinear
7	357,099	1,764,723	Fission+Collinear
7	2,512,693	3,601,581	Fission+Inversion
7	3,633,466	21,527,243	Fission+Collinear
7	21,527,243	42,959,455	Fission+Inversion
7	42,959,455	110,680,435	Fission+Collinear

8	670,939	11,061,887	Collinear
8	11,204,254	65,361,375	Collinear
8	65,708,799	74,712,035	Inversion
8	74,753,191	90,088,231	Inversion
8	90,178,399	110,810,351	Collinear
8	110,810,351	113,041,577	Fission+Fusion
9	694,369	104,143,274	Collinear
10	333,206	6,369,544	Fission+Fusion
10	6,647,026	65,451,273	Fission+Fusion
10	65,451,273	102,765,562	Fission+Fusion
10	102,765,562	103,282,895	Fission+Fusion
11	56,801	9,370,122	Fission+Fusion
11	9,451,181	43,958,071	Fission+Fusion
11	44,280,529	46,168,548	Inversion
11	46,168,548	59,816,851	Fission+Fusion
11	59,830,207	92,296,003	Fission+Fusion
11	92,296,003	103,195,815	Fission+Fusion
11	103,223,387	104,147,843	Collinear
11	104,154,937	105,840,497	Inversion
11	106,159,169	106,982,470	Collinear
12	169,741	87,214,505	Collinear
13	554,884	10,222,205	Inversion
13	11,454,310	17,401,651	Collinear
13	17,995,276	46,973,602	Inversion
13	47,268,723	50,748,545	Inversion
13	51,120,857	83,467,419	Collinear
14	122,750	18,894,568	Collinear
14	19,017,836	44,825,761	Inversion
14	44,825,761	80,922,608	Collinear
14	80,922,608	82,267,568	Inversion
15	1,182,921	34,354,323	Collinear
15	34,378,969	84,003,417	Collinear
16	637,003	41,016,027	Fission+Collinear
16	41,016,027	51,587,821	Fission+Collinear
16	51,587,821	54,789,351	Fission+Collinear
16	54,789,351	81,005,948	Fission+Collinear
17	3,179,044	35,809,155	Fission+Collinear
17	35,809,155	73,139,051	Fission+Collinear
18	495,078	1,098,245	Collinear
18	1,371,975	40,211,040	Collinear

18	40,211,040	65,818,766	Collinear
19	518,056	63,425,567	Collinear
20	207,372	6,607,444	Fission+Fusion
20	6,859,219	71,919,352	Fission+Fusion
21	761,477	23,269,496	Fission+Inversion
21	23,562,461	25,052,205	Fission+Collinear
21	25,163,007	27,217,560	Fission+Collinear
21	27,827,871	29,271,954	Fission+Fusion
21	29,271,954	55,911,356	Fission+Fusion
21	56,186,157	69,862,954	Fission+Collinear
22	120,147	1,282,456	Inversion
22	1,470,526	55,854,666	Collinear
22	56,092,341	56,568,886	Inversion
22	57,173,296	60,713,155	Collinear
23	450,236	52,169,786	Collinear
24	305,472	12,545,173	Inversion
24	12,545,173	62,161,800	Collinear
25	147,382	42,350,435	Collinear
26	726,785	49,637,879	Collinear
26	50,625,508	51,097,954	Inversion
26	51,234,471	51,992,301	Collinear
27	461,316	45,611,885	Fission+Collinear
28	875,359	2,627,229	Inversion
28	2,888,472	45,877,197	Collinear
29	726,520	26,488,536	Fission+Collinear
29	26,497,630	50,930,707	Fission+Collinear
X	381,897	33,985,355	Collinear
X	35,402,189	36,257,020	Collinear
X	36,754,074	37,297,417	Collinear
X	39,926,018	64,236,228	Collinear
X	65,573,288	84,757,216	Inversion
X	84,852,751	94,116,978	Collinear
X	94,401,267	94,949,461	Inversion
X	97,484,801	139,009,125	Inversion

We also performed GOEA in all the types of HSBs (**Figure 3.13**). A total of 10,541 genes were found in collinear HSBs and showed GO terms related to DNA binding, pheromone activity, biological regulation, and DNA-binding transcription factor activity, while the 1,682 genes found in inverted HSBs showed GO terms related to enzyme activity, cytolysis, and metabolic processes. Interestingly, no common GO term was found between the three cases involving interchromosomal rearrangements, instead the fissioned and collinear HSBs presented 3,798 genes with significant GO terms related to chemical activities; fissioned and fused HSBs presented 4,176 genes with significant GO terms related to RNA polymerase and binding; and fissioned and inverted HSBs presented 350 genes with significant GO terms related to signalling, detection of stimulus, and cell communication.

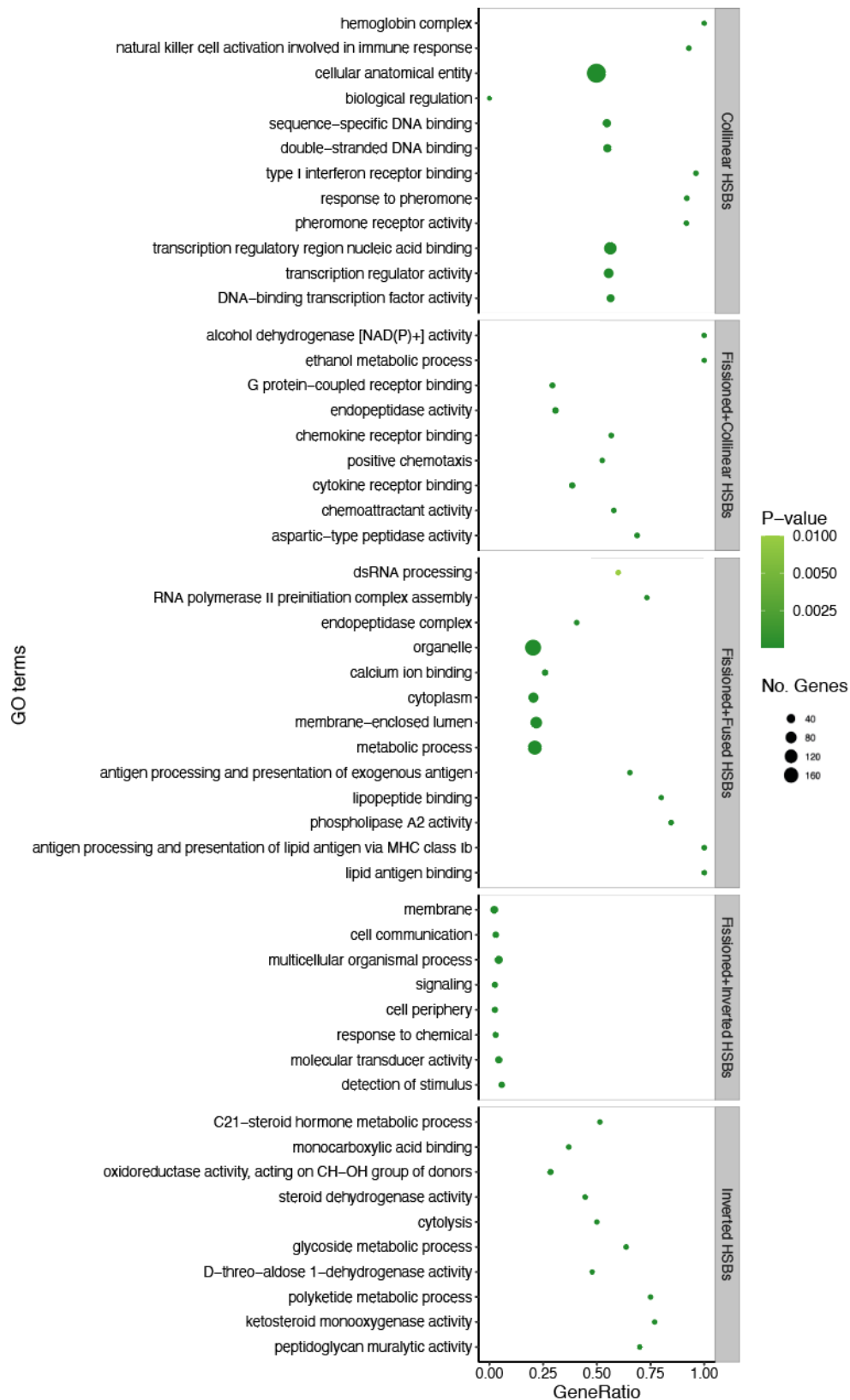


Figure 3.13. GOEA of cattle HSBs according to their classification. X axis represents the gene ratio (number of genes/total number of genes). Y axis represents the GO terms. Bubbles size varies depending on their number of genes in every GO term (No. Genes). Bubbles colour varies depending on their P-value.

3.3 Discussion

In this chapter, we propose a new karyotype for five different ruminant ancestors: the ancestor of all ruminant families (RAK), the ancestor of Pecoran species (PAK), the ancestor of Cervidae, Moschidae and Bovidae families (BOCE), the ancestor of Moschidae and Bovidae families (BOMO) and the ancestor of Bovidae family (BOVIN). RAK, PAK, and BOVIN were reconstructed in previous publications using the same algorithm, DESCHRAMBLER, but a different set of species (M. Farré et al., 2019). In this previous study, the authors compared chromosome evolution from the Pecoran ancestor directly to the bovid ancestor, while in this thesis, by contrast, we reconstructed two intermediate ancestors for the first time (BOCE and BOMO), allowing us to better classify the rearrangements occurring in each of them, and more importantly, to detect cattle-specific (or at least Bovidae-specific) rearrangements.

Additionally to previous findings (M. Farré et al., 2019), our RAK configuration reported new rearrangements: (i) RAK chromosome 1 corresponds to cattle chromosomes 2, 3, and 6, without any fragmentation of chromosome 6; (ii) the fusion of cattle chromosomes 1 and 27; (iii) different pattern of rearrangements in cattle chromosomes 15 and 29; (iv) different pattern of rearrangements involving cattle chromosomes 7, 10, 20, and 21; and (v) many inversions smaller than the minimum resolution of FISH probes. This can be explained because in our reconstruction we used ten more ruminant species and more high-quality genome assemblies.

RAK chromosome 1 differs between our two reconstructions, with the second reconstruction being more similar to previous publications (M. Farré et al., 2019). In this case, chromosome 1 in RAK is supported by our main set of species while in RAK2 we achieved the configuration by manually curating it, suggesting that the configuration proposed by RAK1 might be more accurate. Another controversial chromosome is the one that involves chromosomes 15 and 29 of cattle, which in RAK appears as an only fusion of both chromosomes while in RAK2 these cattle chromosomes appear as different fragments, following previous reconstructions (M. Farré et al., 2019), they appear as fissioned fragments joined alternatively. In this specific case, all extant species genomes support our RAK reconstruction, suggesting that previous reconstructions might be erroneous because they included more genomes assembled at scaffold-level. Finally, we also reported the fusion of cattle chromosome 1 and 27 in both of our reconstructions, in disagreement with previous publications.

Looking at the pecoran ancestor, only one interchromosomal rearrangement is reported in our reconstructions, and is in agreement with previous publications (M. Farré et al., 2019). However, we detected more inversions than before, clearly showing that the selection of species impacts the quality of the reconstructions and directly increases the resolution of detection of CRs. Finally, looking at the Bovidae ancestor of the previous reconstruction (M. Farré et al., 2019) and ours (BOVIN and BOVIN2), BOVIN2 has the same configuration than previous studies but with less fragmentation and subsequently better resolution. However, BOVIN reported a fusion of chromosomes 9 and 14 in cattle, that have not been described before and that is supported by all the bovid species. Despite the fact that these three reconstructions allow a very precise estimation of ruminant evolution, including BOCE and BOMO is fundamental to confirm that the fusion of cattle chromosomes 9 and 14 and the fission of cattle chromosomes 26 and 28 are specific of the Bovidae family, as RAK, PAK, BOCE, and BOMO lack them. Despite these differences, the overall diploid number obtained herein is in agreement with previous studies: our RAK and PAK present the same number, $2n = 46$ and $2n = 58$, respectively, while our BOVIN presents one less pair of chromosomes, $2n = 58$. To assess the reliability of our reconstruction, we performed the same analysis using different sets of species. Our results showed that 88.09% of RAFCs maintained the same structure, and the differences were most likely due to the quality of the genomes of the species included.

Although the majority of reconstructions performed with DESCHRAMBLER nowadays have only used one reference species (Kim et al., 2017; M. Farré et al., 2019; Álvarez-González, Arias-Sardá, et al., 2022), a very recent study (Damas et al., 2022) reconstructed 16 ancestral karyotypes along the mammalian phylogeny using three different reference species. With this new approach for ancestral reconstructions, they described the bovid ancestor. Almost all RACFs agree between our reconstruction and this new publication, except in the fusion of cattle chromosomes 9 and 14 that we reported in BOVIN. This is probably caused by the low number of bovid species they used (only two bovinds). For future work, it would be very interesting to combine the approach used in this recent study (Damas et al., 2022) of reconstructing ancestors by using several reference species, with an exhaustive and extensive selection of species as we did in this chapter.

Taking all these results together, the evolution from RAK to PAK is characterised by interchromosomal rearrangements playing an important role, even though inversions are more

abundant, while the evolution of the rest of ruminant ancestors to current species is driven by inversions.

Overall, our reconstructions presented higher quality ancestral karyotypes compared to the previous ancestors proposed by other studies. This is key not only to describe the CRs that have shaped ruminant genomes, but also to define with high accuracy the EBRs that have originated these CRs. We identified a total of 56 EBRs, with 32 of them described as well-defined EBRs, more EBRs than reported in previous studies (Kim et al., 2017; M. Farré et al., 2019; Damas et al., 2022). Our findings agree with these studies in the correlation of EBRs and genome regions with repetitive sequences. As reported before (M. Farré et al., 2019), we also see an enrichment of ruminant-specific TEs such as BovB, which could provide extra opportunities for non-allelic recombination in the ruminant ancestor germ cells. Furthermore, we also reported an enrichment of more ruminant specific repeat elements such as Bov-tA2, Bov-tA3, and other elements like Charlie elements. And, as reported by (Damas et al., 2022), we also detected the presence of LINEs-L1, and LTRs. However, for the specific case of Bov-tA3, it appeared as negatively correlated in the analysis of well-defined EBRs, which would be expected because our EBRs are narrow and their enrichment could be seen in nearby regions. Contrariwise, not well-defined EBRs, comprising larger regions, showed an enrichment of Bov-tA3, which would confirm our premise of TEs being located in regions near EBRs and not within EBRs. Moreover, we also observed the presence of ERVs, which have been described as conserved repeats in vertebrates due to these elements being generally inactivated by mutations and are then passed vertically to the next generation (Böhne et al., 2008). Overall, we were able to describe the regions of genes that are more prone to break and re-join during evolution (EBRs) and their association with repetitive elements, some of them associated only to ruminant species.

Furthermore, we also associated these EBRs with gene content, and not only we determined the direct association between EBRs and gene-rich regions in the genome, but we also identified the GO terms of the genes located in those regions. We found a total of eight genes within the EBRs, of which only three have a known function: *PIK3R4*, *APTIX*, and *AADACL3*. Of these genes, *AADACL3* is directly related to lipid metabolism (Lu et al., 2020), a distinctive characteristic of ruminants, whose multichambered stomach facilitate the digestion of lipids. Interestingly, *APTIX* is involved in mitochondrial single-strand break repair (SSBR), and whose failure could have implications in genome stability as single-strand breaks (SSBs) may

progress to double-strand breaks (DSBs). These DSBs could then cause DNA instability and/or impaired replication and gene expression from the mitochondrial genome (Iyama & Wilson, 2013; Meagher & Lightowlers, 2014). Our findings agree with the expectation to find genes related to DNA repair mechanisms and genome instability, such as *APTX*, as EBRs are more prone to break and re-join during evolution with the subsequent increase of DSBs that will trigger the activity of these DNA repair mechanisms. Additionally, we found that the well-defined EBRs expanded up- and downstream to reach 50 Kb only showed enrichment in two genes related to lipid glycosylation, a key factor for the milk composition characteristic of bovine species (Poulsen et al., 2019). One of these two genes is *ABO*, a gene that determines the blood group of an individual in humans (Yazer & Palcic, 2005). Interestingly, it has been reported the association between blood groups and the fat percentage of the milk in cattle, as reviewed in (Z. Liu et al., 2019). One of the main differences between human and bovine milk is their oligosaccharides content, with cows having a low amount. One mechanism that bovids use to compensate their low oligosaccharide content in their milk could be related to the overexpression of genes related to lipid glycosylation, such as *ABO*, in their mammary tissue (Z. Liu et al., 2019). In the same vein, when we looked at the GO terms enriched in the not well-defined EBRs, two of them are also related to the specific composition of bovine milk and the glycosylation process in bovids. The first one is related to the alkaline phosphatase activity, in humans reported to be inhibited, in cases of preterm birth and lack of breast milk feeding, impairing the capacity to detoxify proinflammatory bacterial products in premature infants (Y. Yang et al., 2019). The second one is related to the β -N-acetylhexosaminidase activity, related to an increase of the transglycosylation activity in fungi, causing, at the same time, an increase in their oligosaccharide production (Kurakake & Amai, 2022). All these results suggest that EBRs in the ruminant lineage have rearranged genes directly related to milk production and composition. This is in line with a connection between EBRs and the development of new adaptive gene functions (Larkin et al., 2009; Groenen et al., 2012; Capilla et al., 2016). This GOEA opens new opportunities to study gene function and gene expression in regions associated to CRs.

Finally, we were able to identify and classify the HSBs in the cattle genome. Of these HSBs, 48.84% correspond to collinear HSBs and 49.25% correspond to rearranged HSBs, specifically, 8.77% were related to inverted HSBs, and 40.48% were related to HSBs produced by interchromosomal rearrangements, showing a rearrangement rate per My of 1.22 and 0.36, respectively. These results highlight the important role that interchromosomal rearrangements

have played during the evolution of ruminant species. Moreover, we also described the GO terms associated to the specific rearrangements causing these HSBs. In collinear HSBs we found an enrichment in genes related to several transcription-related activities, and furthermore, we also found an enrichment in genes related to olfactory responses to pheromones, which have been described as new adaptative gene functions related to CRs (Larkin et al., 2009; Groenen et al., 2012; Capilla et al., 2016). In inverted HSBs we found an enrichment in genes related to glycoside metabolic processes, peptidoglycan muralytic activity, or monocarboxylic acid binding, which have been identified as key genes in milk fat metabolism (Z. Liu et al., 2019; Mu et al., 2022). Among the different cases of interchromosomal rearrangements, the most interesting results were showed by fissioned+fused HSBs, which are enriched in genes related to RNA polymerase II and lipid-related functions. RNA polymerase II is involved in replication, and it can be related to a stall of the replication fork, which has been related to the development of several tumours (Solvie et al., 2022). On the other side, lipid-related functions are also responsible of the specific bovine milk composition (Toral et al., 2018). Overall, we could not find any enrichment in genes involved in processes related to chromatin conformation or condensation, which would be very important to study the correlation between CRs and chromatin configuration demonstrated in other mammalian species (Marques-Bonet et al., 2007; Fuller et al., 2016; Capilla et al., 2016). For this reason, it would be very interesting to further study these genes alternative networks to try to find any connection with the 3D chromatin structure. For future analysis, it would be very useful to identify repeat content within the HSBs and to compare it with the repeated sequences found in the EBRs associated to that specific CR.



This chapter reports the results from:

Álvarez-González, L., Arias-Sardá, C., Montes-España, L., Marín-Gual, L., Vara, C., Lister, N. C., Cuartero, Y., Garcia, F., Deakin, J., Renfree, M. B., Robinson, T. J., Martí-Renom, M. A., Waters, P. D., Farré, M., & Ruiz-Herrera, A. (2022). Principles of 3D chromosome folding and evolutionary genome reshuffling in mammals. *Cell Reports*, 41(12), 111839.

4. The Marsupialia lineage

4.1 Introduction

The infraclass Marsupialia is formed by 407 species, classified in 91 genera, within 21 families (Acrobatidae, Burramyidae, Caenolestidae, Chaeropodidae, Dasyuridae, Didelphidae, Hypsiprymodontidae, Macropodidae, Microbiotheriidae, Myrmecobiidae, Notoryctidae, Peramelidae, Petauridae, Phalangeridae, Phascolarctidae, Potoroidae, Pseudocheiridae, Tarsipedidae, Thylacinidae, Thylacomyidae and Vombatidae), seven orders (Dasyuromorphia, Didelphimorphia, Diprotodontia, Microbiotheria, Notoryctemorphia, Paucituberculata and Peramelemorphia) and divided into two superorders according to their origin (Ameridelphia, if the species are distributed in the American continent or Australidelphia, if the species are located in Australia) (Burgin et al., 2018; Mammal Diversity Database, 2022).

Some studies suggest that the Ameridelphia superorder is basal to the Australidelphia one, diverging from a common ancestor approximately 80 Mya, during the Cretaceous period (Duchêne et al., 2018). They also pointed that Australian marsupials started diversifying in the early Paleocene (around 66 Mya), during a time when Australia was undergoing a major climatic and floral transformation (**Figure 4.1**).

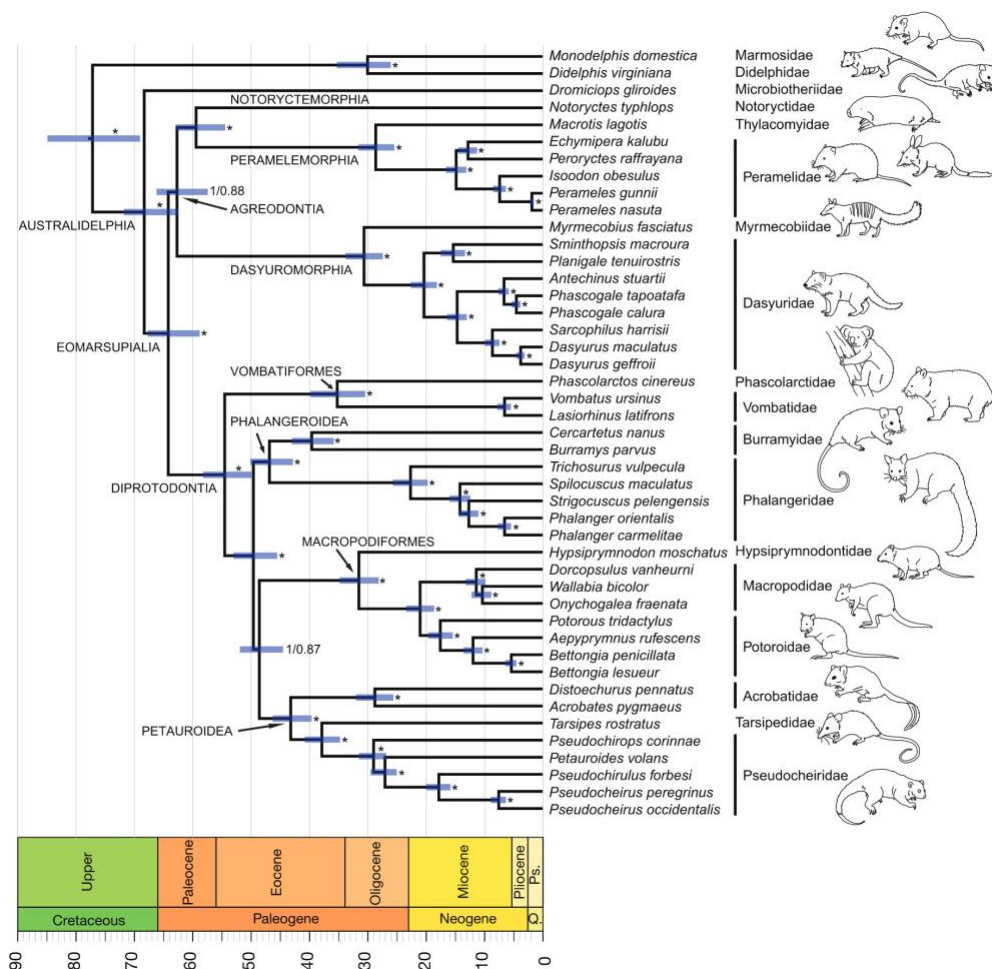


Figure 4.1. Phylogeny of marsupials among the Geological Time Scale (GTM) of Earth. Figure from (Duchêne et al., 2018).

The Marsupialia lineage shows a broad range of species, from little possums to kangaroos, showing both diurnal and nocturnal forms, and having many kinds of diets (including carnivores, specialized insectivores, omnivores, and herbivores) (Mitchell et al., 2014). Marsupials, same as placental mammals, give birth to live young, but the way they do this, leads to a series of characteristics which make marsupials remarkably unique:

- (i) The major difference which characterizes marsupials is the developmental stage in which their young are born. Marsupials give birth to tiny, altricial young, whose development occurs mainly ex utero, typically within the confines of a pouch (Deakin & O'Neill, 2020). This characteristic is the one that give them their name: "marsupial" derives from *marsupium* (in Latin) or pouch. Earlier studies established the equivalence between these young marsupials and an eight- to ten-week-old human embryo (Block, 1960).

- (ii) Another characteristic of marsupials is their sophisticated lactation system with a variable milk composition, changing depending on the developmental stage of the young, and whose length is much longer than in eutherian mammals (Green et al., 1983; Deakin & O'Neill, 2020).
- (iii) Due to their premature birth, the eyes, brain, hindlimbs, reproductive and immune systems, as well as thermoregulation are developed during the young's life in the pouch, contrary to most of mammals whose development occurs in utero (Tyndale-Biscoe, 2005; Deakin & O'Neill, 2020).

Marsupialia comprises species with great diversity in ecology and morphology, and which are distributed across extensive habitats. They can be found in wet tropical rainforests such as New Guinea and the Amazon, as well as in the arid Australian interior (Mitchell et al., 2014). However, despite the great diversity of adaptations displayed by living marsupials, their evolution has been poorly studied due to extensive gaps in the marsupial fossil record. Some studies have related this lack of evidence to several critical periods in marsupials evolutionary history (Archer et al., 1999; Mitchell et al., 2014).

In terms of their genomes, marsupials also show distinct features, such as the packaging of their genomes into fewer and larger chromosomes than most other mammalian species. In contrast to most mammals whose genomes are divided in a big number of small chromosomes, marsupial chromosomes are large in size and small in number. In fact, their chromosomes were one of the first mammalian chromosomes to be observed (Agar, 1923; Sharman, 1961). First studies using traditional cytogenetic methods based on comparisons by chromosome banding in close-related species, were able to determine approximately 60% of all marsupials karyotypes between the 1960s and 1980s (Hayman & Martin, 1969). These studies also suggested that marsupial chromosomes were largely conserved.

Although there is an extensive cytogenetic framework provided from traditional karyotyping studies of the past several decades, nowadays only a few of marsupial genomes are completely sequenced, and even fewer genomes assembled in chromosome- or HiC-scaffold-levels. This lack of data has hindered the attempt to reconstruct their ancestral karyotypes, and only some studies have combined both, cytogenetic and sequencing data, to trace the evolutionary history of marsupial chromosomes (Meredith et al., 2010; Deakin & O'Neill, 2020).

In this thesis, we want to fill this gap of knowledge, by first reconstructing for the first time the ancestral marsupial karyotype at high resolution, and then establish and time the chromosome rearrangements characterising this lineage. To achieve this aim, we used five marsupial genomes (Tasmanian devil, gray short-tailed opossum, red kangaroo, tammar wallaby, and common wombat) (see **Methods Table 2.1**), to reconstruct two different marsupial ancestral karyotypes: MAK, the ancestral karyotype of all marsupial species, and MAUK, the ancestral karyotype of Australian marsupials (Australidelphia ancestor).

4.2 Results

The aims of this chapter are: (i) to reconstruct the ancestral karyotype of all marsupial species; (ii) to identify the chromosome rearrangements (CRs) that have played a role in the evolution from this ancestral karyotype to current marsupial species; (iii) to identify the evolutionary breakpoint regions (EBRs) and the homologous synteny blocks (HSBs) (conserved and/or rearranged) in Tasmanian devil genome; (iv) to perform the gene ontology enrichment analysis (GOEA) of the genes near and within EBRs; (v) to perform the GOEA of the genes within rearranged and/or conserved HSBs; and (vi) to understand the genomic regions where these EBRs are taking place.

4.2.1 Pairwise alignments

The first step was to perform pairwise alignments of the four marsupial species and the five outgroup species against Tasmanian devil. We selected Tasmanian devil as reference genome because it is a descendant species of both ancestors we aimed to reconstruct, the genome assembly is of very high quality, and, within marsupial species, it is well annotated.

Table 4.1. Coverage of net files and HSBs in the alignments of ruminant species against Tasmanian devil. % nets refers to the coverage of all sequences of non-overlapping gapless blocks. % HSBs refers to the coverage of defined HSBs..

Species	Latin name	Size (Gb)	* % nets	* % HSBs
Gray short-tailed opossum	<i>Monodelphis domestica</i>	3.57	99.7	97.7
Tammar wallaby	<i>Macropus eugenii</i>	3.45	99.52	98.3
Red kangaroo	<i>Macropus rufus</i>	3.6	99.61	98.7
Wombat	<i>Vombatus ursinus</i>	3.34	99.43	98.8

*Coverage against Tasmanian devil genome (size=3.08Gb)

The alignment coverage against Tasmanian devil genome was higher than 99% in all species, thanks to the good quality of the assemblies used (see **Methods Table 2.1**), (**Table 3.1**). Moreover, the coverage of the HSBs produced ranged from 97.7% in gray short-tailed opossum to 98.8% in wombat with an average of 98.4%.

These alignments were then visualised in an Evolution Highway (EH) image format (**Figure 4.2**) to detect CRs specific of Tasmanian devil.

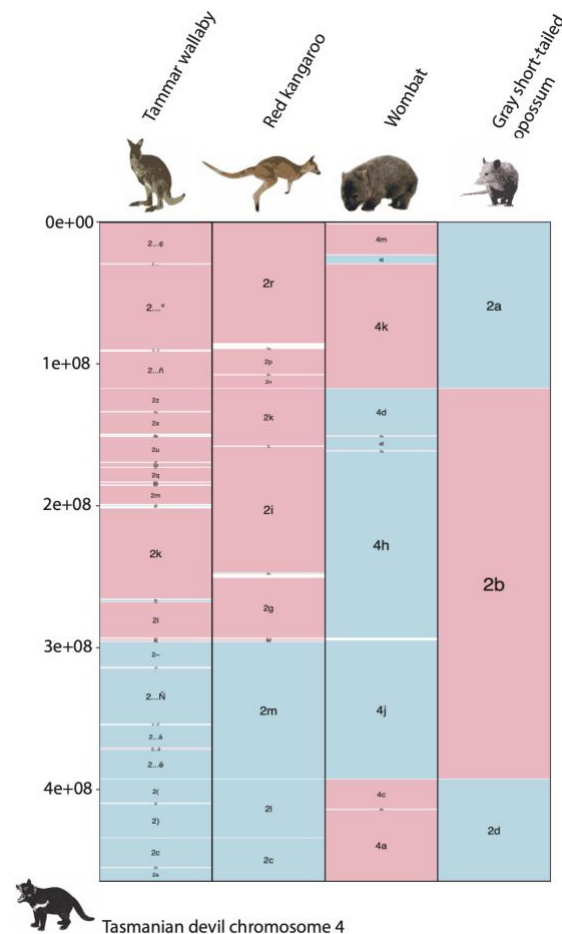


Figure 4.2. EH image of chromosome 4 of Tasmanian devil aligned to four marsupial species. Blue and pink colours indicate same or reverse orientation of the query chromosomes, respectively.

4.2.2 Ancestral karyotype reconstructions

Only five marsupial genomes assembled at chromosome level are publicly available, limiting the number of ancestral karyotype reconstructions that could be generated. Therefore, in this thesis we focused on reconstructing the two main ancestors of the Marsupialia lineage: MAK, the ancestor of all marsupial species, and MAUK, the ancestor of the Australian marsupials (**Figure 4.3**).

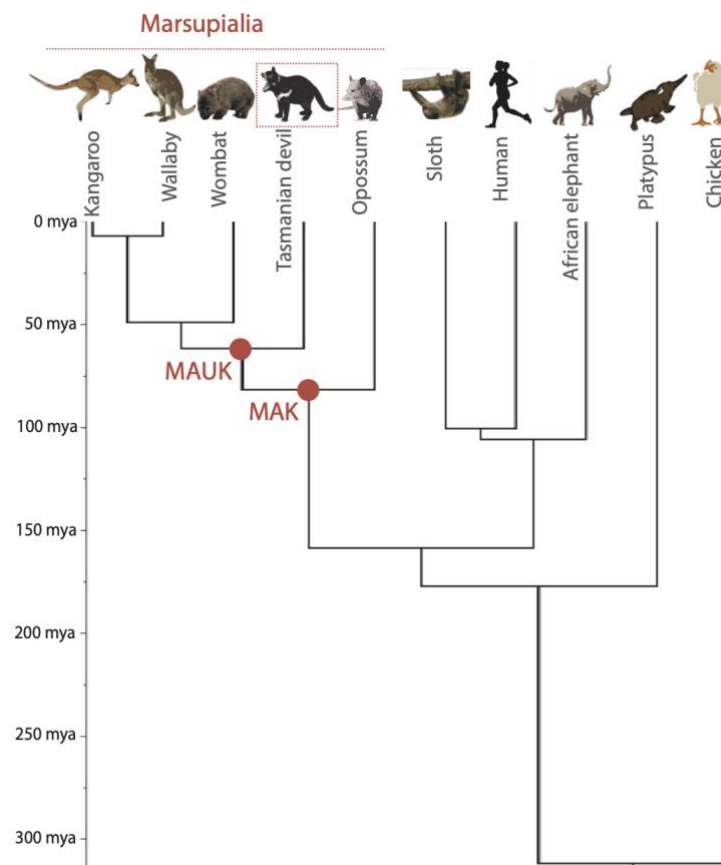


Figure 4.3. Phylogenetic tree of all species used for the marsupial ancestral reconstruction. In red the ancestral nodes reconstructed.

Using DESCHRAMBLER with a resolution of 300 Kb and 0.01 adjacency score, we obtained a raw number of, 14 and 12 RACFs, for MAK and MAUK, respectively. These numbers are higher than previously reported chromosome numbers for the marsupial ancestor (Deakin & O'Neill, 2020), and we proceeded to manually curate them. Because cytogenetic studies focusing on marsupial species always used human-derived painting probes (Deakin & O'Neill, 2020), our manual curation was modified from our original approach, by first aligning our RACFs to the human genome and then comparing their location to cytogenetic studies.

Thereby, comparing MAK aligned to human with this previous reconstruction, we curated the RACFs and reduced their number to 10 and eight for MAK and MAUK, respectively (**Table 4.2**). Although with this curation the maximum RACF size did not change, the minimum size improved significantly from 9,658 Kb to 69,571 Kb in MAK, and from 328 Kb to 71,666 Kb in MAUK.

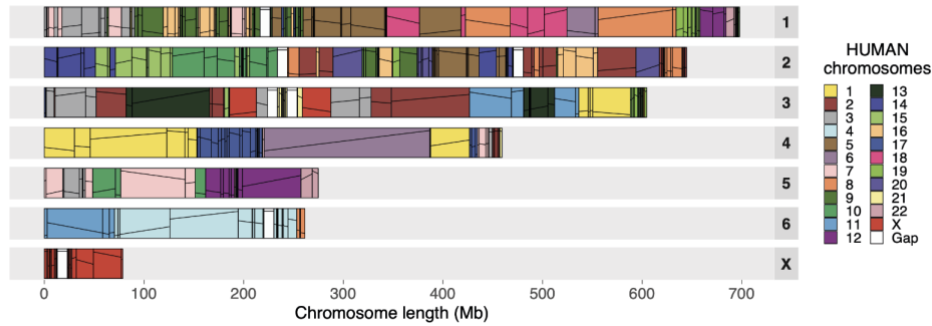
Table 4.2. DESCHRAMBER results. “Mapping number” is the haploid number of chromosomes reported by a previous reconstruction using cytogenetic mapping and computational analysis (Deakin & O’Neill, 2020). The “No. ingroups” indicates the number of marsupial species comprised by each ancestor.

Ancestor	No. RACFs	Mapping number	Max. Size (Kb)	Min. Size (Kb)	Total Size (Kb)	% *	No. ingroups
MAK	14	6+X	690,125	9,658	2,954,102	95.87	5
MAK (curated)	10	6+X	690,125	69,571	2,954,102	95.87	5
MAUK	12	-	698,600	328	3,006,993	97.58	4
MAUK (curated)	8	-	698,600	71,666	3,006,993	97.58	4

*Coverage against Tasmanian devil genome (size=3.08Gb)

Overall, the ten RACFs of MAK corresponded to seven chromosomes, with chromosomes 1 and 2 formed by two and three RACFs, respectively (**Figures 4.4 and 4.5**). Moreover, 85.16% of the syntenic associations between MAK and human were consistent between our reconstruction and the previously published one (Deakin & O’Neill, 2020) (**Figure 4.4**). A total of 14.84% of syntenic associations disagree between our reconstruction and the previous one, probably due to the lower resolution of cytogenetic analysis, where small CRs were undetected (**Figure 4.4**), as such our marsupial ancestor is highly reliable.

A



B

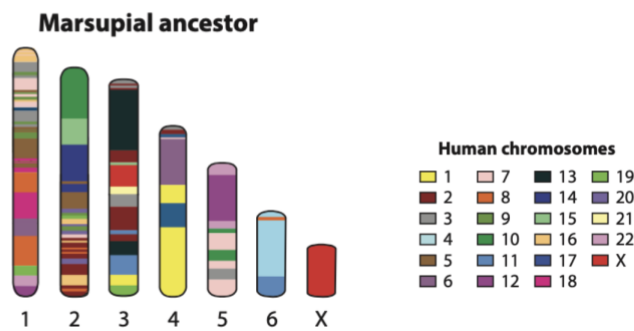


Figure 4.4. Marsupial ancestor comparing to human chromosomes. A. Our marsupial ancestor (MAK) after all RACFs have been manually merged. B. The marsupial ancestor proposed by (Deakin & O'Neill, 2020) and used as a template for our manually curation.

In the case of MAUK, the ancestor of Australian marsupials, the merging was performed in a similar way, without the reference of the ancestor against the human genome. We found that MAUK has the same haploid number of chromosomes as MAK, 6+X chromosomes, with MAUK chromosome 2 split into two different RACFs. Furthermore, in terms of coverage compared to Tasmanian devil genome, MAK and MAUK reconstructions achieved 95.87% and 97.58% respectively, and as expected, the coverage is higher in the closest ancestor to Tasmanian devil (**Table 4.1**).

To identify CRs shaping marsupial genome evolution we cross-compared the reconstructed ancestral karyotypes and the reference species. We focused on determining the CRs between MAK and MAUK, and MAUK and Tasmanian devil (**Figure 4.5**). A total of 14 CRs differentiated MAUK from MAK, with ten inversions, two fissions, and two fusions. These results showed that there are more intrachromosomal rearrangements than interchromosomal rearrangements, but interchromosomal rearrangements have played an important role

modifying chromosome size. For instance, chromosome 1 in MAK has reduced its size in MAUK by a non-reciprocal translocation to chromosome 2 in MAUK. Finally, only three chromosomes were maintained as full HSBs, MAK chromosomes 4, 5, and X (**Figure 4.5**).

Comparing MAUK to Tasmanian devil, the number of CRs is clearly higher than in the comparison between the two ancestors, and not a single chromosome is conserved between MAUK and Tasmanian devil. A total of 25 CRs differentiated MAUK from Tasmanian devil, one fission and 24 inversions. These results highlight inversions as the main CR type leading the marsupial evolution, especially from the Australian ancestor to the current species.

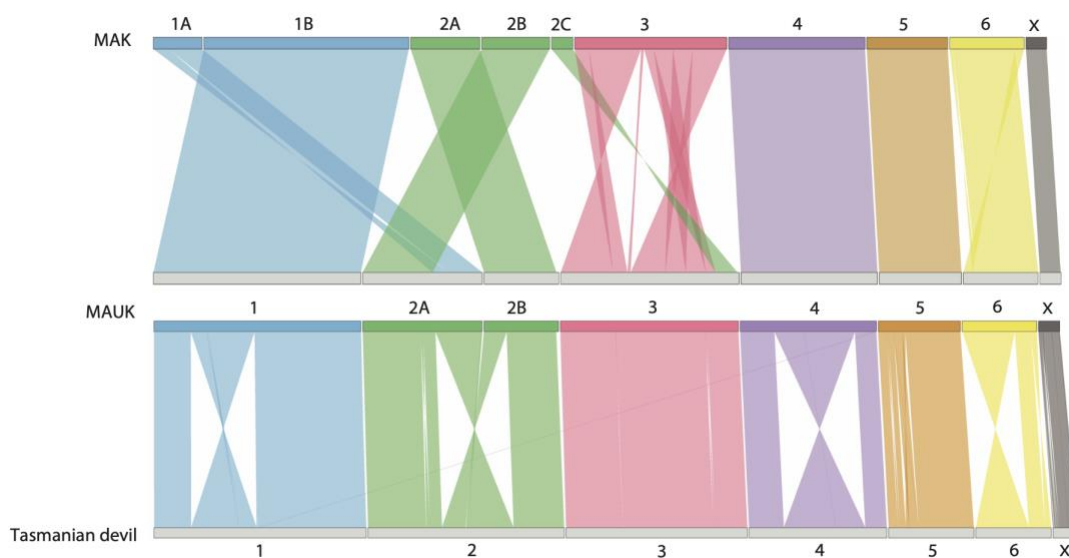


Figure 4.5. Pairwise comparison between MAK, MAUK, and Tasmanian devil.

Comparing both ancestors to Tasmanian devil, it is clear that CRs have played an important role after the radiation of American and Australian marsupials in reshuffling the current marsupial genomes. Moreover, the interchromosomal rearrangement rates of evolution per million years are 0.29 and 0.01 in MAK and MAUK, respectively, reinforcing that interchromosomal rearrangements played a limited role in the genome reorganization in marsupials. Instead, intrachromosomal rearrangements rates of evolution per million years are 0.71 and 0.36 in MAK and MAUK, respectively. Overall, the evolution from MAK to Tasmanian devil has a rate of 1 CR/My, while MAUK to Tasmanian devil has a rate of 0.38 CRs/My (**Table 4.3**).

Altogether, these results indicate, that although the number of rearrangements is lower in the MAK to MAUK evolution, they have occurred faster, while the branch from MAUK to Tasmanian devil genome showed more CRs that have occurred during a longer period of time.

Table 4.3. Summary table of type of CR differentiating every ancestor from Tasmanian devil (TD).

No marsupial species used	5	
No non-marsupial species used	5	
No ancestral genomes reconstructed	2	
	MAK-TD	MAUK-TD
No of CRs	27 inversions	24 inversions
	0 fusions	0 fusions
	2 fissions	1 fission
Rearrangement rate (per my)	1	0.38

4.2.3 Detection and study of evolutionary breakpoint regions (EBRs)

Using the alignment files of Tasmanian devil with MAK and MAUK, EBRs and HSBs were detected and classified in the Tasmanian devil genome. We found a total of 34 EBRs classified in 19 well-defined EBRs (when their size was between 0 and 50 Kb) and 15 not well-defined EBRs (when their size was between 50 and 300 Kb), while 18 were considered gaps. Within the well-defined EBRs, eight were MAK-specific and 11 were Tasmanian devil-specific, and within the not well-defined EBRs, three were MAK-specific and 12 were Tasmanian devil-specific. This classification (**Table 4.4**) reports more EBRs and therefore, more CRs specific from Tasmanian devil than from MAK.

Table 4.4. Classification of Tasmanian devil EBRs.

Chromosome	Start	End	Classification	Type
1	246,754,420	246,754,420	MAK specific	Well-defined EBRs
1	124,965,782	124,974,583	TD specific	Well-defined EBRs
1	346,403,024	346,417,623	TD specific	Well-defined EBRs
1	349,030,717	349,069,400	TD specific	Well-defined EBRs
1	287,296,606	287,335,937	TD specific	Well-defined EBRs
2	113,406,832	113,406,832	MAK specific	Well-defined EBRs
2	168,164,599	168,164,599	MAK specific	Well-defined EBRs
2	209,525,262	209,525,262	TD specific	Well-defined EBRs

2	491,602,898	491,603,004	TD specific	Well-defined EBRs
2	249,485,582	249,488,641	TD specific	Well-defined EBRs
3	229,575,555	229,575,555	MAK specific	Well-defined EBRs
3	240,314,590	240,314,590	MAK specific	Well-defined EBRs
3	357,434,234	357,434,234	MAK specific	Well-defined EBRs
3	427,627,554	427,627,554	MAK specific	Well-defined EBRs
3	530,270,065	530,270,065	MAK specific	Well-defined EBRs
4	392,616,908	392,617,031	TD specific	Well-defined EBRs
4	117,275,408	117,275,630	TD specific	Well-defined EBRs
5	69,043,621	690,58,326	TD specific	Well-defined EBRs
6	180,056,731	180,071,083	TD specific	Well-defined EBRs
1	349,771,217	349,825,269	TD specific	Not well-defined EBRs
3	189,084,773	189,380,950	TD specific	Not well-defined EBRs
3	233,885,942	234,167,636	MAK specific	Not well-defined EBRs
3	503,055,880	503,128,571	TD specific	Not well-defined EBRs
4	293,055,739	293,115,640	TD specific	Not well-defined EBRs
5	2,313,294	2,399,468	TD specific	Not well-defined EBRs
5	57,532,766	57,750,123	TD specific	Not well-defined EBRs
5	101,841,143	101,920,464	TD specific	Not well-defined EBRs
6	136,435,007	136,570,263	MAK specific	Not well-defined EBRs
6	148,571,509	148,627,793	MAK specific	Not well-defined EBRs
6	233,709,463	233,882,534	TD specific	Not well-defined EBRs
X	1,406,978	1,683,726	TD specific	Not well-defined EBRs
X	53,596,421	53,714,762	TD specific	Not well-defined EBRs
X	79,482,820	79,757,310	TD specific	Not well-defined EBRs
X	80,088,726	80,170,948	TD specific	Not well-defined EBRs

A correlation between the number of EBRs and chromosome size is observed for the well-defined EBRs (**Figure 4.6A**), with bigger chromosomes containing more EBRs and therefore, showing a more reorganized configuration. The highest number of EBRs, five, was found in chromosomes 1, 2, and 3. However, because chromosome X is poorly assembled, all EBRs detected within this chromosome were either not well-defined or gaps. Contrariwise, when we expand the EBR selection and study both, well-defined and not well-defined EBRs together (**Figure 4.6B**), no correlation is found between number of EBRs and chromosome size (P-value = 0.156). In this scenario, chromosome 3 is the one containing more EBRs (8), chromosome X contains 4 not well-defined EBRs and chromosome 4 is the one containing less EBRs (2).

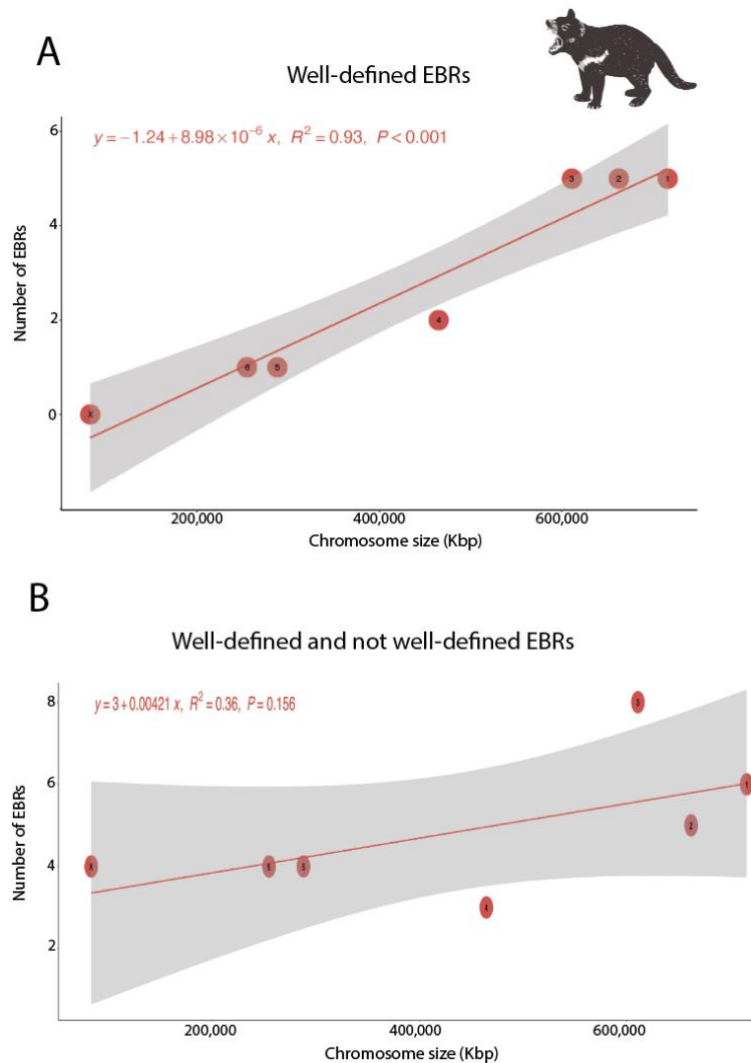


Figure 4.6. Correlation between the number of well-defined (A) and well-defined and not well-defined (B) EBRs and Tasmanian devil chromosome sizes in Kb.

To determine whether EBRs occur in non-random locations of the genome, we first studied their co-location with gene-rich regions. For this purpose, several permutation tests were performed following the same three associations used in the previous chapter: (i) regions within well-defined EBRs; (ii) regions within not well-defined EBRs; and (iii) regions near EBRs (**Figure 4.7**). A clear negative association (Z-score = -3.229 and P-value = 0.003) between well-defined EBRs and Tasmanian devil genes was found (**Figure 4.7A**), and furthermore, regions near these EBRs showed a clear pattern of more gene content than within EBRs (P-value in the 20 Kb downstream region from well-defined EBRs = 0.001 and P-value in the 20 kb upstream region from well-defined EBRs = 0.0009) (**Figure 4.7C**). As previously reported for other species (M. Farré et al., 2019), when EBRs are well-defined sometimes their sizes are

0 or 1 bp, and it is more likely that these well-defined EBRs are surrounded by gene-rich regions than being embedded within a gene. For not well-defined EBRs (**Figure 4.7B**), no statistically significant association was found.

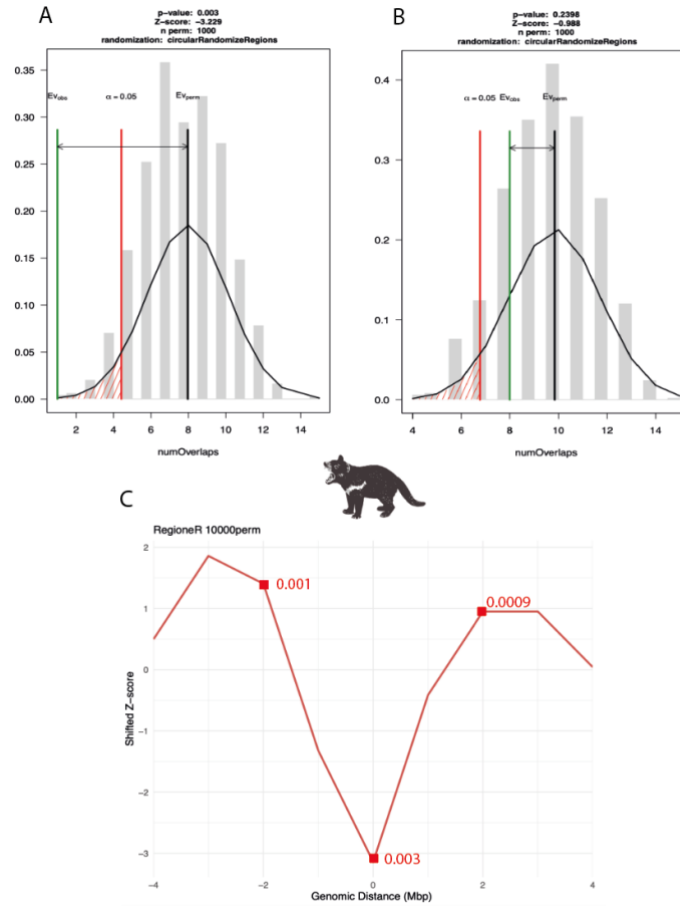


Figure 4.7. Permutation tests performed to establish the correlation between gene content and well-defined EBRs (A), not well-defined EBRs (B), and regions near well-defined EBRs (C) in Tasmanian devil. In C, red number show P-values in the regions marked with a red square.

We then determined the type of genes located within or near EBRs by performing GOEA. For the well-defined EBRs, only one gene was present, *ENSSHAG00000002711*, leading to no significant GO terms. We then expanded all the well-defined EBRs, up- and downstream, until all of them reached 50 Kb in size. After the expansion, nine genes were found but they were unannotated (**Table S2**), which means that no significant GO terms were detected either. On the other hand, when we performed GOEA in the not well-defined EBRs, only 11 genes were included (**Table S3**), and again leading to the same absence of significant GO terms.

Finally, we assessed whether EBRs were in regions with high density of repetitive elements. Neither type of EBRs (well-defined and not well-defined) were associated to repetitive elements in Tasmanian devil.

4.2.4 Detection and study of homologous syntenic blocks (HSBs)

After locating the EBRs, we defined HSBs. These HSBs were classified depending on the type of rearrangement originating them in collinear (when they were conserved between species), inverted, or fissioned HSBs.

In Tasmanian devil, we classified these HSBs according to the rearrangements occurring during the evolution from the ancestor of all marsupial species (MAK) to the current Tasmanian devil genome (**Table 4.5**). A total of 65.67% and 33.58% of Tasmanian devil genome corresponded to collinear and inverted HSBs, respectively. This result remarks the importance of inversions (intrachromosomal rearrangements) in the genome reshuffling occurred during the evolution of the species belonging to the Marsupialia lineage.

Table 4.5. Classification of HSBs in Tasmanian devil genome.

Chromosome	Start	End	Classification
1	41,3176	124,965,782	Collinear
1	124,974,583	287,296,606	Inversion
1	287,335,937	292,600,914	Collinear
1	293,539,669	346,403,024	Inversion
1	346,417,623	349,030,717	Collinear
1	349,069,400	349,771,217	Fission
1	349,825,269	713,935,017	Collinear
1	714,318,760	716,396,363	Inversion
2	345,467	196,055,892	Collinear
2	197,019,960	209,525,262	Inversion
2	209,525,262	249,485,582	Collinear
2	249,488,641	328,841,166	Inversion
2	330,995,254	332,886,993	Inversion
2	333,239,788	336,185,894	Collinear
2	336,554,472	491,602,898	Inversion
2	491,603,004	661,803,265	Collinear
3	85,028	189,084,773	Collinear

3	189,380,950	190,913,832	Inversion
3	191,794,259	495,866,272	Collinear
3	496,230,356	503,055,880	Inversion
3	503,128,571	611,289,136	Collinear
4	90,027	117,275,408	Collinear
4	117,275,630	293,055,739	Inversion
4	293,115,640	392,616,908	Inversion
4	392,617,031	464,644,381	Collinear
5	846,972	2,313,294	Inversion
5	2,399,468	30,813,396	Collinear
5	31,657,846	38,936,826	Inversion
5	39,589,839	57,532,766	Collinear
5	57,750,123	69,043,621	Inversion
5	69,058,326	100,565,821	Collinear
5	101,530,180	101,841,143	Inversion
5	101,920,464	287,795,671	Collinear
6	198,523	180,056,731	Inversion
6	180,071,083	233,709,463	Inversion
6	233,882,534	254,889,130	Inversion
X	31,423	1,406,978	Inversion
X	1,683,726	2,179,691	Collinear
X	4,783,144	10,186,904	Collinear
X	11,686,924	14,107,410	Inversion
X	14,742,900	15,559,472	Collinear
X	15,901,208	16,228,925	Inversion
X	17,563,162	22,991,248	Inversion
X	23,681,581	52,650,129	Collinear
X	53,096,430	53,596,421	Inversion
X	53,714,762	79,482,820	Collinear
X	79,757,310	80,088,726	Inversion
X	80,170,948	82,930,477	Collinear

We then performed GOEA in both collinear and inverted HSBs. The 6,036 genes comprised within inverted HSBs showed GO terms related to signalling receptor activity, response to stimulus, organelles, cellular anatomical entity, or major histocompatibility complex (MHC) protein complex. Additionally, the 11,077 genes in collinear HSBs were related to developmental processes, different metabolic processes or with binding (**Figure 4.8**). Moreover, a small fraction of Tasmanian devil genes was unannotated, representing the 37.03%

of genes reported within collinear HSBs, reflecting the necessity of a better annotation of this genome.

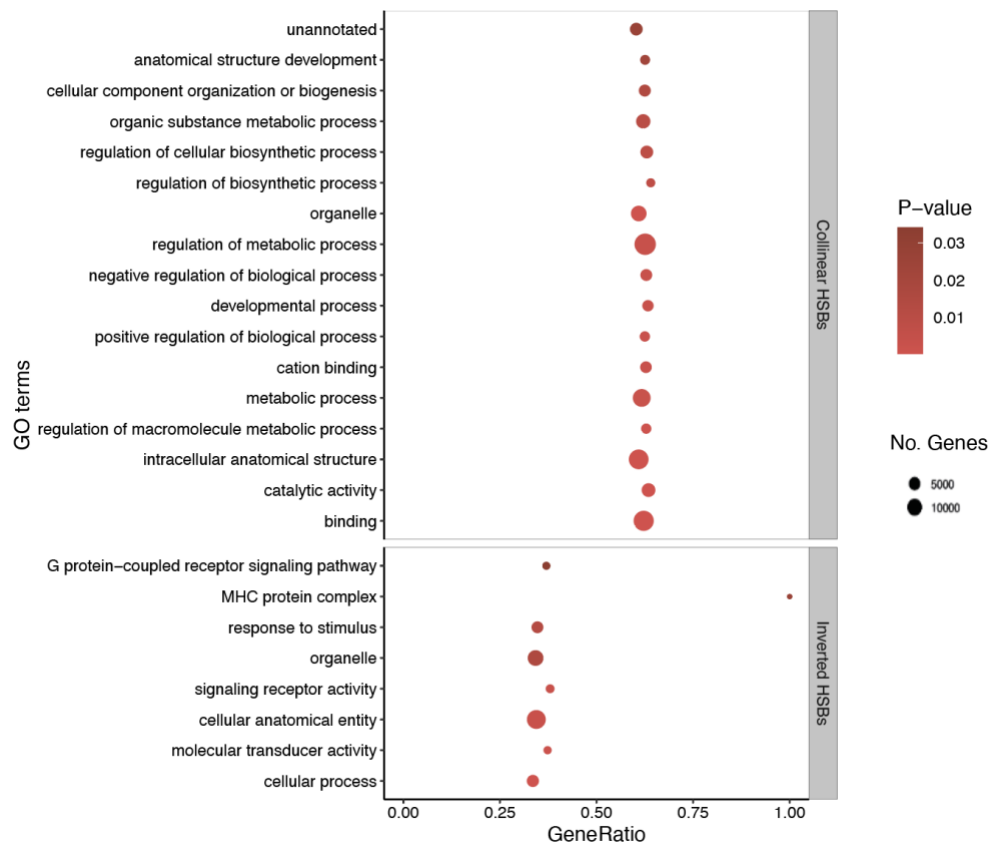


Figure 4.8. GOEA of Tasmanian devil HSBs according to their classification. X axis represents the gene ratio (number of genes/total number of genes). Y axis represents the GO terms. Bubbles size varies depending on their number of genes in every GO term (No. Genes). Bubbles colour varies depending on their P-value.

4.3 Discussion

In this chapter, we proposed a new karyotype for two different marsupial ancestors: the ancestor comprising all marsupial species (MAK) and the ancestor comprising only the Australian marsupial species (MAUK). Here we report for the first time the MAUK ancestor, key to time CRs during the marsupial evolution. This intermediate ancestor is also essential to determine whether the karyotype configurations of current Australian species are similar to their closer ancestor (MAUK), indicating that the majority of CRs have occurred in their evolution from MAK, or whether they are different to MAUK, indicating the main CR episodes occurred later on time. Previous studies (Deakin & O'Neill, 2020; Damas et al., 2022) have only proposed the ancestral configuration of the ancestor of all marsupials, leading to an

overestimation of the number of rearrangements between this ancestor and current marsupial species. In this thesis, by reconstructing a second ancestor closer to current marsupial species, MAUK, we clearly differentiate both, the CRs which have affected genome organization in marsupials when American and Australian marsupials diverged (~ 66 Mya) and the CRs that separate current species from the Australian ancestor. Our results show a bigger number of CRs in the evolution from MAUK to Tasmanian devil, indicating that both the American and the Australian ancestors, are more similar to each other, and that many of the CRs responsible of the evolution of marsupials' genomes have occurred after the divergence of both ancestors. Furthermore, we were also able to identify 14.84% more inversions that due to their small sizes were not previously detected by cytogenetic methodologies. Moreover, a comparison between both ancestors determined the fusion of MAK chromosomes 1 and 2, and 2 and 3, in MAUK chromosomes 2 and 3, respectively. Finally, in terms of chromosome diploid number, MAK and MAUK contain the same number than the previous marsupial ancestral karyotype, $2n = 14$ (Deakin & O'Neill, 2020).

Overall, our results clearly indicate that inversions (intrachromosomal rearrangements) are the main event leading marsupial evolution. Taking these results together with previous studies showing the important role that inversions have played during speciation (Kirkpatrick, 2010; Faria et al., 2019; Dagilis, 2022), we can confirm that inversions are the main type of CRs driving to marsupial speciation.

The high resolution of our reconstruction allowed us to define the EBRs along Tasmanian devil genome and to classify them depending on whether they were ancestor-specific or Tasmanian devil-specific. Our findings agree with previous studies that associate these EBRs with rich-gene regions in the genome (Murphy et al., 2005). However, we were unable to find GO terms for those genes within or near well-defined EBRs. This could be related to the lack of a good annotation for Tasmanian devil genome, as many genes are located but their functions are not known. Within the known genes found within well-defined EBRs, ENSHAG00000021067 (*MTF2*) is related to the metal response element binding transcription factor II, which has been reported to be sufficient to maintain the global chromatin configuration in mammalian cells, together with CTCF, and cohesin (Hsieh et al., 2020). Within the genes found in the not well-defined EBRs, *ERCC6L* (excision repair cross-complementation group 6 like), described in other species as upregulated gene in a variety of malignant tumours, plays a critical oncogenic role (Lu et al., 2022; Huang et al., 2022). Another

gene found in the not well-defined EBRs is *CCNA1* (cyclin A1), also associated to the development and progressions of several tumours (Cao et al., 2019.; El Aliani et al., 2021). Overall, many of the genes found within EBRs are related to tumour environments, which are known to be environments producing high genome instability, which potentially can promote the origin of CRs. The presence of genes related to either chromatin configuration or DSBs repair such as *MFT2* within the EBRs, could be related to structural plasticity detected in lineage-specific chromosomal reorganizations, specifically, could be related to the marsupial-specific anchoring of the centromeres toward the centre of the nucleus favouring intrachromosomal rearrangements (Álvarez-González, Arias-Sardá, et al., 2022).

Additionally, even though repetitive elements have been associated with genomic rearrangements that could be also associated with genome speciation events in different vertebrate lineages (Böhne et al., 2008; M. Farré et al., 2019; Damas et al., 2022), no association between repetitive elements and EBRs in marsupials was found. This can be due to two different facts: (i) the lack of a good TE annotation for the Tasmanian devil genome; or (ii) the actual fact that marsupial EBRs are not enriched in TEs. For future work, it will be necessary to perform a high-quality annotation of Tasmanian devil genome and repeat the permutation analysis to test whether there is an enrichment of TEs in marsupial EBRs.

Finally, we were able to identify and classify the HSBs in the Tasmanian devil genome, and also to describe the GO terms associated to the specific rearrangements causing these HSBs. Surprisingly, many of the genes enriched in both, collinear and inverted HSBs, are oncogenes or are related to tumour environments. Some of the examples we found in collinear HSBs are related to metabolic and biological functions, such as: (i) *RIOK1* (RIO Kinase 1) (Y. Chen et al., 2022); (ii) *XRCC5* (X-ray repair cross-complementing protein 5) (Zhang et al., 2017; Gu et al., 2018; AL-Eitan et al., 2019); (iii) the transcription factor *FOXA2* (Z. Yang et al., 2021; Han et al., 2022); or (iv) *GATA5* (GATA binding protein 5), which acts as tumoral suppressor (P. Liu et al., 2018; Feng et al., 2019; Q. Wang et al., 2022). Additionally, in inverted HSBs, the GO term with the highest gene ratio was the MHC protein complex (major histocompatibility complex), whose transcriptional repression has been observed in a range of cancers, and has recently been identified as a mechanism of resistance to immunotherapy (Burr et al., 2019; Yamamoto et al., 2020; Baeuerle & Wesche, 2022). Interestingly, the specific tumour associated to Tasmanian devil, the devil facial tumour disease (DFTD), which is caused by two genetically independent transmissible cancers that are spread through social interactions, the

devil facial tumour 1 and 2 (DFT1 and DFT2), has been associated to the downregulation of MCH-I in DFT1 cells but not in DFT2 cells (Kayigwe et al., 2022). Overall, we could not find any enrichment in genes involved in processes related to chromatin conformation or condensation, but we found a very high association of EBRs and HSBs in marsupial species to genes related to cancer, which is an indicative of the genome instability associated to bigger chromosome sizes and bigger tendencies to suffer CRs, and more specifically, related to the DFTD specific of Tasmanian devil. For future work, it would be very interesting to further study these genes related to cancer and, moreover, to try to find any connection with them and the 3D chromatin structure.



This chapter reports the results from:

Álvarez-González, L., Arias-Sardá, C., Montes-Espuña, L., Marín-Gual, L., Vara, C., Lister, N. C., Cuartero, Y., Garcia, F., Deakin, J., Renfree, M. B., Robinson, T. J., Martí-Renom, M. A., Waters, P. D., Farré, M., & Ruiz-Herrera, A. (2022). Principles of 3D chromosome folding and evolutionary genome reshuffling in mammals. *Cell Reports*, 41(12), 111839.

5. The Afrotheria lineage

5.1 Introduction

Afrotherians have been considered one of the most ancient placental mammals (Frönicke et al., 2003). This superorder is formed by 90 species, classified in 35 genera within nine families (Chrysochloridae, Dugongidae, Elephantidae, Macroscelididae, Orycteropodidae, Potamogalidae, Procaviidae, Tenrecidae, and Trichechidae), and belonging to six orders (Afrosoricida, Hyracoidea, Macroscelidea, Proboscidea, Sirenia, and Tubulidentata) (Burgin et al., 2018; Mammal Diversity Database, 2022).

The Afrotheria lineage origin seems to be entirely Afro-Arabian. Tectonic movement isolated Afro-Arabia as an "island continent" until the late Cenozoic, when intercontinental dispersal became a possibility (F. Yang et al., 2003). The extreme morphological diversity and niche preference of afrotherian species is thought to result from this long period of isolation ~ 105-125 Mya (Meredith et al., 2011). However, fossil records suggest that the afrotherians did not originate in Africa, on the contrary, they would have had to disperse either from North to South America, or the reverse, in the early Cretaceous around 120 Mya, and then to Africa (Hedges, 2001; O'Leary et al., 2013; Couvreur et al., 2021).

Although Afrotheria has been favoured as the first descendant in the Eutherian tree, another lineage, the Xenarthra, including sloths, anteaters, and armadillos, emerged around 100 Mya within a short interval from Afrotheria (Eizirik, 2001; Delsuc et al., 2004; Svartman et al., 2006). The phylogenetic relationships within Afrotheria orders has been controversial for many years, but it has now reach the consensus claiming that the divergence within Afrosoricida, Macroscelidea, Paenungulata, and Tubulidentata might have occurred very rapidly in the late Cretaceous period (Seiffert, 2007). There is also anatomical and fossil evidence suggesting a close relationship of proboscideans, sirenians, and hyracoids (Novacek, 1992). And aardvark was finally placed within the lineage (Douady et al., 2002) (**Figure 5.1**).

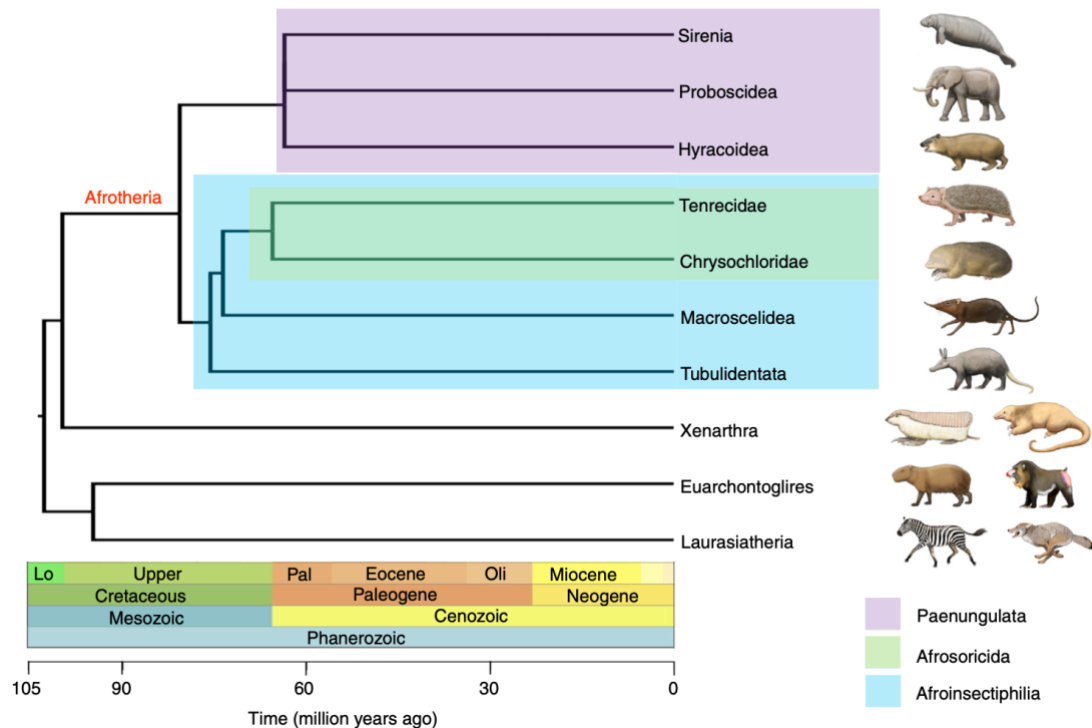


Figure 5.1. Phylogeny of afrotherians among the Geological Time Scale (GTM) of Earth. Figure from (Springer, 2022).

This diverse group of animals show distinct differences in size, preferred habitats, and diets. One of the most remarkable characteristics of this superorder is that there are both marine and terrestrial species within it. The afrotherian order Sirenia, which includes only four living species (the dugong and three manatees), is the only one containing fully aquatic mammals that are also obligate herbivores whose habitats are predominantly warm, near-shore, and relatively shallow waters that accommodate the vegetative growth that comprises their diets (Marsh, 2014; O'Shea, 2014). The fossil record of sirenians begins near the end of the middle Eocene, about 48 Mya, and since that time their evolving diversity and biogeography has ensued to leave fossil occurrences on every continent except Antarctica (Heritage & Seiffert, 2022). These fossils showed a very different sirenian morphology, while nowadays sirenians are equipped with several anatomical adaptations that suit them to wholly aquatic lifeways, such as the lacking of hindlimbs, the earliest sirenians were, however, quadrupedal and amphibious (Domning, 2001; Heritage & Seiffert, 2022).

Within the Afrotheria lineage, the order Proboscidea, including elephants, has been very well described and also has the best fossil record of any afrotherian group, including recently extinct mammoths and mastodons (Callaway, 2021). Elephants have a unique morphology

characterized for structures such as their trunk and tusk and lungs adapted for snorkelling, or well-developed senses of olfaction, hearing, and touch (Shoshani & Tassy, 1996; Sukumar, 2004). Their long lifespans may be of interest for research on longevity, and they are subject to diseases that have analogues in livestock or humans (Roca & O'Brien, 2005). Elephants also have large brains, which enable them to exhibit complex systems of communication and social interaction, and relatively high levels of learning and memory (Blanc et al., 2003; Roca & O'Brien, 2005).

Even though their differences, afrotherians also present derived anatomical characters that are shared by all or most of its members. These include an increase in the number of thoracolumbar vertebrae from 19 to 20 or more, modifications to the allantois (one of the foetal membranes), a deep depression on one of the ankle bones that articulates with the distal end of the shin bone, and a delayed eruption of the permanent dentition (Carter et al., 2004; Springer, 2022).

Although these species bear little resemblance to each other, molecular data offers evidences for their evolutionary relationships (Stanhope et al., 1998). Traditional comparative genomic methods, based on painting probes and FISH, have been performed in different afrotherian species to determine their ancestral karyotype (Frönicke et al., 2003; T. Robinson & Seiffert, 2004; T. J. Robinson et al., 2004). Extending comparative chromosome painting analyses to the remaining Afrotheria and the Xenarthra, facilitated the transfer of mapping data among evolutionarily highly divergent species and contributed to the development of a cohesive, chromosomal character-based phylogenetic framework for all eutherian mammals (F. Yang et al., 2003).

Genome organization within Afrotheria is diverse, with diploid numbers ranging from $2n = 54$ in the African (*Loxodonta africana*) and Asian (*Elephas maximus*) elephants to $2n = 20$ in the armadillo (*Oryzomys azer*) (T. Robinson & Seiffert, 2004). Some studies suggested that relatively little interchromosomal repatterning has occurred subsequent to divergence from the afrotherian last common ancestor, further emphasizing the slow rate of chromosomal evolution in the low-chromosome numbered afrotheria (T. J. Robinson et al., 2004), while other studies emphasised the presence of interchromosomal rearrangements within different afrotherian families (Pardini et al., 2007). When looking at genome composition, some studies have reported endogenous repetitive elements and remarkable cytonuclear genomic patterns in

elephants (Roca & O'Brien, 2005), and other studies based on comparing afrotherian proteomes have been useful to understand afrotherian genome evolution (Yazhini et al., 2021).

Given the position of afrotheria and marsupials near the root of therian mammals and their diverse diploid numbers (reflecting extensive genome reshuffling), the analysis of 3D genome architecture provides a unique opportunity to further understand the mechanisms underpinning mammalian chromosomal evolution (Álvarez-González, Arias-Sardá, et al., 2022).

For the purpose of assessing the functional contribution of chromosomal rearrangements to the evolution of afrotheria, we reconstructed the ancestral karyotype of all the afrotherian species, AFAK, using seven afrotherian genomes: Asian elephant, African elephant, armadillo, cape golden mole, lesser hedgehog tenrec, cape rock hyrax, and West Indian manatee.

5.2 Results

The aims of this chapter are: (i) to reconstruct the ancestral karyotype of all the afrotherian species; (ii) to identify the chromosome rearrangements (CRs) that have played a role in the evolution from this ancestral karyotype to current afrotherian species; (iii) to identify the evolutionary breakpoint regions (EBRs) and the homologous synteny blocks (HSBs) (conserved and/or rearranged) in African elephant genome; (iv) to perform the gene ontology enrichment analysis (GOEA) of the genes near and within EBRs; (v) to perform the GOEA of the genes within rearranged and/or conserved HSBs; and (vi) to study the presence of transposable elements (TEs) near and within EBRs.

5.2.1 Pairwise alignments

The first step was to perform pairwise alignments of the six afrotherian species and the three outgroup species against African elephant. We selected the African elephant as reference genome for several reasons, such as the quality of its assembly, or the availability of genomic information like genes and TEs annotations.

Table 5.1. Coverage of net files and HSBs in the alignments of ruminant species against African elephant. % nets refers to the coverage of all sequences of non-overlapping gapless blocks. % HSBs refers to the coverage of defined HSBs.

Species	Latin name	Size (Gb)	% nets	% HSBs
Cape golden mole	<i>Chrysochloris asiatica</i>	4.21	97.77	94.7
Asian elephant	<i>Elephas maximus</i>	3.21	99.75	99.51
Aardvark	<i>Orycteropus afer</i>	4.42	99.6	97.9
Cape rock hyrax	<i>Procavia capensis</i>	3.61	98.96	93.9
Tenrec	<i>Echinops telfairi</i>	2.95	97.64	95.59
Indian manatee	<i>Trichechus manatus</i>	3.1	99.9	98.6

*Coverage against African elephant genome (size=3.14Gb)

Thanks to the good quality of the assemblies used (see **Methods Table 2.1**), the alignment coverage against African elephant genome ranged from 97% in tenrec and cape golden mole to 99.9% in the Indian manatee (**Table 3.1**). Moreover, the coverage of the HSBs produced ranged from 93.9% in cape rock hyrax to 99.51% in Asian elephant, with an average coverage of 96.7%.

These alignments were then visualised in an Evolution Highway (EH) image format (**Figure 5.2**) to detect CRs not only specific of African elephant, but also specific of the both elephants, so called Proboscidea-specific.

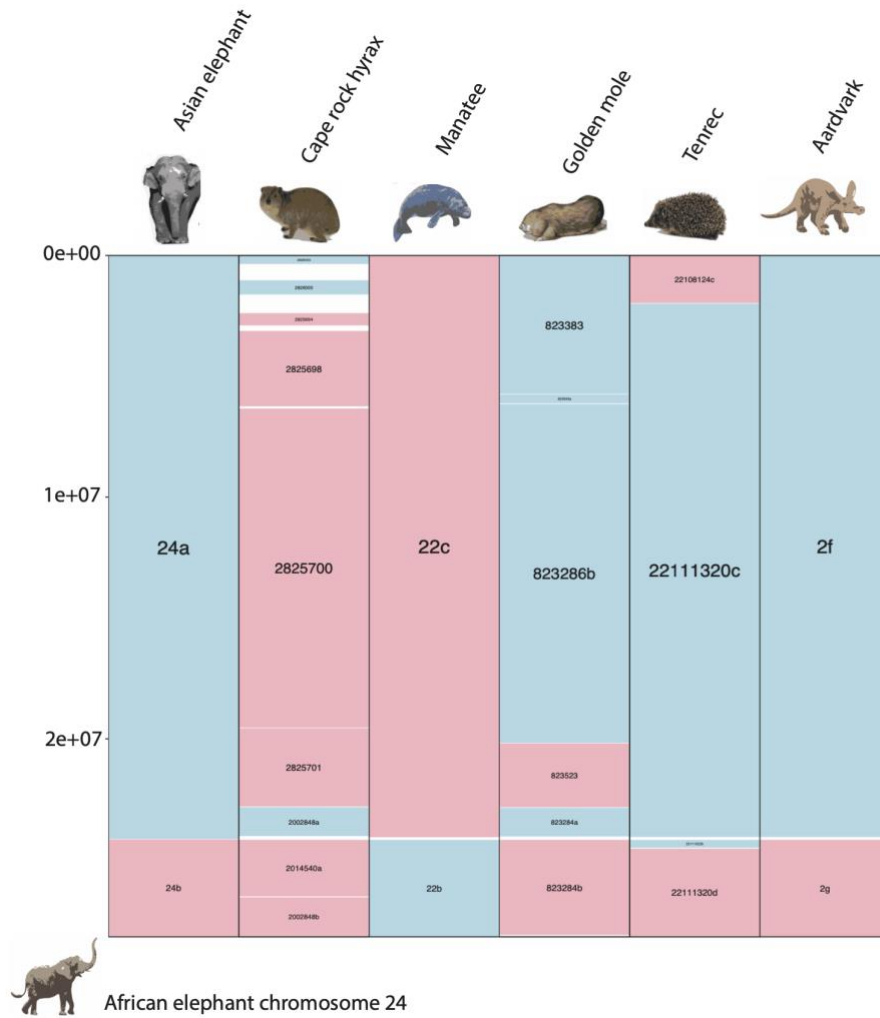


Figure 5.2. EH image of chromosome 24 of African elephant aligned to six afrotherian species. Blue and pink colours indicate same or reverse orientation of the query chromosomes or fragments, respectively.

Overall, looking to the EH images chromosome by chromosome, we can detect African elephant-specific rearrangements, Proboscidea-specific rearrangements, and conserved HSBs among species. However, the best approach to study these CRs that have occurred during the afrotherian evolution, is to reconstruct their ancestral karyotype and compare it with current afrotherian species.

5.2.2 Ancestral karyotype reconstructions

Due to the reduced availability of high-resolution genome assemblies for afrotherian species, only the ancestor of all afrotherian species was reconstructed: AFAK (**Figure 5.3**). Surprisingly, no afrotherian genomes assembled at chromosome level are publicly available,

and for this thesis, we selected the only six species assembled in HiC-scaffold level: both elephants, Indian manatee, cape rock hyrax, golden mole, and aardvark (see **Methods Table 2.1**). Because we only used HiC-scaffold level assemblies, we reconstructed only the ancestor of all afrotherian species (AFAK). In this way, we avoided the possible introduction of unreliable syntenic associations and lack of resolution of EBRs. Moreover, AFAK is the only ancestor which has been studied before (Frönicke et al., 2003) with cytogenetic data that can be used for the posterior curation of RACFs.

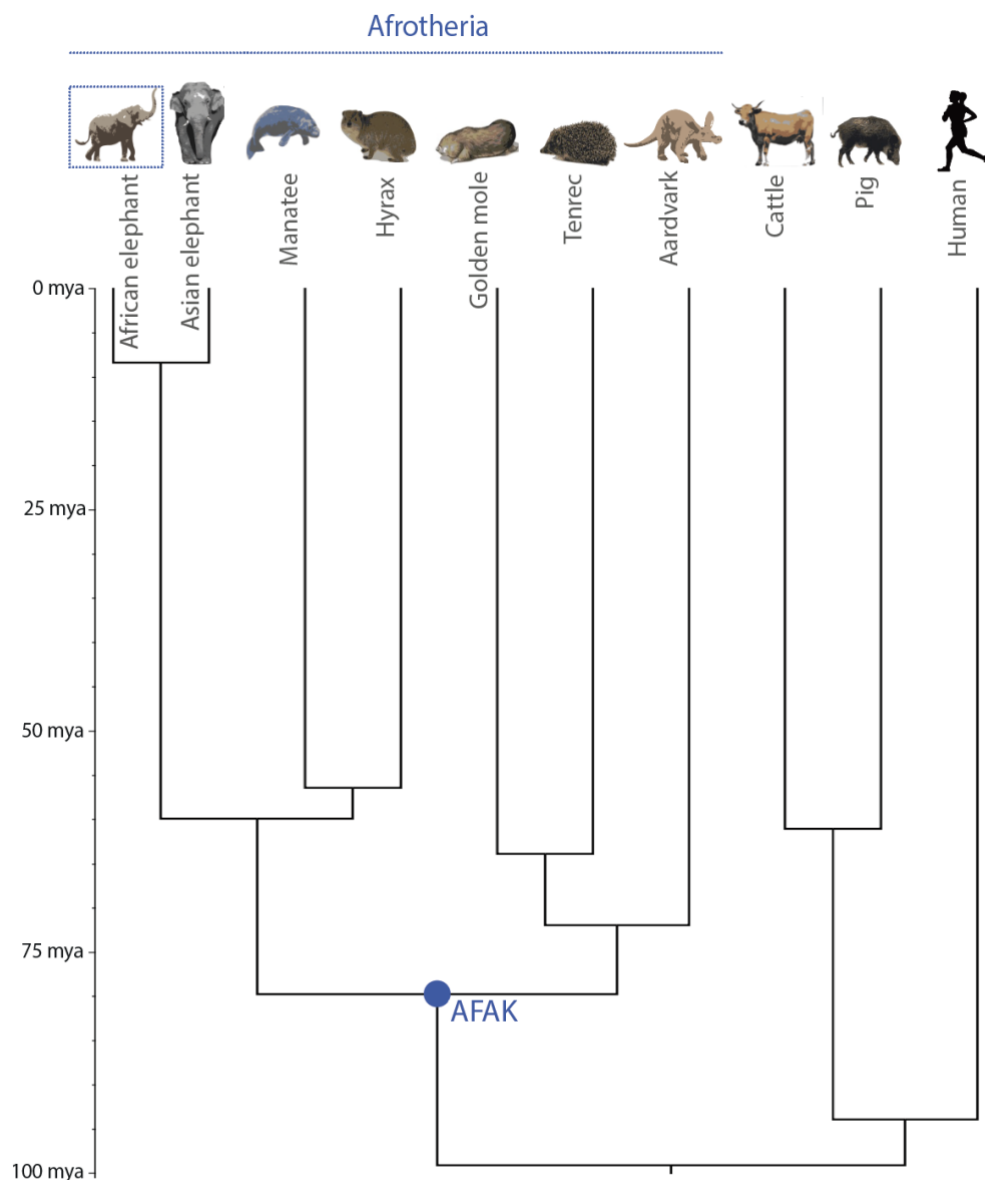


Figure 5.3. Phylogenetic tree of all species included in the afrotherian ancestral reconstruction. In blue the ancestral node reconstructed.

As done in previous chapters, we used DESCHRAMBLER to reconstruct the afrotherian ancestor with a resolution of 300 Kb and an adjacency score of 0.01. The coverage of African elephant genome was very high, 95.54%. We obtained 27 RACFs, higher than the 23+X ancestral haploid number suggested by cross-species chromosome painting between African elephant and human (Frönicke et al., 2003) (**Table 5.2**). We then manually curated the RACFs to increase coverage and minimise overestimation of CRs. RACFs were manually merged by comparing different ancestors and alignments using the most parsimonious scenario (see **Methods section 2.4**).

After the curation, the statistics of the reconstructed ancestors improved (**Table 5.2**). DESCHRAMBLER results showed quite similar chromosome numbers than the previous estimation of the afrotherian ancestor (Frönicke et al., 2003). The number of RACFs was reduced to 25, and both maximum and minimum sizes increased to 270,156 Kb and 3,484 Kb, respectively. Overall, our reconstruction showed a haploid number of 23 + X + one unplaced scaffold (3,484 Kb), the same number of chromosomes reported before (Frönicke et al., 2003).

Table 5.2. DESCHRAMBLER results. “Mapping number” is the number of chromosomes reported by a previous reconstruction which used chromosome painting (Frönicke et al., 2003).

Ancestor	No. RACFs	Mapping number	Max. Size (Kb)	Min. Size (Kb)	Total Size (Kb)	% *
AFAK	27	23+X	259,730	489	2,997,488	95.54
AFAK (curated)	25	23+X	270,156	3,484	2,997,488	95.54

*Coverage against Elephant genome (size=3,14Gb)

We recovered 71.43% of the syntenic associations between human and African elephant that were previously identified by cross-species chromosome painting, as well as all 23 fissions of human chromosomes (Frönicke et al., 2003). And, at the same time, we detected 28.57% of small CRs that had not been described before, indicating the high-resolution of our reconstruction.

Comparing AFAK to African elephant, a total of 58 CRs were found, classified in 31 inversions, 22 fissions, and five fusions (**Figure 5.4 and Table 5.3**). Moreover, there is no

AFAK chromosome that has been maintained collinear during its evolution, as some of them have been fused as entire blocks forming the chromosomes of the African elephant. These results highlight interchromosomal rearrangements as the main cause of genome reshuffling in afrotherian species. However, intrachromosomal rearrangements also played an important role during the evolution of the Afrotheria lineage.

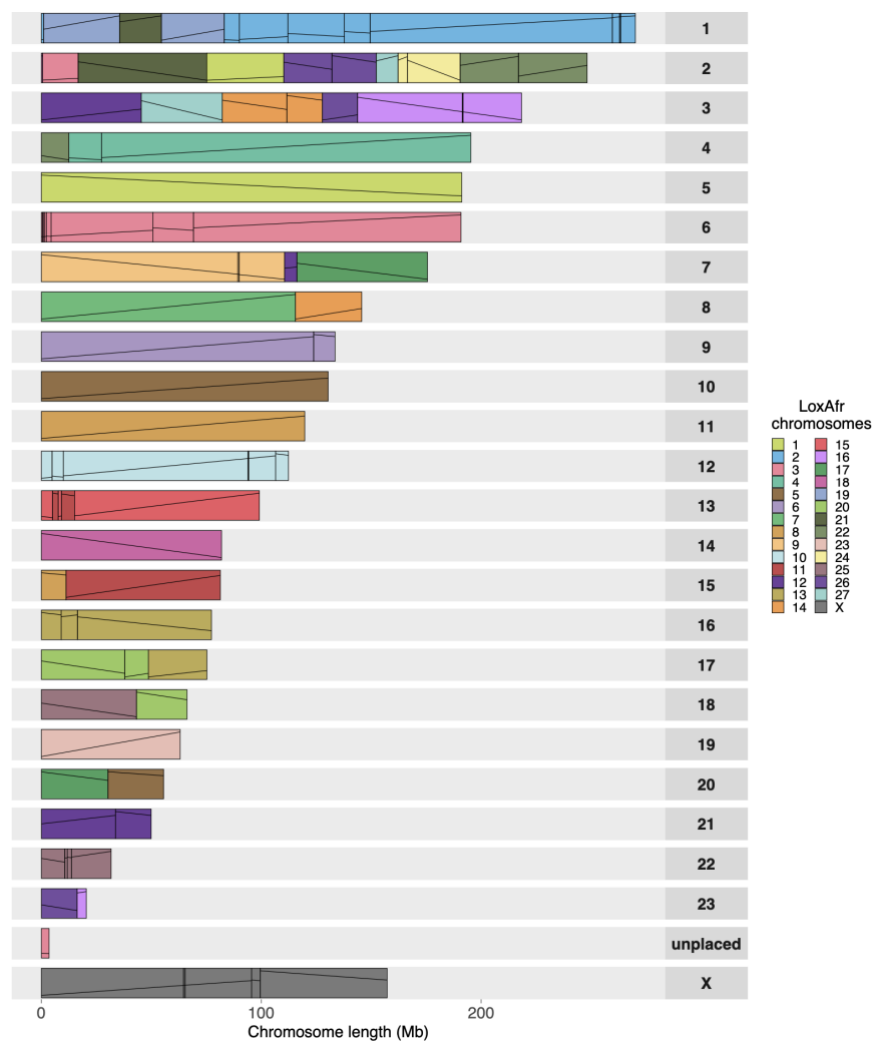


Figure 5.4. Ideogram comparing AFAK and African elephant chromosomes. LoxAfr indicates African elephant (*Loxodonta africana*) chromosomes.

Overall, a rate of 0.58 CRs/My characterised the evolution of AFAK to African elephant, with 0.31 and 0.27 CRs/My of inter- and intrachromosomal rearrangements, respectively (Table 5.3). These rates indicate the equal role of both inter- and intrachromosomal rearrangements in shaping afrotherian karyotypes.

Table 5.3. Summary table of type of CR differentiating every ancestor from African elephant.

No afrotherian species used	7
No afrotherian species used	3
No ancestral genomes reconstructed	1
	AFAK-African elephant
No of CRs	31 inversions
	5 fusions
	22 fissions
Rearrangement rate (per my)	0.58

5.2.4 Detection and study of evolutionary breakpoint regions (EBRs)

A total of 47 EBRs were detected, divided in 32 well-defined EBRs (size 0 - 50 Kb) and 15 not well-defined EBRs (size 50 - 300 Kb), while 18 were considered gaps (>300 Kb). We then phylogenetically classified the EBRs into: (i) Proboscidea-specific, and (ii) African elephant-specific. Within the 32 well-defined EBRs, 30 were Proboscidea-specific while only two were African elephant-specific. Similarly, from the 15 not well-defined EBRs, 14 were Proboscidea specific and only one was African elephant specific (**Table 5.4**). These results show the similarity between both elephant genomes.

Table 5.4. Classification of African elephant EBRs.

Chromosome	Start	End	Classification	Type
1	35,683,919	35,683,919	Proboscidea specific	Well-defined EBRs
2	7,044,311	7,044,311	Proboscidea specific	Well-defined EBRs
2	195,797,255	195,797,255	Proboscidea specific	Well-defined EBRs
2	196,303,304	196,310,952	Proboscidea specific	Well-defined EBRs
3	86,342,620	86,353,146	Proboscidea specific	Well-defined EBRs
6	126,550,135	126,550,135	Proboscidea specific	Well-defined EBRs
10	96,421,332	96,421,332	Proboscidea specific	Well-defined EBRs
10	96,728,078	96,733,561	Proboscidea specific	Well-defined EBRs
11	72,048,915	72,048,915	Proboscidea specific	Well-defined EBRs
11	78,154,021	78,154,021	Proboscidea specific	Well-defined EBRs
12	46,398,607	46,398,607	Proboscidea specific	Well-defined EBRs
12	52,049,514	52,049,514	Proboscidea specific	Well-defined EBRs
12	87,726,135	87,734,379	Proboscidea specific	Well-defined EBRs
14	68,881,911	68,881,911	Proboscidea specific	Well-defined EBRs
15	5,234,776	5,260,684	African elephant specific	Well-defined EBRs
15	5,205,107	5,232,803	Proboscidea specific	Well-defined EBRs

16	75,914,372	75,914,372	Proboscidea specific	Well-defined EBRs
16	27,203,528	27,219,292	Proboscidea specific	Well-defined EBRs
17	59,993,475	59,993,475	Proboscidea specific	Well-defined EBRs
18	82,688,476	82,688,476	Proboscidea specific	Well-defined EBRs
19	29,373,876	29,373,876	Proboscidea specific	Well-defined EBRs
22	12,512,852	12,512,852	Proboscidea specific	Well-defined EBRs
22	44,703,683	44,703,683	Proboscidea specific	Well-defined EBRs
24	24,167,559	24,167,559	African elephant specific	Well-defined EBRs
25	60,222,137	60,262,820	Proboscidea specific	Well-defined EBRs
26	16,042,484	16,042,484	Proboscidea specific	Well-defined EBRs
26	33,616,572	33,616,572	Proboscidea specific	Well-defined EBRs
26	55,588,453	55,588,453	Proboscidea specific	Well-defined EBRs
27	37,203,799	37,203,799	Proboscidea specific	Well-defined EBRs
X	103,444,847	103,444,847	Proboscidea specific	Well-defined EBRs
X	107,607,452	107,607,452	Proboscidea specific	Well-defined EBRs
X	71,327,583	71,350,950	Proboscidea specific	Well-defined EBRs
2	32,136,952	32,252,149	Proboscidea specific	Not well-defined EBRs
3	31,947,210	32,134,917	Proboscidea specific	Not well-defined EBRs
3	25,850,296	26,149,352	Proboscidea specific	Not well-defined EBRs
8	120,487,352	120,564,205	Proboscidea specific	Not well-defined EBRs
10	5,280,341	5,331,628	Proboscidea specific	Not well-defined EBRs
13	28,069,529	28,292,559	Proboscidea specific	Not well-defined EBRs
13	89,182,736	89,439,513	Proboscidea specific	Not well-defined EBRs
13	97,605,695	97,876,176	Proboscidea specific	Not well-defined EBRs
14	36,581,479	36,663,960	Proboscidea specific	Not well-defined EBRs
16	27,521,797	27,656,707	Proboscidea specific	Not well-defined EBRs
18	83,021,036	83,231,934	Proboscidea specific	Not well-defined EBRs
20	11,100,256	11,335,978	Proboscidea specific	Not well-defined EBRs
24	24,064,621	24,160,064	African elephant specific	Not well-defined EBRs
25	58,089,047	58,232,572	Proboscidea specific	Not well-defined EBRs
X	70,317,911	70,510,779	Proboscidea specific	Not well-defined EBRs

Well-defined EBRs are not uniformly distributed along African elephant chromosomes, there is not a correlation between number of EBRs and chromosome size (**Figure 5.5A**). In fact, some of the chromosomes with a medium size (average size of 112 Mb) present the highest number of well-defined EBRs. African elephant chromosomes 2, 12, 26, and X are the ones containing more EBRs (3) while chromosomes 4, 5, 7, 8, 9, 13, 20, 21, and 23 have not well-defined EBRs. Similarly, when we expand the EBR selection and study both, well-defined and not well-defined EBRs together (**Figure 5.5B**), there is not a correlation between number of EBRs and chromosome size. In this scenario, chromosomes 2 and X are the ones containing

more EBRs (4) while chromosomes 4, 5, 7, 9, 21, and 23 remain as syntenic blocks. The lack of EBRs in these chromosomes could be due to: (i) lack of resolution, some of the CRs would be delimited by EBRs higher than 300 Kb, and in these cases we have considered them as gaps instead of EBRs, one example is chromosome X; and (ii) cases where no CRs happened during evolution, such as African elephant chromosomes 7 and 23.

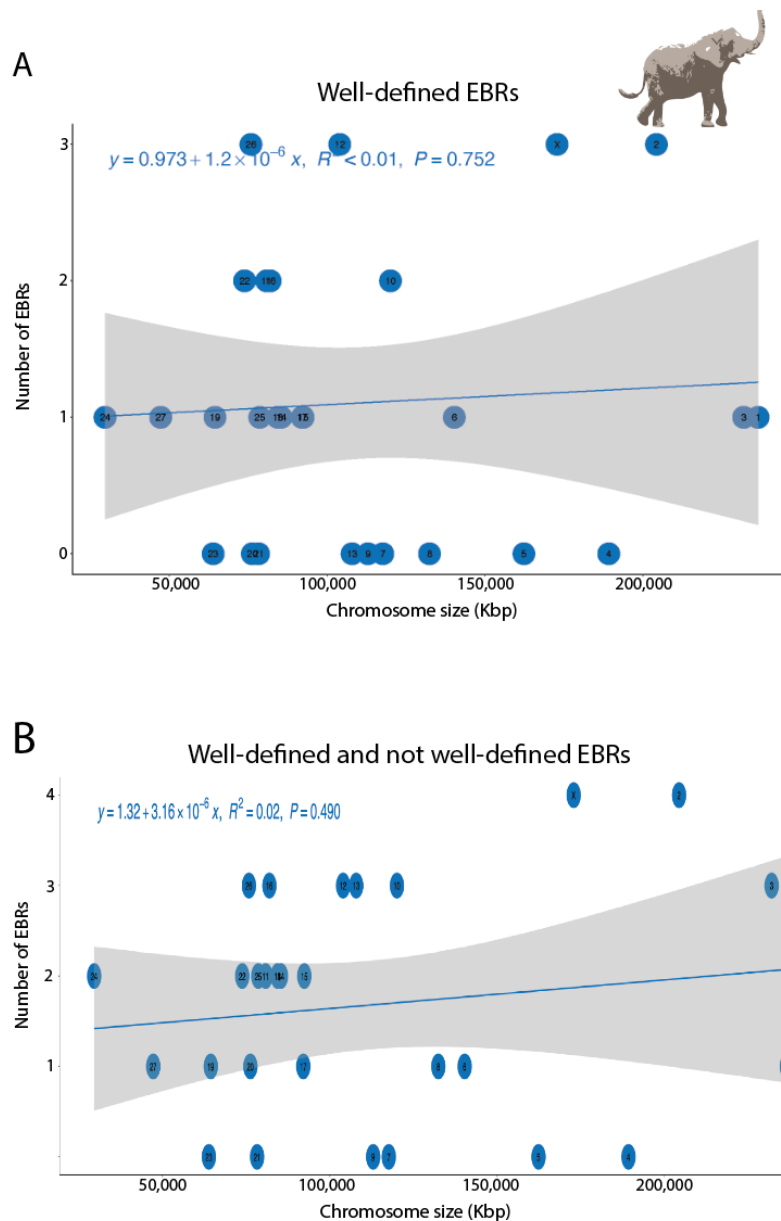


Figure 5.5. Correlation between the number of well-defined (A) and well-defined and not well-defined (B) EBRs and African elephant chromosomes. Chromosome sizes are in Kb.

To determine whether there is a correlation between the distribution of EBRs and gene-rich genome regions, we performed several permutation tests in three different scenarios: (i) regions within well-defined EBRs; (ii) regions within not well-defined EBRs; and (iii) regions near

EBRs (**Figure 5.6**). For the first scenario, there was a clear negative association (Z-score = -2.417 and P-value = 0.006) between well-defined EBRs and genes (**Figure 5.6A**). Instead, no statistically significant association was found between not well-defined EBRs and genes (**Figure 5.6B**). However, regions near the well-defined EBRs showed higher number of genes than the EBRs as such (P-value in the 20 Kb downstream region from well-defined EBRs = 0.0078 and P-value in the 20 kb upstream region from well-defined EBRs = 0.0026) (**Figure 5.6C**). As previously reported (M. Farré et al., 2019), many of the well-defined EBRs are very narrow, sometimes even with sizes of 0 or 1 bp, explaining that EBRs are located in gene-rich regions of the genome without breaking any gene.

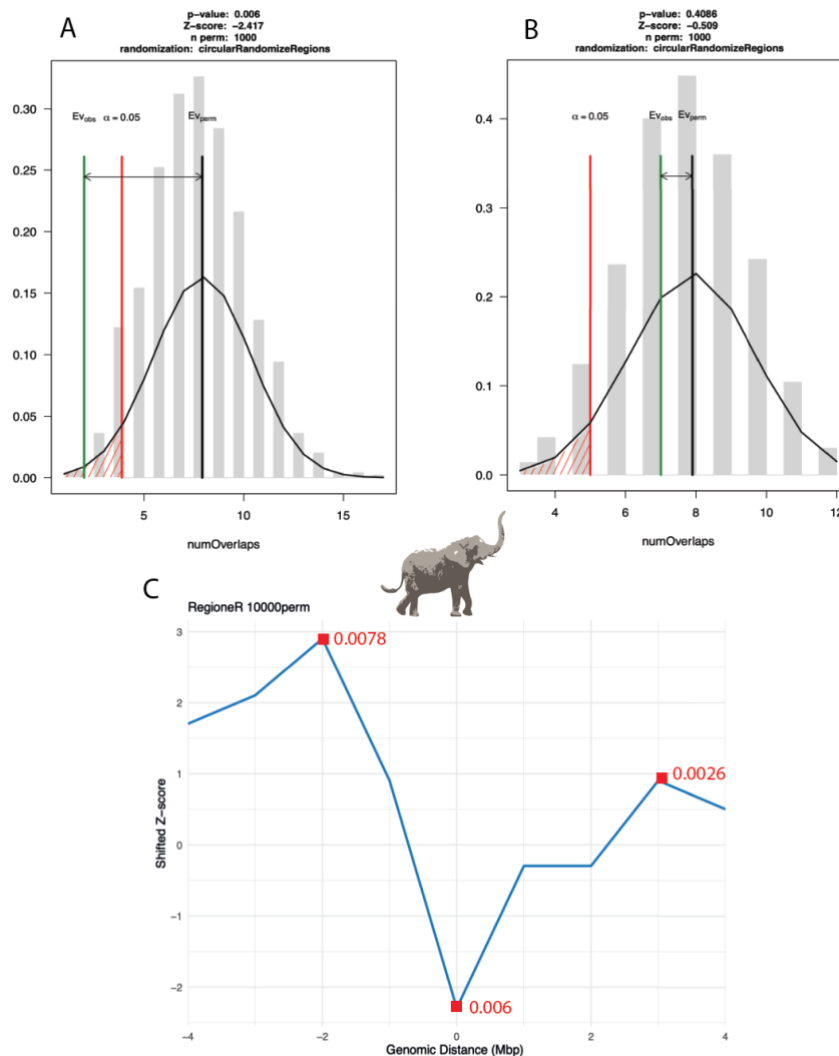
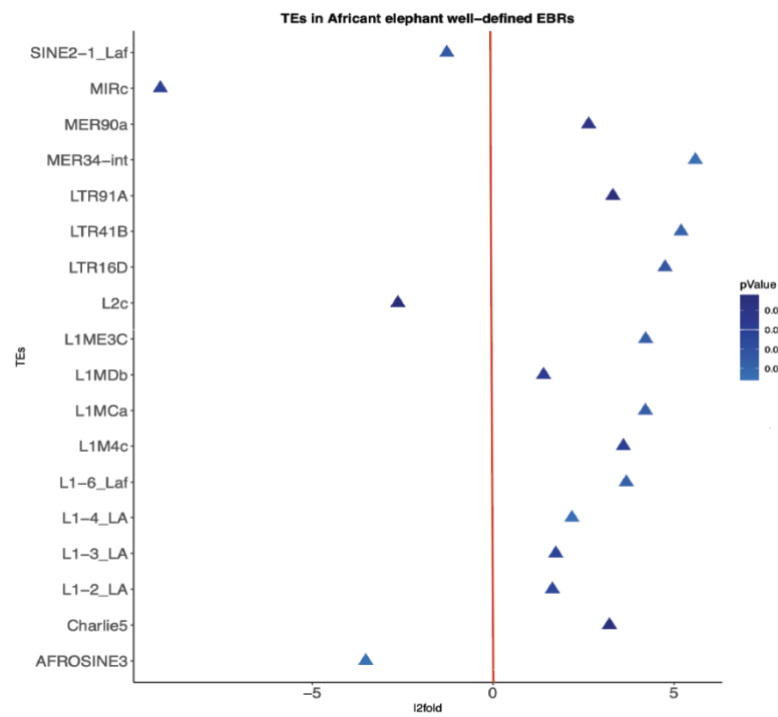


Figure 5.6. Permutation tests performed to establish the association between gene content and well-defined EBRs (A), not well-defined EBRs (B), and regions near well-defined EBRs (C) in African elephant. In C, red number show P-values in the regions marked with a red square.

To determine whether EBRs might have a possible role in rearranging key genes in Afrotheria species, we performed a GOEA. Although no significant GO terms were found for any of the sets of EBRs analysed (see Methods section 2.7), well-defined EBRs contained two genes, *LRCH1* and a novel gene (ENSLAFG00000007181), and nine genes after they were expanded to 50 Kb (*LRCH1*, *ESD*, *AAK1*, *CD1D*, *RHBDD2*, *PLEKHB2*, *MAP9*, *CD70*, and the novel gene ENSLAFG00000007181). Instead, not well-defined EBRs contained 12 genes: *KIAA0825*, *PARD6G*, *ADNP2*, *MC2R*, *ENOSF1*, *TYMS*, *PRXL2A*, *ANKRD13A*, *ABO*, the open reading frame *C12orf76*, and two novel genes (ENSLAFG000000031024 and ENSLAFG000000029660).

Finally, we wanted to know if the correlation of EBRs with genomic regions enriched in TEs, that have been seen in other lineages (M. Farré et al., 2019; Damas et al., 2022), was present in afrotherian species. To do so, we searched for genomic association between repetitive elements and well-defined and not well-defined EBRs. In both cases we found an enrichment in repetitive elements (**Figure 5.7**). One of the most remarkable differences is that well-defined EBRs presented lower variability of repetitive element content (18 different types) than not well-defined EBRs (63 different types), which can be due to the narrow size of the majority of the well-defined EBRs. Among all types of TEs found in this analysis, some of them are afrotherian-specific, such as AFROSINEs and L1-LA elements. The most repeated type of repeats are long interspersed nuclear elements (LINEs; L1), and long terminal repeats (LTRs) and medium reiterated repeats (MER). While AFROSINEs showed lower numbers than expected, indicated with negative values of 12fold, L1-LA elements showed a raise in their presence within both kinds of EBRs.

A



B

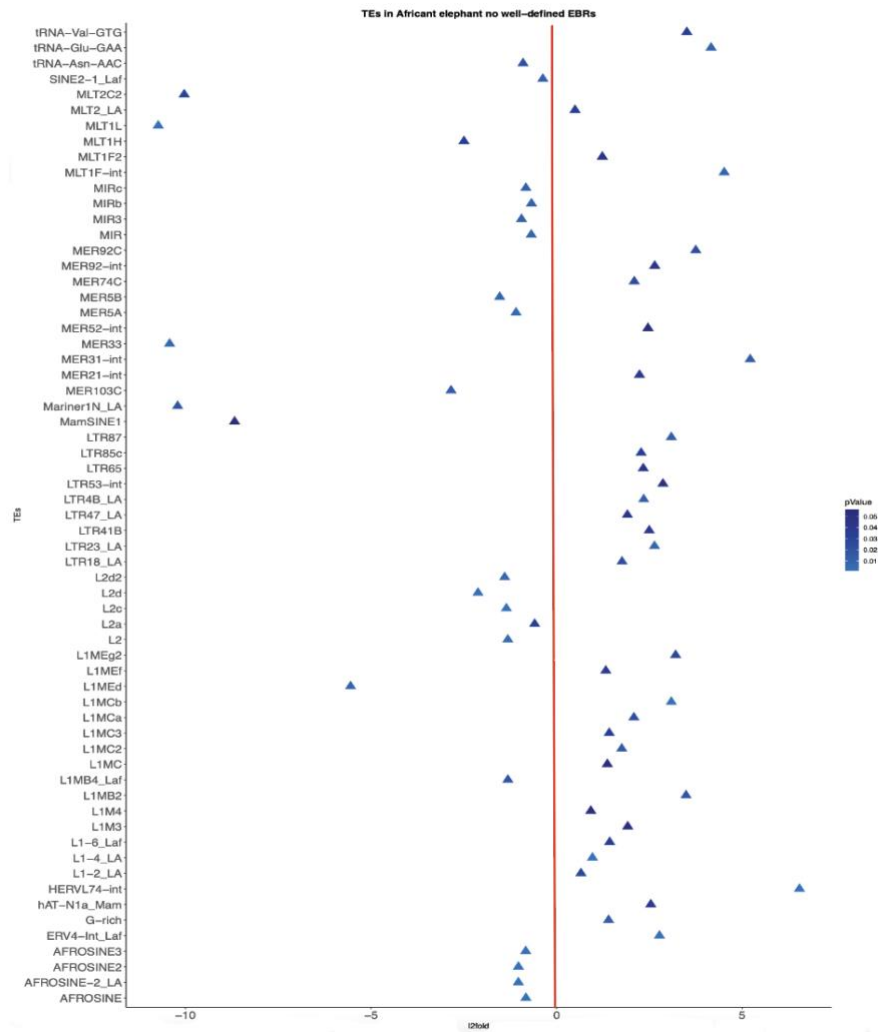


Figure 5.7. Repetitive elements associated with well-defined and not well-defined EBRs (A and B, respectively). X axis represents the log2fold, which indicates less TEs observed than expected in case of negative numbers and more TEs observed than expected in case of positive numbers. The red line highlights the division of negative and positive log2fold values. Y axis represents the different TEs. Triangles colour varies depending on their P-value.

5.2.5 Detection and study of homologous syntenic blocks (HSBs)

The next step was to define and classify HSBs. Depending on the type of CR, we classified CRs in collinear (when they were conserved between species), inverted HSBs, or different combination involving two or more consecutive inter- or intrachromosomal rearrangements (cases explained in **Methods Figure 2.2**).

In African elephant, we classified these HSBs according to the rearrangements occurring during the evolution from AFAK to the current African elephant genome (**Tables 5.5 and S4**).

Within this classification, only 12.6% of African elephant genome corresponds to collinear HSBs, 3.06% to inverted HSBs, and 82.92% to HSBs related to interchromosomal rearrangements (specifically, 20.83% to fissioned and collinear HSBs, 27.37% to fissioned and fused HSBs, 6.93% to fissioned and inverted HSBs, 24.24% to fused and collinear HSBs, and 3.54% to fused and inverted HSBs). Of all types, the biggest block size, 217,599 Kb, is found within the fissioned and inverted HSBs, and the lowest block size, 60,890 Kb, was found within the fused and inverted HSBs (**Table 5.5**). These results together with our previous findings, remark the important role that interchromosomal rearrangements have played in genome reshuffling during Afrotheria species evolution. Furthermore, they highlight that current afrotherian genomes have been shaped through consecutive combination of CRs from their ancestral form.

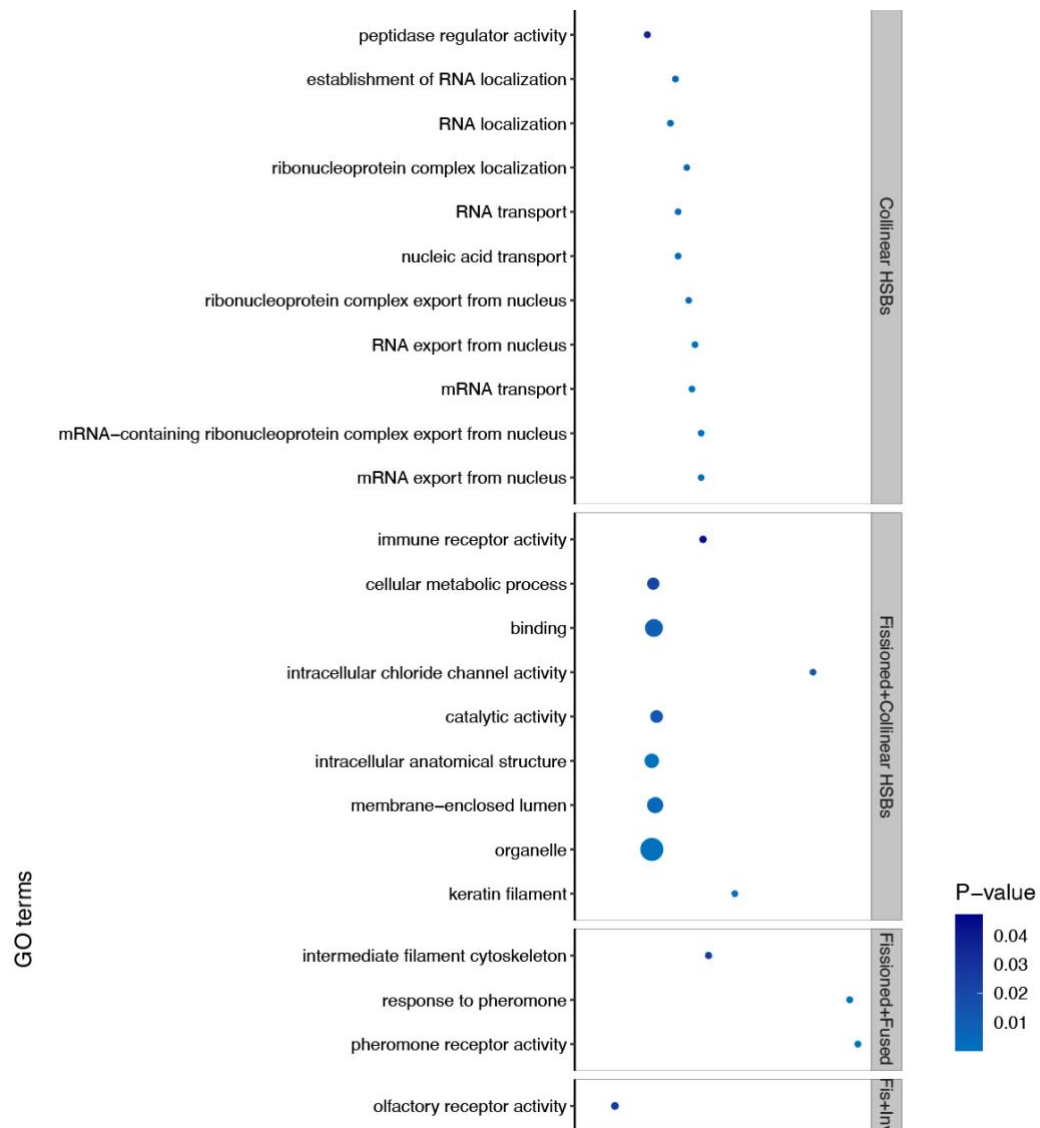
Table 5.5. Summary of African elephant HSBs.

Classification	Total (Kb)	Min. Size (Kb)	Max. Size (Kb)	%*
Collinear	395,352	5,270	126,522	12.60
Fission + Collinear	653,483	675	172,322	20.83
Fission + Fusion	858,746	303	72,011	27.37
Fission + Inversion	217,599	506	217,599	6.94
Fusion + Collinear	760,642	834	200,505	24.25
Fusion + Inversion	111,116	529	60,890	3.54
Inversion	96,037	307	65,082	3.06

* Coverage against African elephant genome (3.14Gb)

We also performed GOEA in all the types of HSBs (**Figure 5.8**). A total of 2,019 genes were found in collinear HSBs and showed GO terms related to RNA mainly, while the 352 genes found in inverted HSBs showed no significant GO terms. Comparing the five cases involving interchromosomal rearrangements, they are enriched in genes with very different GO terms: fissioned and collinear HSBs presented 4,056 genes with significant GO terms related to chemical activities, binding, or organelles, while fissioned and inverted HSBs presented 1,249 genes with the only significant GO term of olfactory receptor activity. Instead, fissioned and fused HSBs presented 5,062 genes related to pheromones and cytoskeleton; fused and inverted HSBs contained 490 genes leading to an enrichment of genes related to different peptidase activities, binding, and cell adhesion, while fused and collinear HSBs contained

3,285 genes leading to an enrichment in genes related to different chemical activities, binding, and organelles.



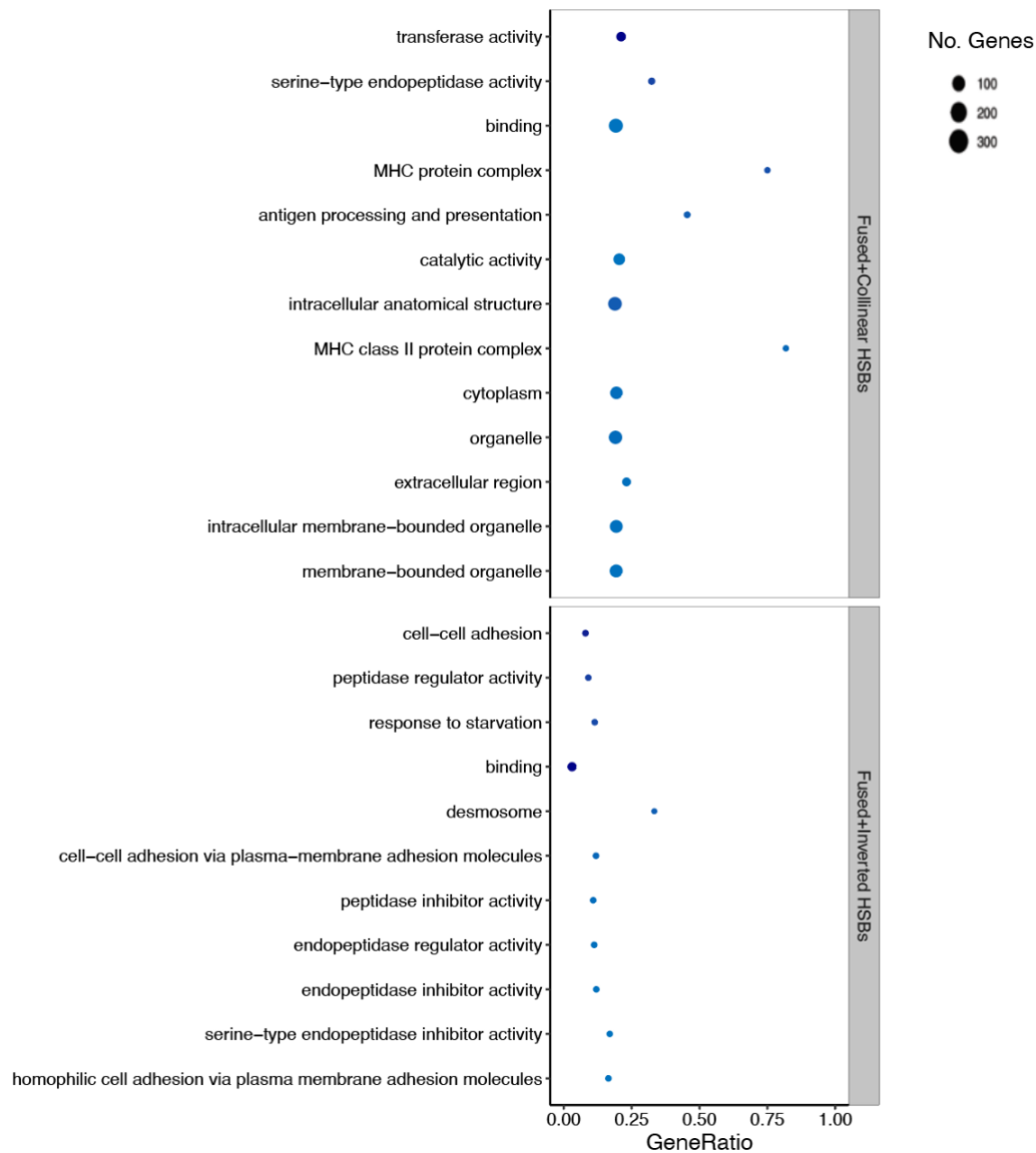


Figure 5.8. GOEA of African elephant HSBs according to their classification. X axis represents the gene ratio (number of genes/total number of genes). Y axis represents the GO terms. Bubbles size varies depending on their number of genes in every GO term (No. Genes). Bubbles colour varies depending on their P-value.

5.3 Discussion

In this chapter, we propose a new karyoor the ancestor of all afrotherian species: AFAK. This is the first time that AFAK is reported using a computational approach supported by cytogenetic data. Previous studies have used reciprocal cross-painting of afrotherian species, like armadillo and African elephant, and human to reconstruct the mammalian ancestral karyotype (F. Yang et al., 2003), or have used cross-painting of African elephant and human to propose the karyotype of the afrotherian ancestor (Frönicke et al., 2003). Our reconstruction presented the same diploid number that reported before (Frönicke et al., 2003), $2n = 48$, and

additionally, we reported the following fusions among African elephant chromosomes (i) chromosomes 2, 19, and 21; (ii) chromosomes 1, 3, 12, 21, 22, 24, and 27; (iii) chromosomes 12,14, 16, 26, and 27; (iv) chromosomes 7 and 22; (v) chromosomes 9, 12, and 17; (vi) chromosomes 14 and 17; (vii) chromosomes 11 and 15; (viii) chromosomes 8 and 11; (ix) chromosomes 1 and 13; (x) chromosomes 1 and 25; (xi) chromosomes 5 and 17, and (xii) chromosomes 16 and 26.

Recent studies have reconstructed several mammalian ancestral genomes, including the afrotherian ancestor (Damas et al., 2022), however for this ancestral karyotype they did not determine neither EBRs nor rearrangements rates. In this thesis, by specifically reconstructing the afrotherian ancestor, we have been able to identify for the first time, EBRs, HSBs, and CRs occurred during the evolution from AFAK to the current karyotype of the African elephant, using DESCHRAMBLER. The comparison of AFAK and the African elephant genome showed an extensive reorganization, with the following CRs: 31 inversions, 22 fissions, and five fusions. These numbers demonstrate that both inter- and intrachromosomal rearrangements have been very important leading afrotherian evolution, showing similar numbers and indicating that both types of CRs equally contributed to the evolution of afrotherian species. It would be very interesting to see if using different reference genomes and combining reconstructions, we would be able to reconstruct more afrotherian ancestors.

Our findings agree with previous studies that associated EBRs location with rich-gene regions in the genome (M. Farré et al., 2019; Damas et al., 2022). Moreover, it has been reported that in other lineages, such as the ruminant lineage, EBRs showed an enrichment of ruminant-specific TEs (M. Farré et al., 2019), and in our results, we also reported an enrichment of afrotherian-specific transposable elements, such as L1-LA elements, known to be useful markers for solving phylogenetic relationships among afrotherian species (Nishihara et al., 2005; Gilbert et al., 2008). However, other afrotherian-specific TE reported before (Nishihara et al., 2005; Gilbert et al., 2008), the AFROSINES, showed a negative association with both, well-defined and not well-defined EBRs. This could be due to AFROSINES being located in regions near EBRs rather than within them. Furthermore, we also could see the presence of more transposable elements that have been reported to be conserved in vertebrates, like Charlie elements and endogenous retrovirus (ERVs) (Damas et al., 2022).

Although we were unable to find GO terms for those genes within or near EBRs, for well-defined EBRs, we found the gene *LRCH1* (leucine-rich repeats and calponin homology domain containing 1), which has been broken by the CR. This gene is a potential translational target to improve T cell immunotherapy against infection and tumours (C. Liu et al., 2020). After expanding the well-defined EBRs to 50 Kb, we found four genes related to chromosome instability and tumorigenesis: (i) *ESD* (esterase D), widely distributed in mammals, playing an important role in drug metabolism, detoxification, and biomarkers, and is closely related to the development of tumours (Yao et al., 2022); (ii) *RHBDD2* (Rhomboid domain containing 2), a rhomboid family member overexpressed during stages of several cancers such as breast cancer or colorectal cancer (Canzoneri et al., 2018; Palma et al., 2020; Ferretti et al., 2021); and (iii) *MAP9* (microtubule associated protein 9), critical for spindle assembly and cytokinesis, and that has been studied because their loss of expression triggers chromosomal instability (S. Wang et al., 2020; Mok et al., 2020). Interestingly, within the genes found in not well-defined EBRs, *ADNP2* (activity-dependent neuroprotective protein) is a regulator of oxidative response, a well-known antioxidant gene related to the neurodevelopment (Malishkevich et al., 2015; Jeong et al., 2022). It has been demonstrated that brain and body size variations appear significantly correlated in Proboscideans. Furthermore, body size has been a trait evolving under selective pressure, co-evolving with brain size during the Proboscidea evolution (Benoit et al., 2019). Finally, another interesting gene found within not well-defined EBRs is *ABO*, a gene which determines the blood group of an individual in humans (Yazer & Palcic, 2005). All in all, our EBR analysis reinforces the relation of CRs and adaptive traits in Afrotheria.

Finally, we performed GOEA, not only over the genes within and near the EBRs but also in genes within the different types of HSBs. As mentioned in previous chapters, we are interested in associating the different type of CRs producing the HSBs to functions related to chromatin configuration or processes such as replication or transcription, as the relationship between CRs, chromatin configuration, and changes in gene expression has been demonstrated in several species (Marques-Bonet et al., 2007; Fuller et al., 2016; Capilla et al., 2016). Within collinear HSBs, the majority of the GO terms found are related to RNA composition and processes. Within the HSBs affected by interchromosomal rearrangements, we found an enrichment in genes related to olfactory responses to pheromones, previously described as new adaptive gene functions related to CRs (Larkin et al., 2009; Groenen et al., 2012; Capilla et al., 2016). Moreover, we also found genes related to tumoral progression and instability in the genome, such as: (i) *APMAP* (adipocyte plasma membrane-associated protein), a glycosyl type II

transmembrane protein that is mainly distributed in the adipocytes and that has been reported to be involved in lipid transport and the development of several types of cancers, such as prostate, colorectal, and cervical cancer (Kamber et al., 2021; Zhu et al., 2021); (ii) *EEF1A2* (eukaryotic protein translation elongation factor 1 α 2), an oncogene that promotes the progression of breast and pancreatic cancer (Jia et al., 2021); and (iii) *WDTC1* (WD40 and tetratricopeptide repeats 1), related to accumulation of adipocytes, fat, and development of obesity, and whose mutation is associated to the development of several cancers such as colorectal cancer (Ducos et al., 2017; X. Wang et al., 2022). Interestingly, genes related to HSBs produced by interchromosomal rearrangements not only are related to tumour development, but to adipocyte activity, and it will be interesting for future work, to use previous studies characterizing the adipose tissue in afrotherians (Nilsson et al., 2014) to detect specific genes for their evolution. Additionally, no significant GO terms were found in inverted HSBs, but it is interesting that many of the genes found within them are from the SERPIN family, which are related to the development of metastasis in the brain (Valiente et al., 2014; Soman & Nair, 2022), highlighting the possibility of finding genes related to brain function within the genes involved in the afrotherian evolution. Overall, we found an association of EBRs and HSBs in afrotherian species to genes related to cancer, which is an indicative of the genome instability and a bigger propension of suffering CRs. Interestingly, we also found an association of EBRs and HSBs to genes related to the adipose tissue and brain function. For future work, it would be very interesting to further study these genes, to find any connection with these genes and tumoral progression and the 3D chromatin structure, and to further study the evolution of the brain and adipose tissue in the afrotherian lineage.



6. Evolution in mammals: Looking for convergence traits in extreme chromosomes

6.1 Introduction

Studying the role played by CRs in genome organization is fundamental to understand the dynamics of chromosomal evolution, and its implication in speciation and macroevolutionary events. For this purpose, mammals are a good study model due to their extreme variability in terms of genome size and chromosome diploid numbers. The three lineages used in this thesis are a good example of this variability:

- (i) Ruminantia is characterized for containing the species with the most extreme chromosome number variability. This variability is also related to the chromosome size of these species, for instance, species with high diploid numbers show smaller chromosome sizes than species with low diploid numbers.
- (ii) Marsupialia is characterised for containing species with massive chromosome sizes compared to other mammalian groups. Moreover, their diploid numbers are quite stable, and the most frequent ones are $2n = 14$ and $2n = 22$.
- (iii) Afrotheria is characterized for containing a great variability of chromosome numbers, which is also reflected in the chromosome size of afrotherian species. Same than in ruminants, species with bigger chromosomes present low diploid numbers, and vice versa.

In general, there is a clear association between chromosome number and chromosome size, which has affected genome reshuffling and chromosome organization during evolution.

The ancestral karyotypes of Ruminantia and Afrotheria have a hypothetical chromosome diploid number of $2n = 48$ (M. Farré et al., 2019) and $2n = 46$ (Frönicke et al., 2003), respectively. To achieve the configuration of current species such as the Indian muntjac ($2n = 6/7$) or white-tailed deer ($2n = 70$) in ruminants, or such as armadillo ($2n = 20$) and African elephant ($2n = 56$) in afrotherians, their ancestral karyotypes must have gone through a great reduction or amplification of their chromosome numbers via fusions and/or fissions, which at the same time, have originated the increase or reduction on their chromosome sizes. For these two lineages, the most parsimonious scenario explaining their evolution would be interchromosomal rearrangements such as fusions and fissions shaping their genomes.

On the other side, the ancestral karyotype of Marsupialia has a hypothetical chromosome diploid number of $2n = 14$ (Deakin & O'Neill, 2020). As mentioned before, marsupial species

show very stable chromosome numbers, but they also present very different phenotypes. The most parsimonious scenario for their evolution, in this case, would be intrachromosomal rearrangements, such as inversions, shaping their genomes. However, the marsupial ancestor possesses less chromosomes than the ruminant and afrotherian ones, which implicates an additional scenario of chromosome number reduction lead by interchromosomal rearrangements, occurring in the evolution from the mammalian ancestor (with a proposed diploid number of $2n = 40$) (Damas et al., 2022) to the marsupial ancestor.

Taking into consideration that the mammalian ancestor karyotype has a diploid number of $2n = 40$ (Damas et al., 2022), the most plausible scenario would be a drastic independent reduction in chromosome number in these three lineages. Seeing that these changes most likely originated independently in each lineage, we could consider this reduction of chromosome number as a convergent trait. This raises the question of how this convergent trait evolved and which genes were then involved to deal with the drastic reduction of chromosomes occurred in the three lineages.

Convergent evolution, which consists in the independent evolution of similar features in different evolutionary lineages (Losos, 2011), is known to be widespread across the mammal tree of life. Many examples can be found such as the evolution of phenotypes of marine mammals (Foote et al., 2015), or the evolution of echolocating genes in diverse mammals like bats or marine mammals (Parker et al., 2013). Therefore, convergent evolution is a good scenario for explaining the extreme reduction of chromosome numbers in several mammalian species and the fixation of this trait across the evolution of these species.

One of the possible ways to study convergent evolution is the detection of positive selection in genes involved in the specific trait under study. Positive selection or adaptation, is the process by which alleles that increase organismal fitness increase in frequency in a population (Hejase et al., 2020). By detecting signatures of this process, the understanding of the genetic basis for important phenotypic traits has been possible in both, human and non-human species. Searching for signatures of positive selection in genomes can reveal an ongoing, recent, or ancient response to adaptation, and can shed light on the study of convergent evolution. Moreover, the ability to detect these signatures of selection, which are known to pinpoint genetic variants responsible for significant changes in gene regulation and expression, or protein synthesis, structure, and function, has become more efficient due to their relevance for

human health studies (Hejase et al., 2020; Abondio et al., 2022). The approach we have used to detect positive selection in this thesis is based on the premise that as positive selection repeatedly targets the same genes, the identification of these genes will show convergent shifts in evolutionary rates (i.e. number of nucleotide or amino acid substitutions per unit time), due to reduced or increased selective constraints on them (Kowalczyk et al., 2019) (see **Introduction section 1.5.5.1**).

Altogether, for identifying signatures of convergent evolution in those mammals that have drastically reduced their chromosome numbers, while having increased their chromosome sizes, we aimed to identify traits related to the CRs which might have played an important role in order to reshuffle genomes into large chromosomes. As seen in **Introduction section 1.3.3.2.1**, CRs are usually originated by erroneous repair of DSBs, related to stalls of the replication fork, or highly specialised cellular processes such as meiosis. Moreover, as seen in **Introduction section 1.3.3.2.3**, some regions of the genome are more prone to suffer breaks and re-joins (EBRs) while others are more prone to be conserved (HSBs), and they are not randomly distributed across the genome, remarking the important role of the 3D organization of the nucleus (including the chromatin state) for the generation of CRs.

In this chapter, we aimed to address how the differences in chromosome number and size found in mammals can be a consequence of convergent evolution. First, we searched for orthologous genes in the three lineages. These orthologous genes, by definition, have evolved from a common ancestral gene by speciation, and, in general, they retain the same function during evolution. The next step was to study the function of these genes, verifying their participation in processes or functions that can be altered by CRs. Finally, we analysed whether these genes have been affected by positive selection, confirming the hypothesis of reduction in chromosome number in mammals as a consequence of convergent evolution.

For the purpose of identifying signatures of convergent evolution in mammals with an extreme range of diploid numbers and chromosome sizes, we selected representative species of the three lineages studied in this thesis, Ruminantia, Marsupialia, and Afrotheria, showing extremely different karyotypes (see **Methods Table 2.2**).

6.2 Results

To understand how chromosome sizes and number evolved in the mammalian clade, we focused on detecting convergent evolution in orthologous genes of the most extreme cases found in the Ruminantia, Marsupialia, and Afrotheria lineages.

6.2.1 Identification of orthologous genes in mammals

A total of 14 species were included in this analysis (see **Methods Table 2.2**). First, we classified all species into two groups, depending on their chromosome number and size: (i) Group 1, containing all species with low diploid numbers ($2n \leq 20$) and large chromosome sizes (average chromosome size >120 Mb); and (ii) Group 2, containing all species with high diploid numbers ($2n > 20$) and small chromosome sizes (average chromosome size <120 Mb). For Group 1 we chose: black and Indian muntjacs, representing ruminants; Tasmanian devil, gray short-tailed opossum, common brushtail possum, and tammar wallaby, representing marsupials; and aardvark, representing afrotherians. For Group 2 we chose: cattle, white-tailed deer, red deer, and Chinese water deer, representing ruminants; and African elephant, to represent afrotherians; however, no marsupial species were selected for this group as they all have bigger chromosome sizes. Interestingly, the extremes of this analysis were found in the Ruminantia lineage, where we can find the Indian muntjac with the smaller diploid number ($2n = 6/7$) and the highest average chromosome size (838.5 Mb), and the white-tailed deer with the highest diploid number ($2n = 70$) and the smallest average chromosome size (66.8 Mb) (see **Methods Table 2.2**).

Then we obtained single-copy orthologous genes using OrthoFinder and all the 14 species. The master phylogenetic tree including all species (**Figure 6.1**) was inferred using the proportion of species trees derived from single-locus gene trees supporting each bipartition as its measure of support (Emms & Kelly, 2015, 2019). In Ruminantia, the lineage showing the most extreme karyotypes, species of Group 1 are not monophyletic, reinforcing our hypothesis that the reduction of diploid number is a convergent trait, and suggesting that positive selection could happen in different ancestors within the lineage.

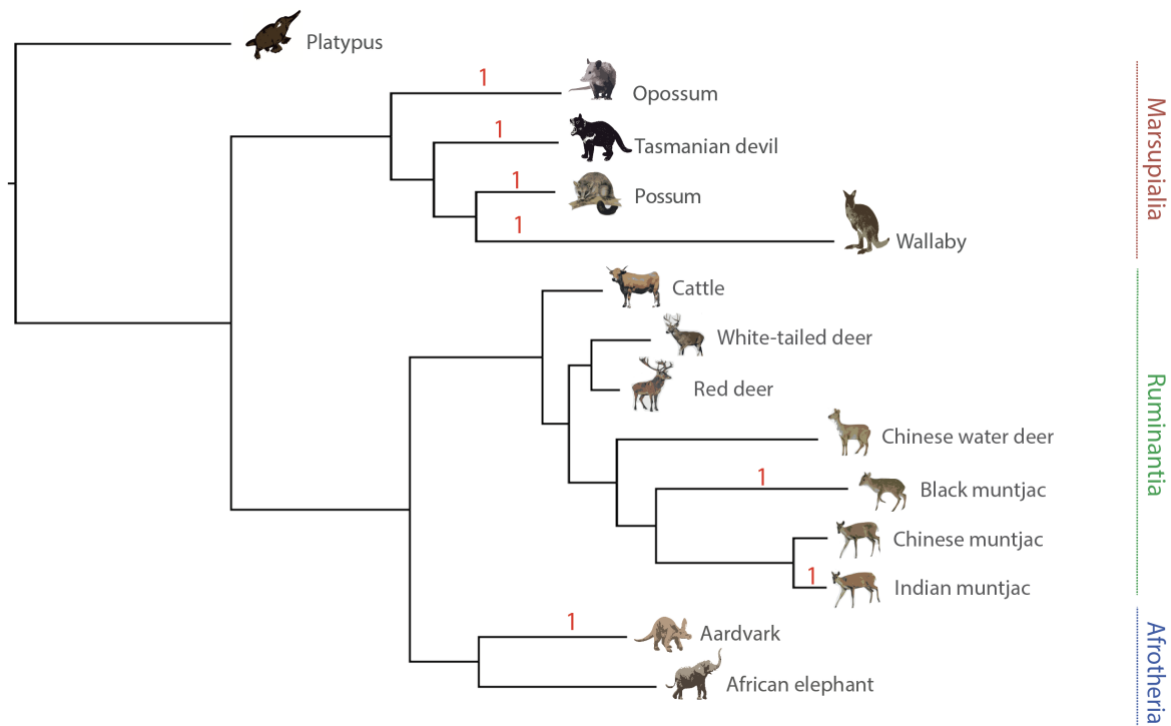


Figure 6.1. Phylogenetic tree inferred for OrthoFinder. Species within Group 1 are indicated in red. OrthoFinder inferred the species from the set of unrooted orthogroup gene trees using STAG (Species Tree inference from All Genes) method, which produces branch lengths according to the molecular phylogenetic distances between species.

A total of 33,827 orthologous genes were identified among the 14 species, including the outgroup platypus, of which 4,747 genes were single-copy orthologs.

6.2.2 Estimation of convergence rates

Once all single-copy orthologs were identified, the next step was searching for convergent rates of evolution. Using RERConverge, we estimated the relative evolutionary rates (RERs) of all branches in the phylogenetic tree for each of the 4,747 single-copy orthologous genes. Intuitively, a gene's RER for a given branch represents how quickly or slowly the gene is evolving on that branch relative to its overall rate of evolution throughout the tree (Chikina et al., 2016; Kowalczyk et al., 2019). For instance, genomic elements, or traits, whose function is important for the development of the convergent phenotype would evolve more slowly, and on the contrary, traits evolving more quickly for a given phenotype would be related to loss of function or relatively decreased functional importance.

We ran RERConverge twice, using first Group 1 and then Group 2, as subsets of ingroup (or foreground) species. After estimating the RERs of all 4,747 genes in both analyses, a total

of 529 genes were found to be statistically significant (P-value adjusted < 0.05) in the analysis using Group 1 as foreground species, while 265 genes were statistically significant for the analysis using Group 2 as foreground species. The P-value adjusted obtained in RERConverge is an estimate of the P-value adjusted for multiple comparisons using the Benjamini-Hochberg procedure (i.e., an estimate of the false discovery rate, or FDR).

Because we are particularly interested in those species with low diploid numbers and large chromosome size, we further analysed those genes in Group 1. We then performed a binary trait analysis using RERConverge and selected Group 1 species as foreground species. To identify those genes potentially related to low diploid numbers, we focused on those common genes among both groups (**Table S5**) showing a high RER score in species in Group 1 and negative RER score in species in Group 2. For instance, one of the genes with the most significant P-value was *JMJD1C* (**Figure 6.2 A and B**). The RERs showed high positive values in species like wallaby, Tasmanian devil, and possum, and high negative values in species such as cattle and Chinese muntjac. These results indicate that *JMJD1C* is a gene related to the evolution of marsupials, as all of them showed positive RERs values, but is not important for the evolution of ruminants or afrotherians. In summary, this gene, may have played a role in the specific evolution of marsupial species, but it is not involved in the convergent evolution of all species showing long chromosomes and low chromosome number. This is due to only those genes with high positive RERs for species of Group 1 are important to explain the convergent evolution of these species.

Another interesting gene which show one of the most significant P-values is *MROH9* (**Figure 6.2C and D**), with high positive RERs values in species like armadillo, Indian and black muntjacs, opossum, and possum. Looking to the phylogenetic tree (**Figure 6.2D**) we can see that all branches of species of Group 1 and/or their ancestral branches show high positive RERs values, indicating that *MROH9* is a gene involved in the convergent evolution of all species showing long chromosomes and low chromosome number.

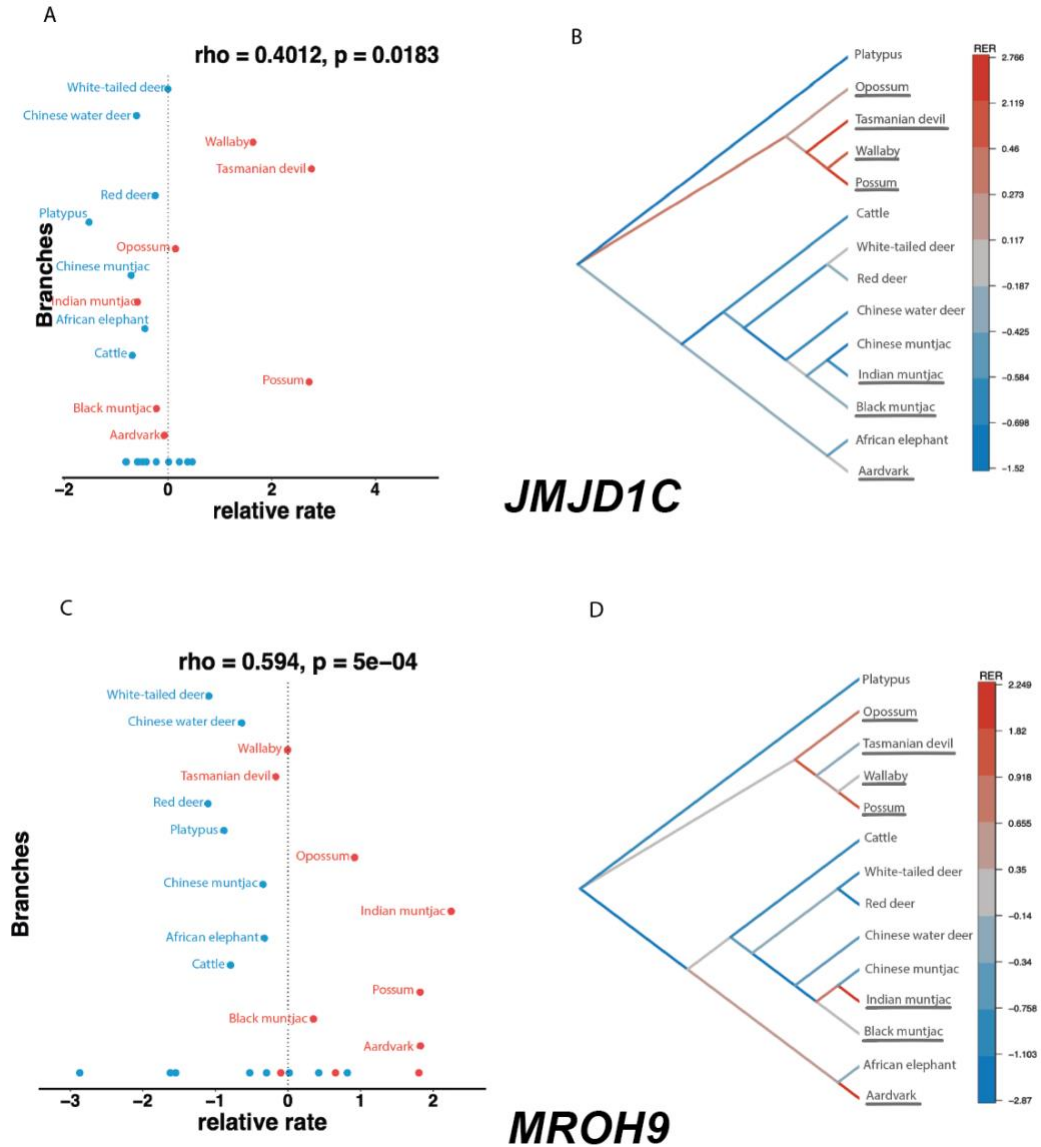


Figure 6.2. RERs of species using Group 1 as ingroup (or foreground) species. **A, C.** Estimated RERs for terminal branches (labeled with species names along the y-axis) and internal branches (plotted in a single row at the base of the y-axis). The foreground branches are highlighted in red, and the outgroup branches in blue. **B, D.** Phylogenetic tree color-coded by RERs values. Underlined species belong to Group 1.

As there were 529 genes with statistically significant RERs for Group 1, we looked for interactions between them that could reduce their number for posterior analysis. To this end, we used GeneMANIA and classified the networks found in physical interactions, co-expression, genetic interactions, co-localization or shared protein domains, among others (see **Methods section 2.11**). Furthermore, we were interested in those genes related to chromatin structure, DNA replication, meiosis, or chromosome instability, like cancer-related genes, because these might be possible factors related to the difference in chromosome size and number in mammals.

After searching for networks connecting these genes, three networks related with DNA repair and chromatin organization were selected for further analysis (see **Methods section 2.11**). The first network was classified as a physical interaction. This network, involving all 20 genes pinpointed in our RER analysis as significant for Group 1, is related to the role of the ubiquitin ligase *RNF123* targeting lamins, which are nuclear proteins important for maintaining nuclear structure and function (Khanna et al., 2018). The second network was classified as a physical interaction too, involves three genes, and is related to *VRK1*, associated with biological processes such as cell cycle, DNA repair, chromatin assembly, and RNA processing (N. Lee et al., 2017). Finally, the third network was directly classified as related to DNA mismatch repair MutS/MSH, a system in charge of the repair of replication errors (Fishel, 2015), and involves three genes.

6.2.3 Detection of positive selection

The 26 genes, with positive RERs, statistically significant, and belonging to one of these three networks, were used to detect positive selection in their evolution in mammals. This analysis was performed with CodeML in EasyCodeML, using the branches of the seven species belonging to Group 1 (opossum, Tasmanian devil, possum, wallaby, black muntjac, Indian muntjac, and armadillo), as well as using the ancestral branch of all marsupials, the ancestral branch of all muntjacs, and the ancestral branch of Indian muntjac (see **Methods Figure 2.4**). Genes with a P-value <0.05 were considered under positive selection (**Table 6.1**).

Table 6.1. EasyCodeML results, showing P-values for every gene in the specific branches of interest. Yellow numbers indicate statistically significant positive selection. Gene KIAA2026 could not be analysed. In grey-scale the three different networks.

	Marsupialia					Ruminantia				Afro-theria
Genes	Opossum	Tasmanian devil	Possum	Wallaby	Ancestral branch	Black muntjac	Indian muntjac	Ancestral branch 1	Ancestral branch 2	Aardvark
<i>AHR</i>	0.997	0.997	<0.05	<0.05	0.058	<0.05	0.991	0.983	<0.05	0.922
<i>BCORL1</i>	1.00	0.444	0.577	<0.05	<0.05	<0.05	1.00	0.999	0.938	1.00
<i>CDIPT</i>	1.00	1.00	1.00	1.00	1.00	1.00	0.998	1.00	1.00	1.00
<i>EDAR</i>	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
<i>GP1BA</i>	<0.05	0.087	0.052	<0.05	<0.05	<0.05	<0.05	<0.05	0.153	<0.05
<i>GRN</i>	0.696	0.996	<0.05	0.957	0.961	<0.05	0.998	<0.05	<0.05	0.989
<i>HACD3</i>	1.00	1.00	1.00	0.688	0.935	1.00	0.998	0.994	0.997	0.614
<i>IQSEC2</i>	0.998	<0.05	0.998	<0.05	<0.05	<0.05	0.995	0.999	<0.05	<0.05
<i>KCNS2</i>	1.00	1.00	1.00	<0.05	1.00	0.591	1.00	1.00	1.00	1.00
<i>KIAA2026</i>	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
<i>LGALSL</i>	1.00	0.710	0.999	0.999	0.432	0.256	0.457	0.998	0.971	0.078
<i>LRP2</i>	<0.05	<0.05	<0.05	<0.05	<0.05	1.00	<0.05	0.058	1.00	<0.05
<i>LTA4H</i>	<0.05	0.781	0.984	<0.05	0.997	0.996	1.00	0.997	0.997	<0.05
<i>NUDT12</i>	0.176	0.872	1.00	0.799	<0.05	<0.05	1.00	0.988	1.00	0.822
<i>PCIF1</i>	1.00	1.00	1.00	1.00	1.00	1.00	0.998	0.999	1.00	1.00
<i>PPP1R32</i>	<0.05	0.152	0.077	<0.05	<0.05	<0.05	0.551	0.999	0.994	0.236
<i>RNF123</i>	1.00	1.00	1.00	0.996	1.00	<0.05	0.998	1.00	1.00	1.00
<i>SPACT1</i>	<0.05	0.095	<0.05	0.623	1.00	1.00	0.652	1.00	2.00	<0.05
<i>TRMT6</i>	0.332	0.997	0.974	<0.05	0.085	0.987	0.995	0.871	0.824	1.00
<i>ZZEF1</i>	0.407	<0.05	0.073	<0.05	0.762	<0.05	0.996	0.998	1.00	<0.05
<i>CLIP2</i>	<0.05	0.989	<0.05	0.997	0.999	0.997	0.997	0.983	0.999	<0.05
<i>DNAH2</i>	0.998	0.994	0.998	<0.05	0.962	<0.05	0.989	0.933	<0.05	0.248
<i>VRK1</i>	0.610	1.00	1.00	0.054	0.054	0.078	1.00	0.999	0.992	<0.05
<i>MSH3</i>	0.057	<0.05	1.00	<0.05	<0.05	0.054	0.177	<0.05	<0.05	0.072
<i>MSH4</i>	0.062	1.00	1.00	1.00	1.00	0.192	0.998	1.00	0.999	1.00
<i>MSH6</i>	<0.05	0.997	0.994	<0.05	0.278	<0.05	<0.05	0.986	<0.05	0.985

Network 1: Physical interaction related to ubiquitin *RNF123* targeting lamins

Network 2: Physical interaction related to *VRK1*

Network 3: DNA mismatch repair

These results show that 19 genes under convergent evolution for the reduction in diploid number were also under positive selection, while six genes were under convergent evolution but not positive selection.

Interestingly, *GP1BA* experienced positive selection in all three lineages presenting convergent reduction of diploid numbers: (i) within Marsupialia: in opossum, wallaby, and the ancestral branch; (ii) within Ruminantia: in both muntjacs and the ancestral node of all muntjacs; and (iii) within Afrotheria: in aardvark. Overall, *GP1BA* is a good candidate to explain convergent evolution of diploid numbers in mammals, as it is also under positive selection in the three lineages. Similar cases are the genes *IQSEC2* and *LRP2*. Additionally, we compared positive selection to the RERs obtained for every gene, and in the case of *IQSEC2* (Table 6.1 and Figure 6.3), same branches experiencing positive selection show high positive RERs: Tasmanian devil, wallaby, the ancestral node of all marsupials, black muntjac, and the ancestral node of Indian and Chinese muntjacs. The only exception in this analysis is aardvark, which shows significant values of positive selection but however its RER is highly negative.

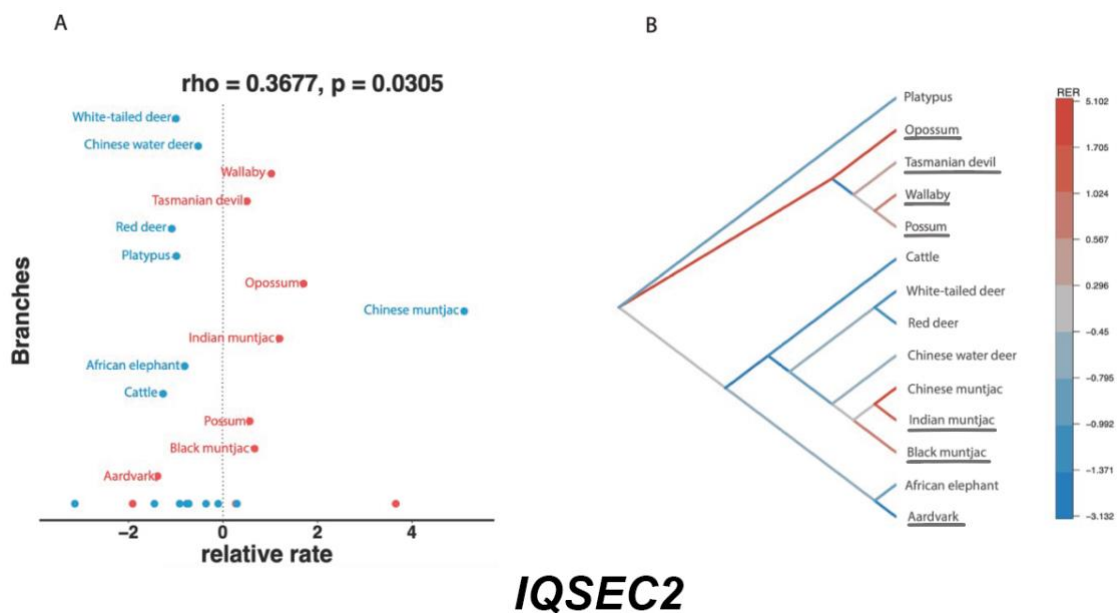


Figure 6.3. RERs of species using Group 1 as ingroup (or foreground) species. **A.** Estimated RERs for terminal branches (labeled with species names along the y-axis) and internal branches (plotted in a single row at the base of the y-axis). The foreground branches are highlighted in red, and the outgroup branches in blue. **B.** Phylogenetic tree color-coded by RERs values. Underlined species belong to Group 1.

Similarly, we also found genes under positive selection in most of the lineages with low diploid numbers, such as *AHR* or *MSH6*. Additionally, we found genes under positive selection in species of two lineages (but not including all of them or the ancestral branches), such as *BCORL1* or *GRN*, or cases with positive selection only affecting species of one of the three lineages, such as *KCNS2* or *RNF123*. For these cases, there is not enough evidence of these genes being affected by positive selection and, therefore, being related to convergent evolution of low diploid number in mammals. Contrariwise, we also found genes without positive selection in any of the species, such as *CDIPT*, *EDAR*, or *PCIF1*.

Altogether, genes related to the role of the ubiquitin ligase *RNF123* targeting lamins, and related to DNA mismatch repair MutS/MSH, have been under positive selection in the three lineages. However, genes related to some important biological processes such as the cell cycle, DNA repair, chromatin assembly, and RNA processing, present positive selection in some branches but not in all species or in all lineages under study.

6.3 Discussion

In this chapter, we proposed convergent evolution as one of the main forces causing differences in mammalian karyotypes. We used three lineages containing species with a wide range of not only diploid numbers, but also chromosome sizes: Ruminantia, Marsupialia, and Afrotheria. Together with previous publications where the ancestral karyotypes of these three groups were described (Frönicke et al., 2003; M. Farré et al., 2019; Deakin & O'Neill, 2020), in this thesis we characterised at high definition these karyotypes, with $2n = 48$, $2n = 14$, and $2n = 48$, respectively. But not only ancestral karyotypes have been defined in these lineages, most recently, the mammalian ancestor was hypothesised to comprise $2n = 20$ chromosomes (Kim et al., 2017; M. Farré et al., 2019; Damas et al., 2022). These studies identified the different CRs involved in the evolution from these ancestors to current species, located EBRs, and associated them with high rich genic regions. But it has not been until recently, that CRs that occurred during mammalian evolution have been associated to changes in chromatin configuration (Álvarez-González, Arias-Sardá, et al., 2022). As part of a collaborative paper, we described how lineage-specific evolutionary genomic reshuffling can influence patterns of higher order chromatin organization in two of the mammalian lineages studied in this thesis.

The main force shaping mammalian karyotypes is chromosome reduction, as we have seen in several species of these lineages. Due to the recent correlation found between this chromosome reduction and different chromatin configuration, we decided to search for genes accomplishing four requirements: (i) being orthologs in all species of the three lineages; (ii) being related to mechanisms associated to CRs, such as DNA replication, meiosis, DNA repair after DSBs, maintenance of chromatin structure, or genome instability; and (iii) being under convergent evolution and positive selection.

Two of the genes with higher significant P-values for being affected by the convergent evolution of reduction of the diploid number and increase in chromosome size were *JMJD1C* and *MROH9*. Interestingly, the histone demethylase *JMJD1C* is a potential druggable target of some leukaemia, which could be a useful biomarker in oncology, and it also regulates lipid synthesis-associated genes. More importantly, it has been associated not only in tumorigenesis and cancer treatments but also to genome-integrity maintenance (Watanabe et al., 2013; Y. Yang et al., 2021; Qi et al., 2022). Contrariwise, there is not much literature about *MROH9*, but it has been found to be upregulated in the pancreatic adenocarcinoma (Z. Wu et al., 2022).

Our results found genes under convergence and positive selection related to the role of the ubiquitin ligase *RNF123* targeting lamins, key nuclear proteins important for maintaining nuclear structure and function (Khanna et al., 2018). It has been reported that lamins are associated to chromosomal territories related to gene-poor chromosomes (Dechat et al., 2008) and that play important roles in the organization of nuclear functions such as DNA replication, transcription and repair, as well as in chromatin organization and gene regulation (Chaturvedi et al., 2012). Within this network, three genes were under positive selection in all species belonging to Group 1: *GP1BA*, *IQSEC2*, and *LRP2*. Of these genes, *IQSEC2* has been found overexpressed in cases of gastric cancer (Meng et al., 2020) and it has been widely studied for its role in epilepsy, intellectual disability, and autism (Brant et al., 2021). Additionally, another gene under positive selection in most of the lineages with low diploid numbers is *AHR*, which influences the major stages of tumorigenesis - initiation, promotion, progression, and metastasis (Murray et al., 2015; Z. Wang et al., 2020).

The second network we studied was only constituted by three genes: *CLIP2*, *DNAH2*, and *VKR1*. *CLIP2* is involved in DNA repair/maintenance, chromosomal instability, and metastasis in the papillary thyroid carcinoma (Selmansberger et al., 2020). In addition, another study

suggests that *DNAH2* may modulate the cell homologous recombination repair partially by increasing the ubiquitination and the enrichment to DNA damage sites of *FANCD2* in patients suffering Fanconi anaemia (Chang et al., 2021). Finally, the last gene belonging to this network is *VRK1*, which regulates the start of cell cycle, chromatin condensation in G2/M, Golgi fragmentation, Cajal body dynamics, and nuclear envelope assembly in mitosis (Campillo-Marcos et al., 2021), as well as it promotes tumour proliferation, migration, and invasion (L. Wang et al., 2021; J. Wu et al., 2022).

Moreover, the third gene network is related to DNA mismatch repair MutS/MSH, which is a system in charge of the repair of replication errors (Fishel, 2015; Lamb et al., 2022). Several studies in mice point that mutations in the DNA mismatch repair as one of the most common causes of genome instability and colorectal cancer predisposition, as well as potential targets to play a role in genome maintenance and tumour suppression (K. Lee et al., 2016).

In summary, we found convergent evolution and positive selection on genes related to chromatin structure and DNA repair, such as *JMJD1C*, *RNF123*, *CLIP2*, *DNAH2*, and *VRK1*, as well as genes related to genome instability and the development and/or progression of several tumours, such as *MROH9*, *IQSEC2*, *AHR*, and all previously mentioned genes related to DNA repair. Therefore, these genes are related to mechanisms associated to the formation of CRs, which would be the main force driving the evolution of mammalian extreme karyotypes and will confirm that the reduction of chromosome number and the increase of chromosome size found in many mammalian species has been driven by convergent evolution.

Overall, the study of convergent evolution and positive selection in extreme karyotypes has pinpointed to potential candidate genes which can be involved in their fast diploid number reduction and chromosome size increase. As this reduction in diploid number used to be correlated to a wide increase in chromosome size, the candidate genes found in this thesis, might be involved in the adaptations which allowed cells to cope with few but large chromosomes during Mammalia evolution.



7. General discussion

7.1. The importance of a good genome assembly

The contribution of this thesis to the study of genome evolution in mammals has been possible thanks to the increasing number of high-quality mammalian genomes. It is due to projects such as the Vertebrates Genome Project (VGP), whose aims are to generate high-quality, complete reference genomes for all extant vertebrate species, and to help to enable a new era of discovery across the life sciences (Rhie et al., 2021).

Through the study of 47 species, of which 36 possessed chromosome- or HiC-scaffold-level assemblies, we reconstructed eight different mammalian ancestors belonging to three different lineages: Ruminantia, Marsupialia, and Afrotheria. These reconstructions made possible the classification of all the CRs occurred from the ancestral karyotype to current species genomes, as well as the identification of EBRs and HSBs in one representative species of every lineage. The accuracy of the location and definition of all EBRs achieved in this thesis was key for their association to genes and TEs.

Finally, the availability of good-quality annotations was essential for the search of orthologous genes in 14 mammalian species, specifically those genes showing high evolutionary rates for traits related to species with low chromosome numbers. The selected genes were probed to be affected by positive selection, and subsequently, confirmed that the wide reduction in diploid number and increase of chromosome size have been originated by convergent evolution in mammals.

Over the past decade, NGS technologies have continued to evolve in order to provide genome reads as long as some entire genomes (Goodwin et al., 2016). At the same time, these new technologies have been improved in order to reduce both, the time and the cost of their use. Moreover, long-reads sequencing has completely eliminated the potential errors that could be introduced during the DNA amplification process of short-reads, as well as the required instruments have reduced their cost and are portable (Hu et al., 2021). A good genome assembly is key to perform a good genome annotation, and chromosome-level assemblies improve the detection not only of genes but also repetitive elements and other copy-number or structural variants (Rice & Green, 2019). Nowadays, these new technologies and new software are being used by international consortiums, such as the Genome 10K Project, the Earth BioGenome Project (EBP), or the Darwin Tree of Life Project (DTOL). Altogether, the availability of new

genome assemblies and the development of new algorithms will allow for the reconstruction of ancestral genomes at further evolutionary depth with even higher resolution, speeding up the research on chromosome evolution even more.

7.2 Reconstructing ancestral karyotypes and CRs

Mammalian genomes have been broadly studied during the past decades, and some studies have reconstructed different ancestors of this group combining both cytogenetics and computational approaches (Murphy et al., 2005; Ruiz-Herrera et al., 2006; Larkin et al., 2009; Lemaitre et al., 2009; Capilla et al., 2016; Kim et al., 2017; M. Farré et al., 2019; Deakin & O'Neill, 2020; Damas et al., 2022; Álvarez-González, Arias-Sardá, et al., 2022). In this thesis, we reconstructed eight different ancestral mammalian karyotypes.

For the Ruminantia lineage, we propose a new karyotype for five different ancestors: RAK, PAK, BOCE, BOMO, and BOVIN. Two of these ancestors, BOCE and BOMO, had not been reported by other studies, and they are key to understand cattle-specific (or at least Bovidae specific) rearrangements. Our proposed diploid number for these ancestors are: $2n = 48$ for RAK, and $2n = 58$ for the rest of the ancestors. For the Marsupialia lineage, we propose new karyotypes for two different ancestors: MAK and MAUK. It was the first time reporting the ancestral configuration of MAUK, which was important to determine CRs differentiating American and Australian marsupials. Both ancestors show a diploid number of $2n = 14$. Finally, for the Afrotheria lineage we proposed a new karyotype, $2n = 48$, for the oldest ancestor, AFAK. Overall, these reconstructions, are key to detect and classify CRs occurring during mammalian evolution as well as to define precisely EBRs and HSBs.

In the Ruminantia lineage, a total of 59 CRs separate RAK and PAK (12 fissions, three fusions and 44 inversions), a total of eight CRs separate PAK and BOVIN (one fusion and seven inversions), and a total of 15 CRs separate BOVIN and cattle (two fusions and 13 inversions). Previous studies determined the number of rearrangements from the ruminant to the pecoran ancestor in 11 inversions, three fusions and 12 fissions, from the pecoran to the Bovidae ancestor in three inversions and one fission, and from the Bovidae ancestor to cattle in 13 inversions (M. Farré et al., 2019). Compared to our results, we detected more inversions. In the Marsupialia lineage, a total of 14 CRs differentiate MAK and MAUK (ten inversions, two fissions, and two fusions), and a total of 25 CRs differentiate MAUK and Tasmanian devil

(one fission and 24 inversions). In the Afrotheria lineage, a total of 58 CRs were found to separate AFAK and African elephant (31 inversions, 22 fissions, and five fusions). It has been previously reported the association of inversions to phenotypic changes and adaptation in multiple species, as well as the possibility of inversions promoting more complex intrachromosomal rearrangements and translocations (Puig et al., 2015). Because of this, it is not surprising that in the three lineages inversions are the most common rearrangement.

Despite inversions being the main type of CRs in the three lineages, interchromosomal rearrangements have played an important role for the Ruminantia and Afrotheria evolution. One of the possible explanations is that the number of interchromosomal interactions observed in different taxa is not directly related to chromosome size but to chromosome position within the cell nucleus. As part of a collaborative paper (Álvarez-González, Arias-Sardá, et al., 2022), we described how lineage-specific evolutionary genomic reshuffling can influence patterns of higher order chromatin organization, using our ancestral reconstructions combined with HiC data of several mammalian species. Our data provide evidence for the existence of different chromosome folding patterns within mammals. The compartmentalization of genomes in the different layers of chromatin organization (including chromosomal territories, compartments (A and B), topologically associated domains (TADs), and looping interactions) has been well documented (Dekker et al., 2013), and how they are conserved across species is key to understand their potential evolutionary genome plasticity. In afrotherians, genomes are organised in well-defined chromosomal territories, with chromosomes showing higher interchromosomal interactions, following the well-known pattern in most mammalian species, including human and mouse. Afrotherians have their chromosomes organized in territories with small chromosomes positioned toward the centre of the cell, which is exemplified by the observation that shorter eutherian chromosomes show high interchromosomal interactions (**Figure 7.1**). Furthermore, the low inter-/intrachromosomal interaction ratios observed in African elephant and armadillo were indicative of relatively highly compacted chromatin as reflected by high distance-dependent interactions and high CTCF density. As a summary, the pattern of chromosome reshuffling observed in modern Afrotheria (which show extensive intra- and interchromosomal reorganizations compared to AFAK) can be related to chromosomes being organized into highly compacted chromosomal territories, where heterologous chromosomes are more frequently in contact. Subsequently, this would favour interchromosomal rearrangements. Importantly, chromatin intermingles at the peripheral regions of chromosomal compartments in mammals, allowing interactions between non-

homologous chromosomes, facilitating the rejoining of broken DNA ends of heterologous chromosomes. This can explain the wide number of interchromosomal rearrangements detected in this lineage (Álvarez-González, Arias-Sardá, et al., 2022). Although there is no HiC data available for ruminant species, we could assume that their chromosomes form well defined territories in the 3D nucleus, as it has been shown that most eutherian mammals have the same 3D nuclear structure as the African elephant, including human and mouse (Hoencamp et al., 2021). As such, a new model emerges whereby species characterised by chromosomal territories show primarily interchromosomal rearrangements while species with centromere clustering predominantly show intrachromosomal rearrangements (**Figure 7.1**). Contrary to other mammals, we show that marsupials display high interactions between all chromosomes, including the X chromosome. Moreover, our findings suggest that genomes organized into low diploid numbers (and hence longer chromosomes), which is the case of all marsupials, present high inter-/intrachromosomal interaction ratios, low interaction values at short and medium genomic distances, and low CTCF density, suggesting that marsupials have a distinct mechanism of chromosome folding. Moreover, long marsupial chromosomes form a “loose” distribution that extends across the nucleus, with presumably lower numbers of longer loops anchored by their centromeres that are facing toward the centre of the cell, favouring intrachromosomal rearrangements (**Figure 7.1**) (Álvarez-González, Arias-Sardá, et al., 2022).

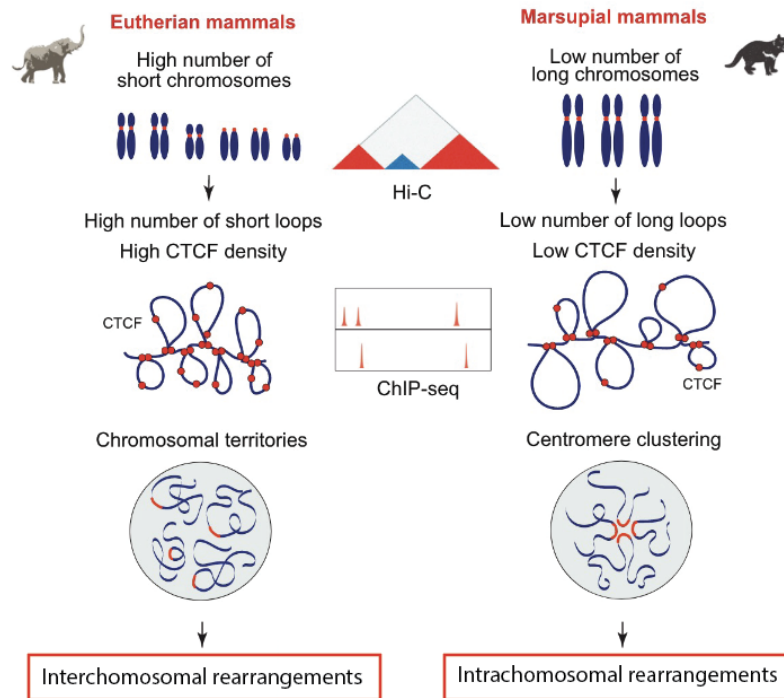


Figure 7.1. Representation of genomic architectural features detected in lineage-specific chromosomal reorganizations in the Tasmanian devil and African elephant. Figure adapted from (Álvarez-González, Arias-Sardá, et al., 2022).

7.3 Contrasting functional roles in EBRs and HSBs

Although some studies have attempted the identification of EBRs in several mammalian species (Larkin et al., 2009), others have focused specifically on cattle (M. Farré et al., 2019) while new studies have used a different approach comparing human EBRs to other species (Damas et al., 2022). In this thesis we have identified new EBRs from three species of different lineages: cattle, Tasmanian devil and African elephant.

Contrary to the expectation of the “Random Breakage Model” of chromosome rearrangements (Nadeau & Taylor, 1984), we see that in cattle and African elephant, with an average chromosome size of 87 Mb and 112 Mb, respectively, there is no association between number of EBRs and chromosome size. Instead, in Tasmanian devil, whose average chromosome size is 440 Mb, this association is highly significant ($P\text{-value} < 0.001$) and big chromosomes present more EBRs than small chromosomes. However, in Tasmanian devil chromosome X no EBRs were found, but we identified several gaps which could not be considered EBRs due to their large size. Some of them probably are EBRs that could not be well-defined because the assembly level of chromosome X is poor. One possible explanation

for these differences among lineages is the inter- and intrachromosomal ratio found in the three lineages. Inversions are the most common rearrangement in all lineages, but more importantly, they are also the main driving force originating genome reshuffling in marsupial species. Moreover, it is well known that the bigger chromosomes found in marsupials can be related to more content in TEs (Kapusta et al., 2017) which, at the same time, is one of the main factors originating CRs due to their association to genome instability triggered by failures in DNA repair mechanisms (Wicker et al., 2007; Richard et al., 2008). Altogether, it is plausible that marsupial chromosomes present more CRs, specifically inversions, than other mammalian lineages, showing a positive correlation between the number of CRs and chromosome size.

After defining EBRs, we studied whether they were located in gene-rich regions as reported in previous studies (Everts-van der Wind et al., 2004; Ma et al., 2006; Larkin et al., 2009) and also whether EBRs are located in genomic regions enriched in TEs (Kim et al., 2017; M. Farré et al., 2019; Damas et al., 2022). All three lineages shared the same results, well-defined EBRs showed a clear negative correlation with gene-rich regions, but there is an increase of genes in regions near the EBRs. This can be explained because many of the well-defined EBRs are 0-1 bp sized, making it unlikely to find genes within them. However, regions near EBRs show an enrichment in genes as previously reported.

Among the genes located within and near the EBRs of ruminant species, we found genes related to lipid metabolism (*AADACL3*), mechanisms for repairing SSBs (*APTX*), and lipid glycosylation and determination of the blood type (*ABO*). Within and near marsupial EBRs we found genes related to either chromatin configuration or DSBs repair such as *MFT2*, as well as genes related to tumoral development and progression, such as *ERCC6L* and *CCNA1*. For the Afrotheria lineage, EBRs were enriched in genes related to chromosome instability and tumorigenesis, such as *ESD*, *RHBDD2*, and *MAP9*, genes related to specific afrotherian traits, such as *ADNP2*, and genes determining the blood type (*ABO*). Comparing the genes found within and near EBRs among the three lineages, marsupials and afrotherians genes are related mostly to genome instability, ruminants and afrotherians shared the gene *ABO*, which has been probed not only to determine the blood type but also to define milk composition in bovids (Z. Liu et al., 2019). The presence of genes related to either chromatin configuration such as *MFT2* in marsupials or to SSBs/DSBs repair mechanisms such as *AADACL3* in ruminants, is related to the structural plasticity detected in lineage-specific chromosomal reorganizations (Álvarez-González, Arias-Sardá, et al., 2022). Overall, many of the genes found within and near the

EBRs are associated to lineage-specific phenotypes, and moreover, ruminants and afrotherians, which possess a very similar genome organization, even shared common genes. All in all reinforcing previous associations of EBRs and adaptive phenotypes.

Further analyses were performed to determine whether these regions near well-defined EBRs were related to a higher TE content, as recent studies have confirmed that EBRs are enriched for segmental duplications (SDs) and TEs (Damas et al., 2022). The TE enrichment in EBRs not only have been seen in mammals but also in other vertebrate lineages, such as in avian genomes (M. Farré et al., 2016). In our analysis, ruminants and afrotherians not only have an enrichment in TEs in EBRs, showed by their enrichment in LINEs-L1 and LTRs, but also they present repetitive elements specific of their lineages. In ruminants, the enrichment of elements like Bov-tA2, Bov-tA3 and other elements like Charlie elements and ERVs was found in our data in agreement to previous studies (Damas et al., 2022). Although Bov-tA2 appears negatively correlated to well-defined EBRs, it appears positively correlated in not well-defined EBRs, suggesting, as mentioned before, that the enrichment of TEs as well as of genes occurred in regions near EBRs. In afrotherians, we also reported an enrichment of afrotherian-specific TEs, such as L1-LA elements (Nishihara et al., 2005; Gilbert et al., 2008), as well as the enrichment in other elements like Charlie elements and ERVs, as reported by (Damas et al., 2022). However, other afrotherian-specific TEs reported before (Nishihara et al., 2005; Gilbert et al., 2008), the AFROSINES, showed a decrease in both well-defined and not well-defined EBRs. This could be due to AFROSINES being located in regions near EBRs but not as close as other repetitive elements. Interestingly, marsupials did not have any enrichment in TEs, which might be explained because: (i) there is no TE enrichment in this lineage; or (ii) the annotation of Tasmanian devil genome might be incomplete for repetitive regions.

Overall, we were able to define more EBRs than previously reported in cattle and trace a new EBR map for Tasmanian devil and African elephant. We classified our EBRs in ancestor- or species-specific, thanks to the high accuracy of our reconstructions. And finally, we demonstrate that regions near EBRs are enriched in gene content in all mammalian lineages, and we also demonstrate that well-defined EBRs and regions near them are enriched in TEs, some of them lineage-specific elements, in ruminant and afrotherian species.

After identifying all EBRs in cattle, Tasmanian devil, and African elephant genomes, we defined and classified all the HSBs present in these species depending on the kind of

rearrangements originating them. Several studies have shown that HSBs are enriched for evolutionary conserved sequences and genes related to basic organismal development (Larkin et al., 2009; M. Farré et al., 2016). Interestingly, only one GO term (binding) is enriched in both collinear and interchromosomal-related HSBs of all three lineages. While in cattle there is an enrichment of genes related to pheromones in collinear HSBs, in African elephant the same GO term is enriched in HSBs related to interchromosomal rearrangements. Although previous studies have reported mammalian HSBs to be enriched in genes related to developmental process, biological adhesion, and meiosis I (M. Farré et al., 2019), we only found genes related to developmental processes in Tasmanian devil, and genes related to cell adhesion in African elephant. Other studies have reported HSBs to contain a significantly higher fraction of genes involved in anatomical morphogenesis and the development of the central nervous system (Damas et al., 2022) but these GO terms were not present in our results. One of the possible explanations for not finding more common GO terms among lineages is because, as genes found in EBRs have proved to be lineage-specific, most probably every lineage possess also specific genes and gene functions in their HSBs, showing these different GO terms in every lineage. However, binding is the only GO term shared by all species and it has been found in different types of HSBs, and this could be due to the 3D nuclear organization. The study of chromatin structure and nucleosome localization have revealed epigenetic mechanisms related to specific cellular processes, and transcription factors or other regulatory proteins binding to open regions of a chromosome have been seen to directly affect intracellular gene replication and transcription and play an important role in transcriptional regulation (Ding & Zhang, 2023), which might be one of the reasons for the conservation of the binding GO term in all the lineages. Further studies of the genes related to the binding GO term shared by all lineages are needed to better understand the importance of their conservation in mammals.

Our results agree with the “Integrative Breakage Model” which explains that CRs and posterior genome reshuffling would occur in regions of the genome that physically interact inside the 3D nuclear space and whose chromatin have an accessible state promoting genome instability (M. Farré et al., 2015). Some studies showed that genome reshuffling could disrupt genes essential for survival, and subsequently be subject to purifying selection (Becker & Lenhard, 2007; Lemaitre et al., 2009; Ruiz-Herrera et al., 2010), and in these cases, EBRs would represent opportunities for the development of novel functions that may promote the adaptation of species. Here we showed that lineage-specific TEs and genes within and near the EBRs are related to lineage-specific phenotypes, suggesting their involvement in adaptation.

Intriguingly, we also see an association of EBRs and genes related to the machinery that repair DSBs, opening a new hypothesis as of whether changes in these genes might be responsible for the incorrect repair of DSBs and the formation of CRs (Brekke et al., 2022).

7.4 The role of DSBs repair mechanisms in the convergent extreme reduction of diploid number in mammals

In this thesis we have seen that mammals possess extreme variability in terms of genome size and chromosome diploid numbers. The three lineages studied here are a good example of this variability, with karyotypes ranging from $2n = 6/7$ in the Indian muntjac to $2n = 70$ in the white-tailed deer in Ruminantia, with massive chromosomes in all Marsupialia species, or with diploid numbers ranging from $2n = 20$ in the aardvark to $2n = 56$ in the African elephant in Afrotheria. Moreover, there is a clear association between chromosome number and chromosome size, which has affected genome reshuffling and organization during evolution. When looking at our ancestral reconstruction, we proposed a diploid number of $2n = 48$ for the ruminant and afrotherian ancestors and a diploid number of $2n = 14$ for the marsupial ancestor. Ruminant and afrotherian ancestral karyotypes must have gone through a great reduction or amplification of their chromosome numbers via fusions and/or fissions, which at the same time, have originated the increase or reduction on their chromosome sizes. However, the marsupial ancestor possesses less chromosomes than the ruminant and afrotherian ones, which implicates an additional scenario of chromosome number reduction lead by interchromosomal rearrangements. Seeing that these changes most likely originated independently in each lineage, we could consider this reduction of chromosome number as a convergent trait, as some studies have demonstrated for the evolution of phenotypes of marine mammals (Foote et al., 2015), or the evolution of echolocating genes in diverse mammals like bats or marine mammals (Parker et al., 2013).

Our results showed convergent evolution in several genes: *JMJD1C*, *MROH9*, *RNF123*, *IQSEC2*, *AHR*, *CLIP2*, *DNAH2*, *VRK1*, and some members of the MSH family. Of these genes: (i) *JMJD1C* not only has been associated to tumorigenesis and cancer treatments but also to genome-integrity maintenance (Watanabe et al., 2013; Y. Yang et al., 2021; Qi et al., 2022); (ii) *MROH9* is upregulated in the pancreatic adenocarcinoma (Z. Wu et al., 2022); (iii) *RNF123* is important for maintaining nuclear structure and function (Khanna et al., 2018) because its

association to nuclear organization functions such as DNA replication, transcription and repair, as well as to chromatin organization and gene regulation (Chaturvedi et al., 2012); (iv) *IQSEC2* has been found overexpressed in cases of gastric cancer (Meng et al., 2020); (v) *AHR* have been found to be associated to the major stages of tumorigenesis - initiation, promotion, progression, and metastasis (Murray et al., 2015; Z. Wang et al., 2020); (vi) *CLIP2* is involved in DNA repair/maintenance, chromosomal instability, and metastasis in the papillary thyroid carcinoma (Selmansberger et al., 2020); (vii) *DNAH2* may modulate the cell homologous recombination repair (Chang et al., 2021); (viii) *VRK1*, regulates the start of cell cycle, chromatin condensation in G2/M, and nuclear envelope assembly in mitosis (Campillo-Marcos et al., 2021), as well as being an oncogene (L. Wang et al., 2021; J. Wu et al., 2022); and (ix) the MSH family is related to the DNA mismatch repair MutS/MSH, a system in charge of the repair of replication errors (Fishel, 2015; Lamb et al., 2022). In summary, we targeted and found genes related to mechanisms associated to the formation of CRs, such as genome instability (related to tumour-related environments), chromatin organization, and DNA repair mechanisms, in species from the three lineages that have been gone through a reduction in chromosome number.

Overall, our results suggested that different karyotypes in mammals were produced by convergent evolution and that changes in chromosome number involve genes related to chromatin configuration and genome instability. Moreover, these genes associated to diploid number reduction in mammals presented functions related to mechanisms of repairing DSBs, also shown in EBRs, and which have been proposed by some studies as drivers of karyotypic evolution (Brekke et al., 2022). Some studies hypothesised the limits in mitotic and meiotic processes that species must overcome for maintaining extremely long chromosome sizes (Schubert & Oud, 1997), and how a uniform chromosome length may facilitate the cell achieving synchronized DNA replication time with the same number of replication forks, correct chromosome configuration on equatorial plate, and accurate migration of homologous chromosomes or sister chromatids to opposite poles (X. Li et al., 2011). Overall, our results highlighted genes both potentially involved in CR formation as well as genes that might have evolved once these large chromosomes were formed. With this in mind, we suggest that DNA repair mechanisms during the replication process are fundamental for the maintenance of long chromosome karyotypes and, therefore, chromosome number and size variation is a vital factor to consider in cell biology and evolution.

7.5 Future directions

The findings from this thesis pose new questions and open up new avenues of research:

1. Does our model associate chromosome positioning in the 3D nucleus and type of chromosome rearrangement hold for other mammals? Or even other animal clades? Generating new ancestral karyotype reconstructions for other clades, as well as HiC-data will be key to understand how chromatin organization affects patterns of genome reshuffling and type of rearrangement.
2. What is the exact relationship of TEs and EBRs? Studies on the potential effects of TEs in gene expression and regulation within EBRs are needed, as well as the comparison of these effects within and between lineages.
3. What role do genes under convergence and positive selection have in diploid number reduction? Our data identified for the first time genes involved in DNA repair and evolving convergently with the diploid number reduction. However, it is very early days in this field, and as such our data open doors to study potential genes related to extreme differences in karyotypes in mammals, and the association of CRs to TEs, genome instability, and chromatin interaction within the 3D nucleus.



8. Bibliography

- Abondio, P., Cilli, E., & Luiselli, D. (2022). Inferring signatures of positive selection in whole-genome sequencing data: An overview of haplotype-based methods. *Genes*, 13(5), 926.
- Agar, W. (1923). The male meiotic phase in two genera of marsupials (*Macropus* and *Petauroides*). *Q. J. Microsc. Sci.*, 67, 183–202.
- AL-Eitan, L. N., Rababa'h, D. M., Alghamdi, M. A., & Khasawneh, R. H. (2019). Genetic association of XRCC5 gene polymorphisms with breast cancer among Jordanian women. *OncoTargets and Therapy*, Volume 12, 7923–7928.
- Álvarez-González, L., Arias-Sardá, C., Montes-Espuña, L., Marín-Gual, L., Vara, C., Lister, N. C., Cuartero, Y., Garcia, F., Deakin, J., Renfree, M. B., Robinson, T. J., Martí-Renom, M. A., Waters, P. D., Farré, M., & Ruiz-Herrera, A. (2022). Principles of 3D chromosome folding and evolutionary genome reshuffling in mammals. *Cell Reports*, 41(12), 111839.
- Álvarez-González, L., Burden, F., Doddamani, D., Malinverni, R., Leach, E., Marín-García, C., Marín-Gual, L., Gubern, A., Vara, C., Paytuví-Gallart, A., Buschbeck, M., Ellis, P. J. I., Farré, M., & Ruiz-Herrera, A. (2022). 3D chromatin remodelling in the germ line modulates genome evolutionary plasticity. *Nature Communications*, 13(1), 2608.
- Archer, M., Arena, R., Bassarova, M., Black, K., Brammall, J., Cooke, B., Creaser, P., Crosby, K., Gillespie, A., Godthelp, H., Gott, M., Kear, B., Krikman, A., Mackness, B., Muirhead, J., Musser, A., Myers, T., Pledge, N., & Wroe, S. (1999). The evolutionary history and diversity of Australian mammals. *Australian Mammalogy*, 21(1), 1.
- Armour, J. A. L. (2006). Tandemly repeated DNA: Why should anyone care? *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 598(1–2), 6–14.
- Armstrong, J., Fiddes, I. T., Diekhans, M., & Paten, B. (2019). Whole-genome alignment and comparative annotation. *Annual Review of Animal Biosciences*, 7(1), 41–64.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: Tool for the unification of biology. *Nature Genetics*, 25(1), 25–29.
- Baeuerle, P. A., & Wesche, H. (2022). T-cell-engaging antibodies for the treatment of solid tumors: Challenges and opportunities. *Current Opinion in Oncology*, 34(5), 552–558.
- Bailey, J. A., & Eichler, E. E. (2006). Primate segmental duplications: Crucibles of evolution, diversity and disease. *Nature Reviews Genetics*, 7(7), 552–564.
- Balmus, G., Trifonov, V. A., Biltueva, L. S., O'Brien, P. C. M., Alkalaeva, E. S., Fu, B., Skidmore, J. A., Allen, T., Graphodatsky, A. S., Yang, F., & Ferguson-Smith, M. A. (2007). Cross-species chromosome painting among camel, cattle, pig and human: Further insights into the putative Cetartiodactyla ancestral karyotype. *Chromosome Research*, 15(4), 499–514.
- Baudoin, N. C., & Bloomfield, M. (2021). Karyotype aberrations in action: The evolution of cancer genomes and the tumor microenvironment. *Genes*, 12(4), 558.

- Becker, T. S., & Lenhard, B. (2007). The random versus fragile breakage models of chromosome evolution: A matter of resolution. *Molecular Genetics and Genomics*, 278(5), 487–491.
- Beckmann, J. S., Estivill, X., & Antonarakis, S. E. (2007). Copy number variants and genetic traits: Closer to the resolution of phenotypic to genotypic variability. *Nature Reviews Genetics*, 8(8), 639–646.
- Benoit, J., Legendre, L. J., Tabuce, R., Obada, T., Maraescul, V., & Manger, P. (2019). Brain evolution in Proboscidea (Mammalia, Afrotheria) across the Cenozoic. *Scientific Reports*, 9(1), 9323.
- Biémont, C., & Vieira, C. (2006). Junk DNA as an evolutionary force. *Nature*, 443(7111), 521–524.
- Blanc, J. J., Barnes, R. F. W., Craig, G. C., Dublin, H. T., Thouless, C. R., Douglas-Hamilton, I., & Hart, J. A. (2003). *African elephant status report 2002: An update from the African elephant database*. IUCN.
- Block, M. (1960). Wound Healing in the New-Born Opossum (*Didelphis virginianam*). *Nature*, 187(4734), 340–341.
- Böhne, A., Brunet, F., Galiana-Arnoux, D., Schultheis, C., & Volff, J.-N. (2008). Transposable elements as drivers of genomic and biological diversity in vertebrates. *Chromosome Research*, 16(1), 203–215.
- Bourque, G., & Pevzner, P. A. (2002). Genome-Scale Evolution: Reconstructing Gene Orders in the Ancestral Species. *Genome Research*, 12, 26–36.
- Bourque, G., Zdobnov, E. M., Bork, P., Pevzner, P. A., & Tesler, G. (2005). Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome Research*, 15(1), 98–110.
- Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M., & Sherlock, G. (2004). GO::TermFinder—Open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, 20(18), 3710–3715.
- Brant, B., Stern, T., Shekhidem, H. A., Mizrahi, L., Rosh, I., Stern, Y., Ofer, P., Asleh, A., Umanah, G. K. E., Jada, R., Levy, N. S., Levy, A. P., & Stern, S. (2021). IQSEC2 mutation associated with epilepsy, intellectual disability, and autism results in hyperexcitability of patient-derived neurons and deficient synaptic transmission. *Molecular Psychiatry*, 26(12), 7498–7508.
- Brekke, T. D., Papadopoulos, A. S. T., Swain, M. T., & Mulley, J. F. (2022). *Sequence features do not drive karyotypic evolution: What are the missing correlates of genome evolution?* [Preprint].
- Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12), 1213–1218.
- Burgin, C. J., Colella, J. P., Kahn, P. L., & Upham, N. S. (2018). How many species of mammals are there? *Journal of Mammalogy*, 99(1), 1–14.

- Burr, M. L., Sparbier, C. E., Chan, K. L., Chan, Y.-C., Kersbergen, A., Lam, E. Y. N., Azidis-Yates, E., Vassiliadis, D., Bell, C. C., Gilan, O., Jackson, S., Tan, L., Wong, S. Q., Hollizeck, S., Michalak, E. M., Siddle, H. V., McCabe, M. T., Prinjha, R. K., Guerra, G. R., ... Dawson, M. A. (2019). An Evolutionarily Conserved Function of Polycomb Silences the MHC Class I Antigen Presentation Pathway and Enables Immune Evasion in Cancer. *Cancer Cell*, 36(4), 385–401.
- Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., Corvelo, A., Clarke, W. E., Musunuri, R., Nagulapalli, K., Fairley, S., Runnels, A., Winterkorn, L., Lowy, E., Human Genome Structural Variation Consortium, Flicek, Germer, G., Brand H., Hall, I. M., Talkowski, M. E., Narzisi, G., Zody, M. C. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* 185(18): 3426–3440.
- Callaway, E. (2021). Mammoth genomes shatter record for oldest ancient DNA. *Nature*, 590, 537–538.
- Campillo-Marcos, I., García-González, R., Navarro-Carrasco, E., & Lazo, P. A. (2021). The human VRK1 chromatin kinase in cancer biology. *Cancer Letters*, 503, 117–128.
- Cantalapiedra, J. L., FitzJohn, R. G., Kuhn, T. S., Fernández, M. H., DeMiguel, D., Azanza, B., Morales, J., & Mooers, A. Ø. (2014). Dietary innovations spurred the diversification of ruminants during the Cenozoic. *Proceedings of the Royal Society B: Biological Sciences*, 281(1776), 20132746.
- Canzonieri, R., Rabassa, M., Gurruchaga, A., Ferretti, V., Palma, S., Isla-Larrain, M., Croce, M., Lacunza, E., & Abba, M. (2018). Alternative splicing variant of RHBDD2 is associated with cell stress response and breast cancer progression. *Oncology Reports*, 40:909-915.
- Cao, J., Dong, J., Wang, Y., & Chen, Y. (2019). The expressions of DNA methyltransferase 1 (DNMT1) and cyclin A1 (CCNA1) in cervical carcinogenesis. *Int J Clin Exp Pathol*, 12(1), 40–49.
- Capilla, L., Sánchez-Guillén, R., Farré, M., Paytuví-Gallart, A., Malinverni, R., Ventura, J., Larkin, D., & Ruiz-Herrera, A. (2016). Mammalian comparative genomics reveals genetic and epigenetic features associated with genome reshuffling in Rodentia. *Genome Biology and Evolution*, 8(12):3703–3717.
- Carter, A. M., Enders, A. C., Künzle, H., Oduor-Okelo, D., & Vogel, P. (2004). Placentation in species of phylogenetic importance: The Afrotheria. *Animal Reproduction Science*, 82–83, 35–48.
- Caspersson, T., Zech, L., & Johansson, C. (1970). Analysis of human metaphase chromosome set by aid of DNA-binding fluorescent agents. *Exp Cell Res*, 62, 490–492.
- Chang, L., Gao, X., Wang, Y., Huang, C., Gao, M., Wang, X., Liu, C., Wu, W., An, W., Wan, Y., Zhang, A., Zhang, Y., Yuan, W., & Zhu, X. (2021). DNAH2 facilitates the homologous recombination repair of Fanconi anemia pathway through modulating FANCD2 ubiquitination. *Blood Science*, 3(3), 71–77.
- Chaturvedi, P., Khanna, R., & Parnaik, V. K. (2012). Ubiquitin ligase RNF123 mediates degradation of heterochromatin protein 1 α and β in Lamin A/C knock-down cells. *PLOS ONE*, 7(10), e47558.

- Chen, L., Qiu, Q., Jiang, Y., Wang, K., Lin, Z., Li, Z., Bibi, F., Yang, Y., Wang, J., Nie, W., Su, W., Liu, G., Li, Q., Fu, W., Pan, X., Liu, C., Yang, J., Zhang, C., Yin, Y., Wang, W. (2019). Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. *Science*, 364(6446).
- Chen, Y., Zhou, S., Wan, K., Yu, L., Zhao, C., Deng, H., Ou, Q., Qin, J., Hu, J., & Hou, Z. (2022). RIOK1 mediates p53 degradation and radioresistance in colorectal cancer through phosphorylation of G3BP2. *Oncogene*, 41(25), 3433–3444.
- Chikina, M., Robinson, J. D., & Clark, N. L. (2016). Hundreds of genes experienced convergent shifts in selective pressure in marine mammals. *Molecular Biology and Evolution*, 33(9), 2182–2192.
- Claussen, U. (2005). Chromosomics. *Cytogenetic and Genome Research*, 111(2), 101–106. <https://doi.org/10.1159/000086377>
- Cordaux, R., & Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, 10(10), 691–703.
- Couvreur, T. L. P., Dauby, G., Blach-Overgaard, A., Deblauwe, V., Dessein, S., Droissart, V., Hardy, O. J., Harris, D. J., Janssens, S. B., Ley, A. C., Mackinder, B. A., Sonké, B., Sosef, M. S. M., Stévant, T., Svenning, J., Wieringa, J. J., Faye, A., Missoup, A. D., Tolley, K. A., Sepulchre, P. (2021). Tectonics, climate and the diversification of the tropical African terrestrial flora and fauna. *Biological Reviews*, 96(1), 16–51.
- Cremer, T., & Cremer, C. (2006). Rise, fall and resurrection of chromosome territories: A historical perspective Part II. Fall and resurrection of chromosome territories during the 1950s to 1980s. Part III. Chromosome territories and the functional nuclear architecture: Experiments and models from the 1990s to the present. *Eur. J. Histochem*, 50, 223–272.
- Cremer, T., & Cremer, M. (2010). Chromosome Territories. *Cold Spring Harbor Perspectives in Biology*, 2(3), a003889–a003889.
- Cunningham, F., Allen, J. E., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Austine-Orimoloye, O., Azov, A. G., Barnes, I., Bennett, R., Berry, A., Bhai, J., Bignell, A., Billis, K., Boddu, S., Brooks, L., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., ... Flicek, P. (2022). Ensembl 2022. *Nucleic Acids Research*, 50(D1), D988–D995.
- Dagilis, A. J. (2022). What inversion lengths can tell us about their evolution. *Molecular Ecology*, 31(13), 3513–3515.
- Damas, J., Corbo, M., Kim, J., Turner-Maier, J., Farré, M., Larkin, D. M., Ryder, O. A., Steiner, C., Houck, M. L., Hall, S., Shiue, L., Thomas, S., Swale, T., Daly, M., Korlach, J., Uliano-Silva, M., Mazzoni, C. J., Birren, B. W., Genereux, D. P., ... Zhang, X. (2022). Evolution of the ancestral mammalian karyotype and syntenic regions. *Proceedings of the National Academy of Sciences*, 119(40), e2209139119.
- Damas, J., O'Connor, R., Farré, M., Lenis, V. P. E., Martell, H. J., Mandawala, A., Fowler, K., Joseph, S., Swain, M. T., Griffin, D. K., & Larkin, D. M. (2017). Upgrading short-read animal genome

- assemblies to chromosome level using comparative genomics and a universal probe set. *Genome Research*, 27(5), 875–884.
- Darwin, C. (1859). *On the origin of species by means of natural selection, or, the preservation of favoured races in the struggle for life*. J. Murray.
- Davis, E. B., Brakora, K. A., & Lee, A. H. (2011). Evolution of ruminant headgear: A review. *Proceedings of the Royal Society B: Biological Sciences*, 278(1720), 2857–2865.
- Dayhoff, M. O., & Ledley, R. S. (1962). Comprotein: A computer program to aid primary protein structure determination. *ACM Journal*, 262–274.
- De Grouchy, J., Turleau, C., Roubin, M., & Klein, M. (1972). Karyotypic evolution of man and chimpanzees. A comparative study of band topographies after controlled denaturalization. *Ann Génétique*, 15, 79–84.
- Deakin, J. E., Delbridge, M. L., Koina, E., Harley, N., Alsop, A. E., Wang, C., Patel, V. S., & Graves, J. A. (2013). Reconstruction of the ancestral marsupial karyotype from comparative gene maps. *BMC Evolutionary Biology*, 13(1), 258.
- Deakin, J. E., & O'Neill, R. J. (2020). Evolution of marsupial genomes. *Annual Review of Animal Biosciences*, 8(1), 25–45.
- Deakin, J. E., Potter, S., O'Neill, R., Ruiz-Herrera, A., Cioffi, M. B., Eldridge, M. D. B., Fukui, K., Marshall Graves, J. A., Griffin, D., Grutzner, F., Kratochvíl, L., Miura, I., Rovatsos, M., Srikulnath, K., Wapstra, E., & Ezaz, T. (2019). Chromosomics: Bridging the gap between genomes and chromosomes. *Genes*, 10(8), 627.
- Dechat, T., Pfliegerhaa, K., Sengupta, K., Shimi, T., Shumaker, D. K., Solimando, L., & Goldman, R. D. (2008). Nuclear lamins: Major factors in the structural organization and function of the nucleus and chromatin. *Genes & Development*, 22(7), 832–853.
- Deininger, P. L., & Batzer, M. A. (1999). Alu repeats and human disease. *Molecular Genetics and Metabolism*, 67(3), 183–193.
- Dekker, J., Marti-Renom, M. A., & Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes: Interpreting chromatin interaction data. *Nature Reviews Genetics*, 14(6), 390–403.
- Delsuc, F., Vizcaíno, S. F., & Douzery, E. J. (2004). Influence of Tertiary paleoenvironmental changes on the diversification of South American mammals: A relaxed molecular clock study within xenarthrans. *BMC Evolutionary Biology*, 13.
- Demment, M. W., & Van Soest, P. J. (1985). A nutritional explanation for body-size patterns of ruminant and nonruminant herbivores. *The American Naturalist*, 125(5), 641–672.
- Derrien, T., Andre, C., Galibert, F., & Hitte, C. (2007). AutoGRAPH: An interactive web server for automating and visualizing comparative genome maps. *Bioinformatics*, 23(4), 498–499.
- Dessimoz, C., Zoller, S., Manousaki, T., Qiu, H., Meyer, A., & Kuraku, S. (2011). Comparative genomics approach to detecting split-coding regions in a low-coverage genome: Lessons from

- the chimaera *Callorhynchus milii* (Holocephali, Chondrichthyes). *Briefings in Bioinformatics*, 12(5), 474–484.
- Ding, T., & Zhang, H. (2023). Novel biological insights revealed from the investigation of multiscale genome architecture. *Computational and Structural Biotechnology Journal*, 21, 312–325.
- Domning, D. P. (2001). The earliest known fully quadrupedal sirenian. *Nature*, 413(6856), 625–627.
- Douady, C. J., Chatelier, P. I., Madsen, O., de Jong, W. W., Catzeflis, F., Springer, M. S., & Stanhope, M. J. (2002). Molecular phylogenetic evidence confirming the Eulipotyphla concept and in support of hedgehogs as the sister group to shrews☆. *Molecular Phylogenetics and Evolution*, 25(1), 200–209.
- Duchêne, D. A., Bragg, J. G., Duchêne, S., Neaves, L. E., Potter, S., Moritz, C., Johnson, R. N., Ho, S. Y. W., & Eldridge, M. D. B. (2018). Analysis of phylogenomic tree space resolves relationships among marsupial families. *Systematic Biology*, 67(3), 400–412.
- Ducos, E., Vergès, V., Dugé de Bernonville, T., Blanc, N., Giglioli-Guivarc'h, N., & Dutilleul, C. (2017). Remarkable evolutionary conservation of antiobesity ADIPOSE/WDTC1 homologs in animals and plants. *Genetics*, 207(1), 153–162.
- Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., Shamim, M. S., Machol, I., Lander, E. S., Aiden, A. P., & Aiden, E. L. (2017). De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, 356(6333), 92–95.
- Dutrillaux, B., & Couturier, J. (1981). The ancestral karyotype of platyrrhine monkeys. *Cytogenetic and Genome Research*, 30(4), 232–242.
- Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5), 1792–1797.
- Eizirik, E. (2001). Molecular dating and biogeography of the early placental mammal radiation. *Journal of Heredity*, 92(2), 212–219.
- El Aliani, A., El-Abid, H., El Mallali, Y., Attaleb, M., Ennaji, M. M., & El Mzibri, M. (2021). Association between gene promoter methylation and cervical cancer development: Global distribution and a meta-analysis. *Cancer Epidemiology, Biomarkers & Prevention*, 30(3), 450–459.
- Elsik, C. G., The Bovine Genome Sequencing and Analysis Consortium, Tellam, R. L., Worley, K. C., Gibbs, R. A., Muzny, D. M., Weinstock, G. M., Adelson, D. L., Eichler, E. E., Elnitski, L., Guigó, R., Hamernik, D. L., Kappes, S. M., Lewin, H. A., Lynn, D. J., Nicholas, F. W., Reymond, A., Rijnkels, M., Skow, L. C., Zhao, F.-Q. (2009). The genome sequence of taurine cattle: A window to ruminant biology and evolution. *Science*, 324(5926), 522–528.
- Emms, D. M., & Kelly, S. (2015). OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, 16(1), 157.

- Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20(1), 238.
- Everts-van der Wind, A., Kata, S. R., Band, M. R., Rebeiz, M., Larkin, D. M., Everts, R. E., Green, C. A., Liu, L., Natarajan, S., Goldammer, T., Lee, J. H., McKay, S., Womack, J. E., & Lewin, H. A. (2004). A 1463 gene cattle–human comparative map with anchor points defined by human genome sequence coordinates. *Genome Research*, 14(7), 1424–1437.
- Faria, R., Johannesson, K., Butlin, R. K., & Westram, A. M. (2019). Evolving inversions. *Trends in Ecology & Evolution*, 34(3), 239–248.
- Faria, R., & Navarro, A. (2010). Chromosomal speciation revisited: Rearranging theory with pieces of evidence. *Trends in Ecology & Evolution*, 25(11), 660–669.
- Farré, M., Bosch, M., López-Giráldez, F., Ponsà, M., & Ruiz-Herrera, A. (2011). Assessing the role of Tandem Repeats in shaping the genomic architecture of great apes. *PLOS ONE*, 6(11), e27239.
- Farré, M., Kim, J., Proskuryakova, A. A., Zhang, Y., Kulemzina, A. I., Li, Q., Zhou, Y., Xiong, Y., Johnson, J. L., Perelman, P. L., Johnson, W. E., Warren, W. C., Kukekova, A. V., Zhang, G., O’Brien, S. J., Ryder, O. A., Graphodatsky, A. S., Ma, J., Lewin, H. A., & Larkin, D. M. (2019). Evolution of gene regulation in ruminants differs between evolutionary breakpoint regions and homologous synteny blocks. *Genome Research*, 29(4), 576–589.
- Farré, M., Micheletti, D., & Ruiz-Herrera, A. (2013). Recombination Rates and Genomic Shuffling in Human and Chimpanzee—A New Twist in the Chromosomal Speciation Theory. *Molecular Biology and Evolution*, 30(4), 853–864.
- Farré, M., Narayan, J., Slavov, G. T., Damas, J., Auvil, L., Li, C., Jarvis, E. D., Burt, D. W., Griffin, D. K., & Larkin, D. M. (2016). Novel insights into chromosome evolution in birds, archosaurs, and reptiles. *Genome Biology and Evolution*, 8(8), 2442–2451.
- Farré, M., Robinson, T. J., & Ruiz-Herrera, A. (2015). An Integrative Breakage Model of genome architecture, reshuffling and evolution: The *Integrative Breakage Model* of genome evolution, a novel multidisciplinary hypothesis for the study of genome plasticity. *BioEssays*, 37(5), 479–488.
- Farré, X., Molina, R., Barteri, F., Timmers, P. R. J. H., Joshi, P. K., Oliva, B., Acosta, S., Esteve-Altava, B., Navarro, A., & Muntané, G. (2021). Comparative analysis of mammal genomes unveils key genomic variability for human life span. *Mol Bio Evol*, 38(11), 4948–4961.
- Feng, H., Zhu, M., Zhang, R., Wang, Q., Li, W., Dong, X., Chen, Y., Lu, Y., Liu, K., Lin, B., Guo, J., & Li, M. (2019). GATA5 inhibits hepatocellular carcinoma cells malignant behaviours by blocking expression of reprogramming genes. *Journal of Cellular and Molecular Medicine*, 23(4), 2536–2548.
- Ferguson-Smith, M. A., & Trifonov, V. (2007). Mammalian karyotype evolution. *Nature Reviews Genetics*, 8(12), 950–962.

- Fernández, M. H., & Vrba, E. S. (2005). A complete estimate of the phylogenetic relationships in Ruminantia: A dated species-level supertree of the extant ruminants. *Biological Reviews*, 80(2), 269–302.
- Ferretti, V., Canzoneri, R., Palma, S., Lacunza, E., Aldaz, C., & Abba, M. (2021). RHBDD2-WWOX protein interaction during proliferative and differentiated stages in normal and breast cancer cells. *Oncology Reports*, 46(2), 157.
- Fishel, R. (2015). Mismatch Repair. *Journal of Biological Chemistry*, 290(44), 26395–26403.
- Foote, A. D., Liu, Y., Thomas, G. W. C., Vinař, T., Alföldi, J., Deng, J., Dugan, S., van Elk, C. E., Hunter, M. E., Joshi, V., Khan, Z., Kovar, C., Lee, S. L., Lindblad-Toh, K., Mancina, A., Nielsen, R., Qin, X., Qu, J., Raney, B. J., Gibbs, R. A. (2015). Convergent evolution of the genomes of marine mammals. *Nature Genetics*, 47(3), 272–275.
- Frönicke, L., & Wienberg, J. (2001). Comparative chromosome painting defines the high rate of karyotype changes between pigs and bovids. *Mammalian Genome*, 12(6), 442–449.
- Frönicke, L., Wienberg, J., Stone, G., Adams, L., & Stanyon, R. (2003). Towards the delineation of the ancestral eutherian genome organization: Comparative genome maps of human and the African elephant (*Loxodonta africana*) generated by chromosome painting. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1522), 1331–1340.
- Fuller, Z. L., Haynes, G. D., Richards, S., & Schaeffer, S. W. (2016). Genomics of natural populations: How differentially expressed genes shape the evolution of chromosomal inversions in *Drosophila pseudoobscura*. *Genetics*, 204(1), 287–301.
- Gao, F., Chen, C., Arab, D. A., Du, Z., He, Y., & Ho, S. Y. W. (2019). EasyCodeML: A visual tool for analysis of selection using CodeML. *Ecology and Evolution*, 9(7), 3891–3898.
- Gauthier, J., Vincent, A. T., Charette, S. J., & Derome, N. (2019). A brief history of bioinformatics. *Briefings in Bioinformatics*, 20(6), 1981–1996.
- Gayon, J. (2016). From Mendel to epigenetics: History of genetics. *Comptes Rendus Biologies*, 339(7–8), 225–230.
- Genome 10K Community of Scientists. (2009). Genome 10K: A proposal to obtain whole-genome sequence for 10 000 vertebrate species. *Journal of Heredity*, 100(6), 659–674.
- Gilbert, C., Pace, J. K., & Waters, P. D. (2008). Target site analysis of RTE1_LA and its AfroSINE partner in the elephant genome. *Gene*, 425(1–2), 1–8.
- Glas, R. (1999). Cross-species chromosome painting between human and marsupial directly demonstrates the ancient region of the mammalian X. *Mammalian Genome*, 10(11), 1115–1116.
- Goldman, A. D., & Landweber, L. F. (2016). What Is a Genome? *PLOS Genetics*, 12(7), e1006181.
- Goodstadt, L., Heger, A., Webber, C., & Ponting, C. P. (2007). An analysis of the gene complement of a marsupial, *Monodelphis domestica*: Evolution of lineage-specific genes and giant chromosomes. *Genome Research*, 17(7), 969–981.

- Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6), 333–351.
- Graphodatsky, A. S., Trifonov, V. A., & Stanyon, R. (2011). The genome diversity and karyotype evolution of mammals. *Molecular Cytogenetics*, 4(1), 22.
- Graphodatsky, A. S., Yang, F., Dobigny, G., Romanenko, S. A., Biltueva, L. S., Perelman, P. L., Beklemisheva, V. R., Alkalaeva, E. Z., Serdukova, N. A., Ferguson-Smith, M. A., Murphy, W. J., & Robinson, T. J. (2008). Tracking genome organization in rodents by Zoo-FISH. *Chromosome Research*, 16(2), 261–274.
- Green, B., Griffiths, M., & MC Leckie, R. (1983). Qualitative and Quantitative Changes in Milk Fat during Lactation in the Tammar Wallaby (*Macropus eugenii*). *Australian Journal of Biological Sciences*, 36(6), 455.
- Griffin, D. K., Robertson, L. B., Tempest, H. G., Vignal, A., Fillon, V., Crooijmans, R. P., Groenen, M. A., Deryusheva, S., Gaginskaya, E., Carré, W., Waddington, D., Talbot, R., Völker, M., Masabanda, J. S., & Burt, D. W. (2008). Whole genome comparative studies between chicken and turkey and their implications for avian genome evolution. *BMC Genomics*, 9(1), 168.
- Groenen, M. A. M., Archibald, A. L., Uenishi, H., Tuggle, C. K., Takeuchi, Y., Rothschild, M. F., Rogel-Gaillard, C., Park, C., Milan, D., Megens, H.-J., Li, S., Larkin, D. M., Kim, H., Frantz, L. A. F., Caccamo, M., Ahn, H., Aken, B. L., Anselmo, A., Anthon, C., Schook, L. B. (2012). Analyses of pig genomes provide insight into porcine demography and evolution. *Nature*, 491(7424), 393–398.
- Gu, Z., Li, Y., Yang, X., Yu, M., Chen, Z., Zhao, C., Chen, L., & Wang, L. (2018). Overexpression of CLC-3 is regulated by XRCC5 and is a poor prognostic biomarker for gastric cancer. *Journal of Hematology & Oncology*, 11(1), 115.
- Guttenbach, M., Nanda, I., Feichtinger, W., Masabanda, J. S., Griffin, D. K., & Schmid, M. (2003). Comparative chromosome painting of chicken autosomal paints 1–9 in nine different bird species. *Cytogenetic and Genome Research*, 103(1–2), 173–184.
- Hagen, J. B. (2000). The origins of bioinformatics. *Nature Reviews Genetics*, 1(3), 231–236.
- Han, M., Li, F., Zhang, Y., Dai, P., He, J., Li, Y., Zhu, Y., Zheng, J., Huang, H., Bai, F., & Gao, D. (2022). FOXA2 drives lineage plasticity and KIT pathway activation in neuroendocrine prostate cancer. *Cancer Cell*, 40(11), 1306–1323.
- Harris, R. S. (2007). *Improved pairwise alignment of genomic DNA*. The Pennsylvania State University.
- Hayman, D., & Martin, P. (1969). Cytogenetics of marsupials. In *Comparative Mammalian Cytogenetics* (pp. 191–217). New York: Springer.
- Hedges, S. B. (2001). Afrotheria: Plate tectonics meets genomics. *Proceedings of the National Academy of Sciences*, 98(1), 1–2.
- Heger, A., Webber, C., Goodson, M., Ponting, C. P., & Lunter, G. (2013). GAT: A simulation framework for testing the association of genomic intervals. *Bioinformatics*, 29(16), 2046–2048.

- Hejase, H. A., Dukler, N., & Siepel, A. (2020). From summary statistics to gene trees: Methods for inferring positive selection. *Trends in Genetics*, 36(4), 243–258.
- Helgen, K. M. (2011). The mammal family tree. *Science*, 334(6055), 458–459.
- Heritage, S., & Seiffert, E. R. (2022). Total evidence time-scaled phylogenetic and biogeographic models for the evolution of sea cows (Sirenia, Afrotheria). *PeerJ*, 10, e13886.
- Hieter, P., & Boguski, M. (1997). Functional Genomics: It's All How You Read It. *Science*, 278(5338), 601–602.
- Hoencamp, C., Dudchenko, O., Elbatsh, A. M. O., Brahmachari, S., & Raaijmakers, J. A. (2021). 3D genomics across the tree of life identifies condensin II as a determinant of architecture type. *Science*, 372(6545), 984–989.
- Hsieh, T.-H. S., Cattoglio, C., Slobodyanyuk, E., Hansen, A. S., Rando, O. J., Tjian, R., & Darzacq, X. (2020). Resolving the 3D landscape of transcription-linked mammalian chromatin folding. *Molecular Cell*, 78(3), 539–553.e8.
- Hu, T., Chitnis, N., Monos, D., & Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Human Immunology*, 82(11), 801–811.
- Huang, X., Jiang, L., Lu, S., Yuan, M., Lin, H., Li, B., Wen, Z., & Zhong, Y. (2022). Overexpression of ERCC6L correlates with poor prognosis and confers malignant phenotypes of lung adenocarcinoma. *Oncology Reports*, 48(1), 131.
- Huber, D., Voith von Voithenberg, L., & Kaigala, G. V. (2018). Fluorescence in situ hybridization (FISH): History, limitations and what to expect from micro-scale FISH? *Micro and Nano Engineering*, 1, 15–24.
- IUCN. (2022). *The IUCN Red List of Threatened Species*.
- Iyama, T., & Wilson, D. M. (2013). DNA repair mechanisms in dividing and non-dividing cells. *DNA Repair*, 12(8), 620–636.
- Jeong, E., Martina, J. A., Contreras, P. S., Lee, J., & Puertollano, R. (2022). The FACT complex facilitates expression of lysosomal and antioxidant genes through binding to TFEB and TFE3. *Autophagy*, 18(10), 2333–2349.
- Jia, L., Ge, X., Du, C., Chen, L., Zhou, Y., Xiong, W., Xiang, J., Li, G., Xiao, G., Fang, L., & Li, Z. (2021). EEF1A2 interacts with HSP90AB1 to promote lung adenocarcinoma metastasis via enhancing TGF- β /SMAD signalling. *British Journal of Cancer*, 124(7), 1301–1311.
- Jones, B. R., Rajaraman, A., Tannier, E., & Chauve, C. (2012). ANGES: Reconstructing ANcestral GENomeS maps. *Bioinformatics*, 28(18), 2388–2390.
- Joron, M., Frezal, L., Jones, R. T., Chamberlain, N. L., Lee, S. F., Haag, C. R., Whibley, A., Becuwe, M., Baxter, S. W., Ferguson, L., Wilkinson, P. A., Salazar, C., Davidson, C., Clark, R., Quail, M. A., Beasley, H., Glithero, R., Lloyd, C., Sims, S., ffrench-Constant, R. H. (2011). Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*, 477(7363), 203–206.

- Kamber, R. A., Nishiga, Y., Morton, B., Banuelos, A. M., Barkal, A. A., Vences-Catalán, F., Gu, M., Fernandez, D., Seoane, J. A., Yao, D., Liu, K., Lin, S., Spees, K., Curtis, C., Jerby-Arnon, L., Weissman, I. L., Sage, J., & Bassik, M. C. (2021). Inter-cellular CRISPR screens reveal regulators of cancer cell phagocytosis. *Nature*, 597(7877), 549–554.
- Kapusta, A., Suh, A., & Feschotte, C. (2017). Dynamics of genome size evolution in birds and mammals. *Proceedings of the National Academy of Sciences*, 114(8).
- Kasai, F., O'Brien, P. C. M., Pereira, J. C., & Ferguson-Smith, M. A. (2018). Marsupial chromosome DNA content and genome size assessed from flow karyotypes: Invariable low autosomal GC content. *Royal Society Open Science*, 5(8), 171539.
- Kayigwe, A. N., M. Darby, J., Lyons, A. B., L. Patchett, A., Lisowski, L., Liu, G.-S., & S. Flies, A. (2022). A human adenovirus encoding IFN- γ can transduce Tasmanian devil facial tumour cells and upregulate MHC-I. *Journal of General Virology*, 103(11).
- Keilwagen, J., Hartung, F., & Grau, J. (2019). GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data. In M. Kollmar (Ed.), *Gene Prediction* (Vol. 1962, pp. 161–177). Springer New York.
- Keilwagen, J., Wenk, M., Erickson, J. L., Schattat, M. H., Grau, J., & Hartung, F. (2016). Using intron position conservation for homology-based gene prediction. *Nucleic Acids Research*, 44(9), e89–e89.
- Kempfer, R., & Pombo, A. (2020). Methods for mapping 3D chromosome architecture. *Nature Reviews Genetics*, 21(4), 207–226.
- Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S., & Karolchik, D. (2010). BigWig and BigBed: Enabling browsing of large distributed datasets. *Bioinformatics*, 26(17), 2204–2207.
- Khan, A. I., Dinh, D. M., Schneider, D., Lenski, R. E., & Cooper, T. F. (2011). Negative epistasis between beneficial mutations in an evolving bacterial population. *Science*, 332(6034), 1193–1196.
- Khanna, R., Krishnamoorthy, V., & Parnaik, V. K. (2018). E3 ubiquitin ligase RNF 123 targets lamin B1 and lamin-binding proteins. *The FEBS Journal*, 285(12), 2243–2262.
- Kim, J., Farré, M., Auvil, L., Capitanu, B., Larkin, D. M., Ma, J., & Lewin, H. A. (2017). Reconstruction and evolutionary history of eutherian chromosomes. *Proceedings of the National Academy of Sciences*, 114(27).
- Kim, J., Larkin, D. M., Cai, Q., Asan, Zhang, Y., Ge, R.-L., Auvil, L., Capitanu, B., Zhang, G., Lewin, H. A., & Ma, J. (2013). Reference-assisted chromosome assembly. *Proceedings of the National Academy of Sciences*, 110(5), 1785–1790.
- Kirkpatrick, M. (2010). How and Why Chromosome Inversions Evolve. *PLOS Biology*, 8(9), e1000501.
- Kolmogorov, M., Armstrong, J., Raney, B. J., Streeter, I., Dunn, M., Yang, F., Odom, D., Flicek, P., Keane, T. M., Thybert, D., Paten, B., & Pham, S. (2018). Chromosome assembly of large and complex genomes using multiple references. *Genome Research*, 28(11), 1720–1732.

- Kowalczyk, A., Meyer, W. K., Partha, R., Mao, W., Clark, N. L., & Chikina, M. (2019). RERconverge: An R package for associating evolutionary rates with convergent traits. *Bioinformatics*, 35(22), 4815–4817.
- Krefting, J., Andrade-Navarro, M. A., & Ibn-Salem, J. (2018). Evolutionary stability of topologically associating domains is associated with conserved gene regulation. *BMC Biology*, 16(1), 87.
- Kuhn, R. M., Haussler, D., & Kent, W. J. (2013). The UCSC genome browser and associated tools. *Briefings in Bioinformatics*, 14(2), 144–161.
- Kulemzina, A. I., Perelman, P. L., Grafodatskaya, D. A., Nguyen, T. T., Thompson, M., Roelke-Parker, M. E., & Graphodatsky, A. S. (2014). Comparative chromosome painting of pronghorn (*Antilocapra americana*) and saola (*Pseudoryx nghetinhensis*) karyotypes with human and dromedary camel probes. *BMC Genetics*, 15(1), 68.
- Kulemzina, A. I., Trifonov, V. A., Perelman, P. L., Rubtsova, N. V., Volobuev, V., Ferguson-Smith, M. A., Stanyon, R., Yang, F., & Graphodatsky, A. S. (2009). Cross-species chromosome painting in Cetartiodactyla: Reconstructing the karyotype evolution in key phylogenetic lineages. *Chromosome Research*, 17(3), 419–436.
- Kulemzina, A. I., Yang, F., Trifonov, V. A., Ryder, O. A., Ferguson-Smith, M. A., & Graphodatsky, A. S. (2011). Chromosome painting in Tragulidae facilitates the reconstruction of Ruminantia ancestral karyotype. *Chromosome Research*, 19(4), 531–539.
- Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Molecular Biology and Evolution*, 34(7), 1812–1819.
- Kurakake, M., & Amai, Y. (2022). Characterization of a β -N-acetylhexosaminidase with transglycosylation activity from *Metarhizium* sp. A34. *Journal of Food Science*, 87(4), 1466–1474.
- Lamb, N. A., Bard, J. E., Loll-Krippelber, R., Brown, G. W., & Surtees, J. A. (2022). Complex mutation profiles in mismatch repair and ribonucleotide reductase mutants reveal novel repair substrate specificity of MutS homolog (MSH) complexes. *Genetics*, 221(4).
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczy, J., LeVine, R., McEwan, P., Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 860–921.
- Larkin, D. M., Pape, G., Donthu, R., Auvil, L., Welge, M., & Lewin, H. A. (2009). Breakpoint regions and homologous syntenic blocks in chromosomes have different evolutionary histories. *Genome Research*, 19(5), 770–777.
- Lee, J., Han, K., Meyer, T. J., Kim, H.-S., & Batzer, M. A. (2008). Chromosomal inversions between human and chimpanzee lineages caused by retrotransposons. *PLOS ONE*, 3(12), e4047.
- Lee, K., Tosti, E., & Edelmann, W. (2016). Mouse models of DNA mismatch repair in cancer research. *DNA Repair*, 38, 140–146.

- Lee, N., Kim, D.-K., Han, S. H., Ryu, H. G., Park, S. J., Kim, K.-T., & Choi, K. Y. (2017). Comparative Interactomes of VRK1 and VRK3 with Their Distinct Roles in the Cell Cycle of Liver Cancer. *Molecules and Cells*, 40(9), 621–631.
- Lemaitre, C., Zaghloul, L., Sagot, M.-F., Gautier, C., Arneodo, A., Tannier, E., & Audit, B. (2009). Analysis of fine-scale mammalian evolutionary breakpoints provides new insight into their relation to genome organisation. *BMC Genomics*, 10(1), 335.
- Lewin, H. A., Graves, J. A. M., Ryder, O. A., Graphodatsky, A. S., & O'Brien, S. J. (2019). Precision nomenclature for the new genomics. *GigaScience*, 8(8), giz086.
- Li, X., Zhu, C., Lin, Z., Wu, Y., Zhang, D., Bai, G., Song, W., Ma, J., Muehlbauer, G. J., Scanlon, M. J., Zhang, M., & Yu, J. (2011). Chromosome size in diploid eukaryotic species centers on the average length with a conserved boundary. *Molecular Biology and Evolution*, 28(6), 1901–1911.
- Li, Y., Liu, Z., Shi, P., & Zhang, J. (2010). The hearing gene Prestin unites echolocating bats and whales. *Current Biology*, 20(2), R55–R56.
- Liu, C., Xu, X., Han, L., Wan, X., Zheng, L., Li, C., Liao, Z., Xiao, J., Zhong, R., Zheng, X., Wang, Q., Li, Z., Chen, H., Wei, B., & Wang, H. (2020). LRCH1 deficiency enhances LAT signalosome formation and CD8⁺ T cell responses against tumors and pathogens. *Proceedings of the National Academy of Sciences*, 117(32), 19388–19398.
- Liu, G. E., Ventura, M., Cellamare, A., Chen, L., Cheng, Z., Zhu, B., Li, C., Song, J., & Eichler, E. E. (2009). Analysis of recent segmental duplications in the bovine genome. *BMC Genomics*, 10(1), 571.
- Liu, P., Zhou, T.-F., Qiu, B.-A., Yang, Y.-X., Zhu, Y.-J., An, Y., Zhao, W.-C., Wu, Y.-T., Ma, P.-F., Li, J.-B., & Xia, N.-X. (2018). Methylation-Mediated Silencing of GATA5 Gene Suppresses Cholangiocarcinoma Cell Proliferation and Metastasis. *Translational Oncology*, 11(3), 585–592.
- Liu, Z., Wang, T., Pryce, J. E., MacLeod, I. M., Hayes, B. J., Chamberlain, A. J., Jagt, C. V., Reich, C. M., Mason, B. A., Rochfort, S., & Cocks, B. G. (2019). Fine-mapping sequence mutations with a major effect on oligosaccharide content in bovine milk. *Scientific Reports*, 9(1), 2137.
- Lodish, H. (Ed.). (2008). *Molecular cell biology* (6. ed). Freeman.
- Longo, M. S., Carone, D. M., NISC Comparative Sequencing Program, Green, E. D., O'Neill, M. J., & O'Neill, R. J. (2009). Distinct retroelement classes define evolutionary breakpoints demarcating sites of evolutionary novelty. *BMC Genomics*, 10(1), 334.
- Losos, J. B. (2011). Convergence, adaptation, and constraint. *Evolution*, 65(7), 1827–1840.
- Lu, Z., Fei, L., & Hou, G. (2022). A pan-cancer analysis of the oncogenic role of ERCC6L. *BMC Cancer*, 22(1), 1347.

- Lu, Z., Yue, Y., Yuan, C., Liu, J., Chen, Z., Niu, C., Sun, X., Zhu, S., Zhao, H., Guo, T., & Yang, B. (2020). Genome-wide association study of body weight traits in chinese fine-wool sheep. *Animals*, 10(1), 170.
- Lupski, J. R. (1998). Genomic disorders: Structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends in Genetics*, 14(10), 417–422.
- Ma, J., Zhang, L., Suh, B. B., Raney, B. J., Burhans, R. C., Kent, W. J., Blanchette, M., Haussler, D., & Miller, W. (2006). Reconstructing contiguous regions of an ancestral genome. *Genome Research*, 16(12), 1557–1565.
- Malishkevich, A., Leyk, J., Goldbaum, O., Richter-Landsberg, C., & Gozes, I. (2015). ADNP/ADNP2 expression in oligodendrocytes: Implication for myelin-related neurodevelopment. *Journal of Molecular Neuroscience*, 57(2), 304–313.
- Mammal Diversity Database. (2022). *Mammal Diversity Database (1.9)*.
- Marques-Bonet, T., Sánchez-Ruiz, J., Armengol, L., Khaja, R., Bertranpetit, J., Lopez-Bigas, N., Rocchi, M., Gazave, E., & Navarro, A. (2007). On the association between chromosomal rearrangements and genic evolution in humans and chimpanzees. *Genome Biology*, 8(10), R230.
- Marsh, H. (2014). Family Dugongidae (Dugong). In *Handbook of the Mammals of the World* (Barcelona: Lynx Edicon, Vols 4-Sea Mammals, pp. 564–573).
- Meagher, M., & Lightowlers, R. N. (2014). The role of TDP1 and APTX in mitochondrial DNA repair. *Biochimie*, 100, 121–124.
- Meng, C., Xia, S., He, Y., Tang, X., Zhang, G., & Zhou, T. (2020). Discovery of prognostic signature genes for overall survival prediction in gastric cancer. *Computational and Mathematical Methods in Medicine*, 2020, 1–9.
- Meredith, R. W., Mendoza, M. A., Roberts, K. K., Westerman, M., & Springer, M. S. (2010). A Phylogeny and Timescale for the Evolution of Pseudocheiridae (Marsupialia: Diprotodontia) in Australia and New Guinea. *Journal of Mammalian Evolution*, 17(2), 75–99.
- Mitchell, K. J., Pratt, R. C., Watson, L. N., Gibb, G. C., Llamas, B., Kasper, M., Edson, J., Hopwood, B., Male, D., Armstrong, K. N., Meyer, M., Hofreiter, M., Austin, J., Donnellan, S. C., Lee, M. S. Y., Phillips, M. J., & Cooper, A. (2014). Molecular phylogeny, biogeography, and habitat preference evolution of marsupials. *Molecular Biology and Evolution*, 31(9), 2322–2330.
- Mok, E. H. K., Leung, C. O. N., & Lee, T. K. W. (2020). MAP9/ERCC3 signaling cascade: A new insight on understanding the chromosomal instability in hepatocellular carcinoma. *EBioMedicine*, 54, 102709.
- Montazerinezhad, S., Emamjomeh, A., & Hajieghrari, B. (2020). Chromosomal abnormality, laboratory techniques, tools and databases in molecular Cytogenetics. *Molecular Biology Reports*, 47(11), 9055–9073.

- Morgan, T. H., Sturtevant, A. H., Muller, H. J., & Bridges, C. B. (1915). *The mechanism of Mendelian heredity*. Henry Holt and Company.
- Mu, T., Hu, H., Ma, Y., Wen, H., Yang, C., Feng, X., Wen, W., Zhang, J., & Gu, Y. (2022). Identifying key genes in milk fat metabolism by weighted gene co-expression network analysis. *Scientific Reports*, 12(1), 6836.
- Muffato, M. (2010). *Reconstruction de génomes ancestraux chez les vertébrés*. Université d'Evry Val d'Essonne.
- Muffato, M., Louis, A., Nguyen, N. T. T., Lucas, J., Berthelot, C., & Crollius, H. R. (2022). Reconstruction of hundreds of reference ancestral genomes across the eukaryotic kingdom [Preprint].
- Muller, H., Gil, J., & Drinnenberg, I. A. (2019). The impact of centromeres on spatial genome architecture. *Trends in Genetics*, 35(8), 565–578.
- Murphy, W. J., Foley, N. M., Bredemeyer, K. R., Gatesy, J., & Springer, M. S. (2021). Phylogenomics and the genetic architecture of the placental mammal radiation. *Annual Review of Animal Biosciences*, 9(1), 29–53.
- Murphy, W. J., Larkin, D. M., der Wind, A. E., Bourque, G., Tesler, G., Auvil, L., Beever, J. E., Chowdhary, B. P., Galibert, F., Gatzke, L., Hitte, C., Meyers, S. N., Milan, D., Ostrander, E. A., Pape, G., Parker, H. G., Raudsepp, T., Rogatcheva, M. B., Schook, L. B., Lewin, H. A. (2005). Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science*, 309(5734), 613–617.
- Murray, I. A., Patterson, A. D., & Perdew, G. H. (2015). AH Receptor ligands in cancer: friend and foe. *Nat Rev Cancer*, 14(12), 801–814.
- Nadeau, J. H., & Taylor, B. A. (1984). Lengths of chromosomal segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Sciences*, 81(3), 814–818.
- Navarro, A., & Barton, N. H. (2003). Chromosomal speciation and molecular divergence—accelerated evolution in rearranged chromosomes. *Science*, 300(5617), 321–324.
- Nilsson, E. M., Fainberg, H. P., Choong, S. S., Giles, T. C., Sells, J., May, S., Stansfield, F. J., Allen, W. R., Emes, R. D., Mostyn, A., Mongan, N. P., & Yon, L. (2014). Molecular characterization of adipose tissue in the African elephant (*Loxodonta africana*). *PLOS ONE*, 9(3), e91717.
- Nishihara, H., Satta, Y., Nikaido, M., Thewissen, J. G. M., Stanhope, M. J., & Okada, N. (2005). A retroposon analysis of afrotherian phylogeny. *Molecular Biology and Evolution*, 22(9), 1823–1833.
- Novacek, J. (1992). Mammalian phylogeny: Shaking the tree. *Nature*, 356, 121–125.
- O'Brien, S. J., Menninger, J. C., & Nash, W. G. (2005). *Atlas of Mammalian Chromosomes*. John Wiley & sons.
- O'Leary, M. A., Bloch, J. I., Flynn, J. J., Gaudin, T. J., Giallombardo, A., Giannini, N. P., Goldberg, S. L., Kraatz, B. P., Luo, Z.-X., Meng, J., Ni, X., Novacek, M. J., Perini, F. A., Randall, Z. S.,

- Rougier, G. W., Sargis, E. J., Silcox, M. T., Simmons, N. B., Spaulding, M., Cirranello, A. L. (2013). The placental mammal ancestor and the post-k-pg radiation of placentals. *Science*, 339(6120), 662–667.
- O’Neill, R. J. W., Eldridge, M. D. B., Toder, R., Ferguson-Smith, M. A., O’Brien, P. C., & Graves, J. A. M. (1999). *Chromosome evolution in kangaroos (Marsupialia: Macropodidae): Cross species chromosome painting between the tammar wallaby and rock wallaby spp. With the 2n = 22 ancestral macropodid karyotype*. 42, 6.
- O’Shea, T. (2014). Family Trichechidae (Manatees). In *Handbook of the Mammals of the World* (Barcelona: Lynx Edicions, Vols 4-Sea Mammals, pp. 548–562).
- Palma, S., Raffa, C. I., Garcia-Fabiani, M. B., Ferretti, V. A., Zwenger, A., Perez Verdera, P. V., Llontop, A., Rojas Bilbao, E., Cuartero, V., Abba, M. C., & Lacunza, E. (2020). RHBDD2 overexpression promotes a chemoresistant and invasive phenotype to rectal cancer tumors via modulating UPR and focal adhesion genes. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1866(8), 165810.
- Pardini, A. T., O’Brien, P. C. M., Fu, B., Bonde, R. K., Elder, F. F. B., Ferguson-Smith, M. A., Yang, F., & Robinson, T. J. (2007). Chromosome painting among Proboscidea, Hyracoidea and Sirenia: Support for Paenungulata (Afrotheria, Mammalia) but not Tethytheria. *Proceedings of the Royal Society B: Biological Sciences*, 274(1615), 1333–1340.
- Parker, J., Tsagkogeorga, G., Cotton, J. A., Liu, Y., Provero, P., Stupka, E., & Rossiter, S. J. (2013). Genome-wide signatures of convergent evolution in echolocating mammals. *Nature*, 502(7470), 228–231.
- Partha, R., Chauhan, B. K., Ferreira, Z., Robinson, J. D., Lathrop, K., Nischal, K. K., Chikina, M., & Clark, N. L. (2017). Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling. *ELife*, 6, e25884.
- Pauling, L., & Zuckerkandl, E. (1963). Chemical paleogenetics: Molecular “restoration studies” of extinct forms of life. *Acta Chem Scand*, 17, S9-16.
- Peng, Q., Pevzner, P. A., & Tesler, G. (2006). The Fragile Breakage versus Random Breakage Models of Chromosome Evolution. *PLOS Computational Biology*, 2(2), e14.
- Pertea, G., & Pertea, M. (2020). GFF Utilities: GffRead and GffCompare. *F1000Research*, 9, 304.
- Pevzner, P., & Tesler, G. (2003). Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proceedings of the National Academy of Sciences*, 100(13), 7672–7677.
- Pierce, B. A. (2020). *Genetics: A conceptual approach* (Seventh edition). Macmillan International, Higher Education.
- Portin, P., & Wilkins, A. (2017). The evolving definition of the term “gene”. *Genetics*, 205(4), 1353–1364.

- Poulsen, N. A., Robinson, R. C., Barile, D., Larsen, L. B., & Buitenhuis, B. (2019). A genome-wide association study reveals specific transferases as candidate loci for bovine milk oligosaccharides synthesis. *BMC Genomics*, 20(1), 404.
- Puig, M., Casillas, S., Villatoro, S., & Cáceres, M. (2015a). Human inversions and their functional consequences. *Briefings in Functional Genomics*, 14(5), 369–379.
- Purvis, A., Fritz, S. A., Rodríguez, J., Harvey, P. H., & Grenyer, R. (2011). The shape of mammalian phylogeny: Patterns, processes and scales. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1577), 2462–2477.
- Qi, D., Wang, J., Zhao, Y., Yang, Y., Wang, Y., Wang, L., Wang, Z., Xu, X., & Hu, Z. (2022). JMJD1C-regulated lipid synthesis contributes to the maintenance of MLL-rearranged acute myeloid leukemia. *Leukemia & Lymphoma*, 63(9).
- Rambaut, A. (2009). *FigTree, a graphical viewer of phylogenetic trees*. University of Edinburgh.
- Redi, C. A., & Capanna, E. (2012). Genome size evolution: sizing mammalian genomes. *Cytogenetic and Genome Research*, 137(2–4), 97–112.
- Rens, W., O'Brien, P. C. M., Yang, F., Solanky, N., Perelman, P., Graphodatsky, A. S., Ferguson, M. W. J., Svartman, M., Leo, A. A. D., Graves, J. A. M., & Ferguson-Smith, M. A. (2001). Karyotype relationships between distantly related marsupials from South America and Australia. *Chromosome Research*, 9, 301–308.
- Rey, C., Lanore, V., Veber, P., Guéguen, L., Lartillot, N., Sémon, M., & Boussau, B. (2019). Detecting adaptive convergent amino acid evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1777), 20180234.
- Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Functamman, A., Kim, J., Lee, C., Ko, B. J., Chaisson, M., Gedman, G. L., Cantin, L. J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, 592(7856), 737–746.
- Rice, E. S., & Green, R. E. (2019). New approaches for genome assembly and scaffolding. *Annual Review of Animal Biosciences*, 7(1), 17–40.
- Richard, G.-F., Kerrest, A., & Dujon, B. (2008). Comparative genomics and molecular dynamics of dna repeats in eukaryotes. *Microbiology and Molecular Biology Reviews*, 72(4), 686–727.
- Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, 29(1), 24–26.
- Robinson, T. J., Fu, B., Ferguson-Smith, M. A., & Yang, F. (2004). Cross-species chromosome painting in the golden mole and elephant-shrew: Support for the mammalian clades Afrotheria and Afroinsectiphillia but not Afroinsectivora. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 271(1547), 1477–1484.
- Robinson, T., & Seiffert, E. (2004). Afrotherian origins and interrelationships: New views and future prospects. *Curr. Top. Dev. Biol.*, 63, 37–60.

- Roca, A. L., & O'Brien, S. J. (2005). Genomic inferences from Afrotheria and the evolution of elephants. *Current Opinion in Genetics & Development*, 15(6), 652–659.
- Rowley, M. J., & Corces, V. G. (2018). Organizational principles of 3D genome architecture. *Nature Reviews Genetics*, 19(12), 789–800.
- Ruiz-Herrera, A., Castresana, J., & Robinson, T. J. (2006). Is mammalian chromosomal evolution driven by regions of genome fragility? *Genome Biology*, 7(12), R115.
- Ruiz-Herrera, A., Farré, M., Ponsà, M., & Robinson, T. J. (2010). Selection against Robertsonian fusions involving housekeeping genes in the house mouse: Integrating data from gene expression arrays and chromosome evolution. *Chromosome Research*, 18(7), 801–808.
- Ruiz-Herrera, A., Farré, M., & Robinson, T. J. (2012). Molecular cytogenetic and genomic insights into chromosomal evolution. *Heredity*, 108(1), 28–36.
- Ruiz-Herrera, A., Ponsa, M., & Garc, F. (2002). Fragile sites in human and *Macaca fascicularis* chromosomes are breakpoints in chromosome evolution. *Chromosome Research*, 10, 33–44.
- Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., Sherry, S. T. (2022). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 50(D1), D20–D26.
- Scherthan, H., Cremer, T., Arnason, U., Weier, H.-U., Lima-de-Faria, A., & Fröncke, L. (1994). Comparative chromosome painting discloses homologous segments in distantly related mammals. *Nature Genetics*, 6(4), 342–347.
- Schibler, L., Roig, A., Mahe, M.-F., Laurent, P., Hayes, H., Rodolphe, F., & Cribiu, E. P. (2006). High-resolution comparative mapping among man, cattle and mouse suggests a role for repeat sequences in mammalian genome evolution. *BMC Genomics*, 7(1), 194.
- Schubert, I., & Lysak, M. A. (2011). Interpretation of karyotype evolution should consider chromosome structural constraints. *Trends in Genetics*, 27(6), 207–216.
- Schubert, I., & Oud, J. L. (1997). There is an upper limit of chromosome size for normal development of an organism. *Cell*, 88(4), 515–520.
- Seiffert, E. R. (2007). A new estimate of afrotherian phylogeny based on simultaneous analysis of genomic, morphological, and fossil evidence. *BMC Evolutionary Biology*, 7(1), 224.
- Selmansberger, M., Michna, A., Braselmann, H., Höfig, I., Schorpp, K., Weber, P., Anastasov, N., Zitzelsberger, H., Hess, J., & Unger, K. (2020). Transcriptome network of the papillary thyroid carcinoma radiation marker CLIP2. *Radiation Oncology*, 15(1), 182.
- Sharman, G. (1961). The mitotic chromosomes of marsupials and their bearing on taxonomy and phylogeny. *Aust.J.Zool*, 9(1), 38–60.
- Sharp, A. J., Hansen, S., Selzer, R. R., Cheng, Z., Regan, R., Hurst, J. A., Stewart, H., Price, S. M., Blair, E., Hennekam, R. C., Fitzpatrick, C. A., Segraves, R., Richmond, T. A., Guiver, C., Albertson, D. G., Pinkel, D., Eis, P. S., Schwartz, S., Knight, S. J. L., & Eichler, E. E. (2006).

- Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nature Genetics*, 38(9), 1038–1042.
- She, X., Cheng, Z., Zöllner, S., Church, D. M., & Eichler, E. E. (2008). Mouse segmental duplication and copy number variation. *Nature Genetics*, 40(7), 909–914. <https://doi.org/10.1038/ng.172>
- Shoshani, J., & Tassy, P. (1996). The Proboscidea: Evolution and Palaeoecology of Elephants and their Relatives. *Oxford University Press*.
- Siva, N. (2008). 1000 Genomes project. *Nature Biotechnology*, 26(3), 256–256.
- Slate, J., Van Stijn, T. C., Anderson, R. M., McEwan, K. M., Maqbool, N. J., Mathias, H. C., Bixley, M. J., Stevens, D. R., Molenaar, A. J., Beever, J. E., Galloway, S. M., & Tate, M. L. (2002). A deer (subfamily Cervinae) genetic linkage map and the evolution of ruminant genomes. *Genetics*, 160(4), 1587–1597.
- Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., Arnaiz, O., Awedh, M. H., Baldock, R., Barbiera, G., Bardou, P., Beck, T., Blake, A., Bonierbale, M., Brookes, A. J., Bucci, G., Buetti, I., Burge, S., Cabau, C., Kasprzyk, A. (2015). The BioMart community portal: An innovative alternative to large, centralized data repositories. *Nucleic Acids Research*, 43(W1), W589–W598.
- Solvie, D., Baluapuri, A., Uhl, L., Fleischhauer, D., Endres, T., Papadopoulos, D., Aziba, A., Gaballa, A., Mikicic, I., Isaakova, E., Giansanti, C., Jansen, J., Jungblut, M., Klein, T., Schüle-Völke, C., Maric, H., Dose, S., Sauer, M., Beli, P., Eilers, M. (2022). MYC multimers shield stalled replication forks from RNA polymerase. *Nature*, 612(7938), 148–155.
- Soman, A., & Nair, S. A. (2022). Unfolding the cascade of SERPINA3: Inflammation to cancer. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1877(5), 188760.
- Song, S., Liu, L., Edwards, S. V., & Wu, S. (2012). Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proceedings of the National Academy of Sciences*, 109(37), 14942–14947.
- Speicher, M. R., & Carter, N. P. (2005). The new cytogenetics: Blurring the boundaries with molecular biology. *Nature Reviews Genetics*, 6(10), 782–792.
- Springer, M. S. (2022). Afrotheria. *Current Biology*, 32(5), R205–R210.
- Stadler, T. (2011). Mammalian phylogeny reveals recent diversification rate shifts. *Proceedings of the National Academy of Sciences*, 108(15), 6187–6192.
- Stanhope, M. J., Waddell, V. G., Madsen, O., de Jong, W., Hedges, S. B., Cleven, G. C., Kao, D., & Springer, M. S. (1998). Molecular evidence for multiple origins of Insectivora and for a new order of endemic African insectivore mammals. *Proceedings of the National Academy of Sciences*, 95(17), 9967–9972.
- Stanyon, R., Consigliere, S., Muller, S., Morescalchi, A., Neusser, M., & Wienberg, J. (2000). Fluorescence in situ hybridization (FISH) maps chromosomal homologies between the dusky titi and squirrel monkey. *American Journal of Primatology*, 50(2), 95–107.

- Storz, J. F. (2016). Causes of molecular convergence and parallelism in protein evolution. *Nature Reviews Genetics*, 17(4), 239–250.
- Sukumar, R. (2004). The living elephants: Evolutionary ecology, behavior, and conservation. *Journal of Mammalogy*, 85(3), 581–582.
- Svartman, M., Stone, G., & Stanyon, R. (2006). The ancestral eutherian karyotype is present in Xenarthra. *PLOS Genetics*, 2(7), e109.
- Tempel, S. (2012). Using and Understanding RepeatMasker. In Y. Bigot (Ed.), *Mobile Genetic Elements* (Vol. 859, pp. 29–51). Humana Press.
- Tesler, G. (2002). GRIMM: Genome rearrangements web server. *Bioinformatics*, 18(3), 492–493.
- The Darwin Tree of Life Project Consortium (2021). Sequence locally, think globally: The Darwin Tree of Life Project. *PNAS Special feature: Perspective*, 119 (4):e2115642118.
- The Gene Ontology Consortium, Carbon, S., Douglass, E., Good, B. M., Unni, D. R., Harris, N. L., Mungall, C. J., Basu, S., Chisholm, R. L., Dodson, R. J., Hartline, E., Fey, P., Thomas, P. D., Albou, L.-P., Ebert, D., Kesling, M. J., Mi, H., Muruganujan, A., Huang, X., Elser, J. (2021). The Gene Ontology resource: Enriching a GOLD mine. *Nucleic Acids Research*, 49(D1), D325–D334.
- Thorvaldsdottir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2), 178–192.
- Toder, R., O'Neill, R. J. W., & Graves, J. A. M. (1998). Chromosome Painting in Marsupials. *ILAR Journal*, 39(2–3), 92–95.
- Toder, R., O'Neill, R. J. W., Wienberg, J., O'Brien, P. C. M., Voullaire, L., & Marshall-Graves, J. A. (1997). Comparative chromosome painting between two marsupials: Origins of an XX/XY1Y2 sex chromosome system. *Mammalian Genome*, 8(6), 418–422.
- Toral, P. G., Monahan, F. J., Hervás, G., Frutos, P., & Moloney, A. P. (2018). Review: Modulating ruminal lipid metabolism to improve the fatty acid composition of meat and milk. Challenges and opportunities. *Animal*, 12, s272–s281.
- Tyndale-Biscoe, C. (2005). *Life of Marsupials*. CSIRO Publishing.
- Valiente, M., Obenaus, A. C., Jin, X., Chen, Q., Zhang, X. H.-F., Lee, D. J., Chaff, J. E., Kris, M. G., Huse, J. T., Brogi, E., & Massagué, J. (2014). Serpins promote cancer cell survival and vascular co-option in brain metastasis. *Cell*, 156(5), 1002–1016.
- Veyrunes, F., Dobigny, G., Yang, F., O'Brien, P. C. M., Catalan, J., Robinson, T. J., & Britton-Davidian, J. (2006). Phylogenomics of the genus *Mus* (Rodentia; Muridae): Extensive genome repatterning is not restricted to the house mouse. *Proceedings of the Royal Society B: Biological Sciences*, 273(1604), 2925–2934.
- Wang, B., Chen, L., & Wang, W. (2019). Genomic insights into ruminant evolution: From past to future prospects. *Zool. Res.*, 40(6), 476–487.

- Wang, L., Zhai, R., Shen, H., Song, G., Wan, F., & Li, Q. (2021). VRK1 promotes proliferation, migration, and invasion of gastric carcinoma cells by activating β -catenin. *Neoplasma*, 68(05), 1005–1014.
- Wang, Q., Liu, Z., Zhai, G., Yu, X., Ke, S., Shao, H., & Guo, J. (2022). Overexpression of GATA5 Inhibits Prostate Cancer Progression by Regulating PLAGL2 via the FAK/PI3K/AKT Pathway. *Cancers*, 14(9), 2074.
- Wang, S., Huang, J., Li, C., Zhao, L., Wong, C. C., Zhai, J., Zhou, Y., Deng, W., Zeng, Y., Gao, S., Zhang, Y., Wang, G., Guan, X. Y., Wei, H., Wong, S. H., He, H. H., Shay, J. W., & Yu, J. (2020). MAP9 loss triggers chromosomal instability, initiates colorectal tumorigenesis, and is associated with poor survival of patients with colorectal cancer. *Clinical Cancer Research*, 26(3), 746–757.
- Wang, X., Cai, Q., Ping, J., Diaz-Zabala, H., Xia, Y., & Guo, X. (2022). The putative oncogenic role of *WDT1* in colorectal cancer. *Carcinogenesis*, 43(6), 594–600.
- Wang, Y., Zhang, C., Wang, N., Li, Z., Heller, R., Liu, R., Zhao, Y., Han, J., Pan, X., Zheng, Z., Dai, X., Chen, C., Dou, M., Peng, S., Chen, X., Liu, J., Li, M., Wang, K., Liu, C., Qiu, Q. (2019). Genetic basis of ruminant headgear and rapid antler regeneration. *Science*, 364(6446), eaav6335.
- Wang, Z., Snyder, M., Kenison, J. E., Yang, K., Lara, B., Lydell, E., Bennani, K., Novikov, O., Federico, A., Monti, S., & Sherr, D. H. (2020). How the AHR became important in cancer: The role of chronically active AHR in cancer aggression. *International Journal of Molecular Sciences*, 22(1), 387.
- Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C. T., Maitland, A., Mostafavi, S., Montojo, J., Shao, Q., Wright, G., Bader, G. D., & Morris, Q. (2010). The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*, 38(suppl_2), W214–W220.
- Watanabe, S., Watanabe, K., Akimov, V., Bartkova, J., Blagoev, B., Lukas, J., & Bartek, J. (2013). JMJD1C demethylates MDC1 to regulate the RNF8 and BRCA1-mediated chromatin response to DNA breaks. *Nature Structural & Molecular Biology*, 20(12), 1425–1433.
- Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., & Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, 8(12), 973–982.
- Wilson, G. M., Flibotte, S., Missirlis, P. I., Marra, M. A., Jones, S., Thornton, K., Clark, A. G., & Holt, R. A. (2006). Identification by full-coverage array CGH of human DNA copy number increases relative to chimpanzee and gorilla. *Genome Research*, 16(2), 173–181.
- Wolffe, A. (1998). *Chromatin: Structure and function*. Elsevier Ltd.

- Wolffe, A. P., & Matzke, M. A. (1999). Epigenetics: Regulation through repression. *Science*, 286(5439), 481–486.
- Wu, J., Li, T., Ji, H., Chen, Z., & Zhai, B. (2022). VRK1 predicts poor prognosis and promotes bladder cancer growth and metastasis in vitro and in vivo. *Frontiers in Pharmacology*, 13, 874235.
- Wu, Z., Huang, X., Cai, M., Huang, P., & Guan, Z. (2022). Novel necroptosis-related gene signature for predicting the prognosis of pancreatic adenocarcinoma. *Aging*, 14(2), 869–891.
- Yamamoto, K., Venida, A., Yano, J., Biancur, D. E., Kakiuchi, M., Gupta, S., Sohn, A. S. W., Mukhopadhyay, S., Lin, E. Y., Parker, S. J., Banh, R. S., Paulo, J. A., Wen, K. W., Debnath, J., Kim, G. E., Mancias, J. D., Fearon, D. T., Perera, R. M., & Kimmelman, A. C. (2020). Autophagy promotes immune evasion of pancreatic cancer by degrading MHC-I. *Nature*, 581(7806), 100–105.
- Yang, F., Alkalaeva, E. Z., Perelman, P. L., Pardini, A. T., Harrison, W. R., O'Brien, P. C. M., Fu, B., Graphodatsky, A. S., Ferguson-Smith, M. A., & Robinson, T. J. (2003). Reciprocal chromosome painting among human, aardvark, and elephant (superorder Afrotheria) reveals the likely eutherian ancestral karyotype. *Proceedings of the National Academy of Sciences*, 100(3), 1062–1066.
- Yang, F., Müller, S., Just, R., Ferguson-Smith, M. A., & Wienberg, J. (1997). Comparative Chromosome Painting in Mammals: Human and the Indian Muntjac (*Muntiacus muntjak vaginalis*). *Genomics*, 39(3), 396–401.
- Yang, Y., Rader, E., Peters-Carr, M., Bent, R. C., Smilowitz, J. T., Guillemin, K., & Rader, B. (2019). Ontogeny of alkaline phosphatase activity in infant intestines and breast milk. *BMC Pediatrics*, 19(1), 2.
- Yang, Y., Zhang, X., Zhang, X., Wang, Y., Wang, X., Hu, L., Zhao, Y., Wang, H., Wang, Z., Wang, H., Wang, L., Dirks, W. G., Drexler, H. G., Xu, X., & Hu, Z. (2021). Modulators of histone demethylase JMJD1C selectively target leukemic stem cells. *FEBS Open Bio*, 11(1), 265–277.
- Yang, Z., Jiang, X., Zhang, Z., Zhao, Z., Xing, W., Liu, Y., Jiang, X., & Zhao, H. (2021). HDAC3-dependent transcriptional repression of FOXA2 regulates FTO/m6A/MYC signaling to contribute to the development of gastric cancer. *Cancer Gene Therapy*, 28(1–2), 141–155.
- Yao, W., Yang, Y., Chen, X., Cui, X., Zhou, B., Zhao, B., Lin, Z., & Miao, J. (2022). Activation of Esterase D by FPD5 inhibits growth of A549 lung cancer cells via JAB1/p53 pathway. *Genes*, 13(5), 786.
- Yazer, M. H., & Palcic, M. M. (2005). The importance of disordered loops in ABO glycosyltransferases. *Transfusion Medicine Reviews*, 19(3), 210–216.
- Yazhini, A., Srinivasan, N., & Sandhya, S. (2021). Signatures of conserved and unique molecular features in Afrotheria. *Scientific Reports*, 11(1), 1011.
- Zachos, J., Pagani, M., Sloan, L., Thomas, E., & Billups, K. (2001). Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science*, 292(5517), 686–693.

- Zhang, Z., Zheng, F., Yu, Z., Hao, J., Chen, M., Yu, W., Guo, W., Chen, Y., Huang, W., Duan, Z., & Deng, W. (2017). XRCC5 cooperates with p300 to promote cyclooxygenase-2 expression and tumor growth in colon cancers. *PLOS ONE*, 12(10), e0186900.
- Zhu, X., Xiang, Z., Zou, L., Chen, X., Peng, X., & Xu, D. (2021). APMAP promotes epithelial-mesenchymal transition and metastasis of cervical cancer cells by activating the Wnt/ β -catenin pathway. *Journal of Cancer*, 12(20), 6265–6273.



9. Supplementary data

Table S1. Information of all ruminant species. Size (Gb) refers to the total genome size. A.L. refers to the assembly level, and the species can be assembled in chromosomes (Chr), c-scaffolds (C-sca), super-scaffolds (S-sca), or scaffolds (Sca). The 2n refers to the diploid number, and in some species can be variable depending on their sex (female/male). IUCN refers to their status in the IUCN red list, which can be extinct (EX), extinct in the wild (EXW), critically endangered (CE), endangered (EN), vulnerable (VU), near threatened (NT), or least concern (LC). NA refers to not available information.

Species	Latin Name	Family	2n	Size (Gb)	IUCN	A.L.
Pronghorn	<i>Antilocapra americana</i>	Antilocapridae	58	2.95	LC	Sca
Impala	<i>Aepyceros melampus</i>	Bovidae	58/60	2.63	LC	Sca
Hartebeest	<i>Alcelaphus buselaphus</i>	Bovidae	40	3.88	LC	Sca
Hirola	<i>Beatragus hunteri</i>	Bovidae	44	2.69	CE	Sca
Black wildebeest	<i>Connochaetes gnou</i>	Bovidae	58	NA	LC	NA
Brindled gnu/Blue wildebeest	<i>Connochaetes taurinus</i>	Bovidae	58	2.65	LC	S-sca
Topi	<i>Damaliscus lunatus</i>	Bovidae	36	3.12	LC	Sca
Bontebok	<i>Damaliscus pygargus</i>	Bovidae	38	NA	LC	NA
Springbok	<i>Antidorcas marsupialis</i>	Bovidae	56	3.02	LC	Sca
Blackbuck	<i>Antilope cervicapra</i>	Bovidae	30/33	NA	LC	NA
Beira	<i>Dorcatragus megalotis</i>	Bovidae	NA	NA	VU	NA
Red-fronted gazelle	<i>Eudorcas rufufrons</i>	Bovidae	58	NA	VU	NA
Thomson's gazelle	<i>Eudorcas thomsonii</i>	Bovidae	58	2.9	LC	Sca
Arabian gazelle	<i>Gazella arabica</i>	Bovidae	NA	NA	VU	NA
Chinkara	<i>Gazella bennettii</i>	Bovidae	49/52	NA	LC	NA
Queen of Sheba's gazelle	<i>Gazella bilkis</i>	Bovidae	NA	NA	EX	NA
Cuvier's gazelle	<i>Gazella cuvieri</i>	Bovidae	32/33	NA	VU	NA
Dorcas gazelle	<i>Gazella dorcas</i>	Bovidae	31	NA	VU	NA
Neumann's gazelle	<i>Gazella erlangeri</i>	Bovidae	34	NA	NA	NA
Mountain gazelle	<i>Gazella gazella</i>	Bovidae	34/35	NA	EN	NA
Rhim gazelle	<i>Gazella leptoceros</i>	Bovidae	33	NA	EN	NA
Arabian Sand gazelle	<i>Gazella marica</i>	Bovidae	32	NA	VU	NA
Saudi gazelle	<i>Gazella saudiya</i>	Bovidae	46/53	NA	EX	NA
Speke's gazelle	<i>Gazella spekei</i>	Bovidae	32/33	NA	EN	NA
Goitered gazelle	<i>Gazella subgutturosa</i>	Bovidae	32/31	NA	VU	NA
Gerenuk	<i>Litocranius walleri</i>	Bovidae	60	2.98	NT	S-sca
Guenther's dik-dik	<i>Madoqua guentheri</i>	Bovidae	48/50	NA	LC	NA
Kirk's dik-dik	<i>Madoqua kirkii</i>	Bovidae	46/48	2.65	LC	Sca
Salt's dik-dik	<i>Madoqua saltiana</i>	Bovidae	NA	NA	LC	NA
Dama gazelle	<i>Nanger dama</i>	Bovidae	38/40	NA	CE	NA
Grant's gazelle	<i>Nanger granti</i>	Bovidae	32/31	3.03	LC	Sca
Soemmerring's gazelle	<i>Nanger soemmerringii</i>	Bovidae	36	NA	VU	NA
Bate's dwarf antelope	<i>Neotragus batesi</i>	Bovidae	NA	NA	LC	NA

Suni	<i>Neotragus moschatus</i>	Bovidae	52/56	2.67	LC	Sca
Royal antelope	<i>Neotragus pygmaeus</i>	Bovidae	36	2.89	LC	Sca
Klipspringer	<i>Oreotragus oreotragus</i>	Bovidae	60	3.25	LC	Sca
Oribi	<i>Ourebia ourebi</i>	Bovidae	NA	2.52	LC	Sca
Chiru	<i>Pantholops hodgsonii</i>	Bovidae	60	2.69	NT	Sca
Mongolian gazelle	<i>Procapra gutturosa</i>	Bovidae	60	NA	LC	NA
Tibetan gazelle	<i>Procapra picticaudata</i>	Bovidae	58	NA	NT	NA
Przewalski's gazelle	<i>Procapra przewalskii</i>	Bovidae	NA	2.69	EN	Sca
Steenbok	<i>Raphicerus campestris</i>	Bovidae	30	3.14	LC	Sca
Cape grysbok	<i>Raphicerus melanotis</i>	Bovidae	NA	NA	LC	NA
Sharpe's grysbok	<i>Raphicerus sharpei</i>	Bovidae	NA	NA	LC	NA
Saiga	<i>Saiga tatarica</i>	Bovidae	60	2.88	CE	Sca
Ancient bison	<i>Bison antiquus</i>	Bovidae	NA	NA	NA	NA
American bison	<i>Bison bison</i>	Bovidae	60	2.83	NT	S-sca
European bison	<i>Bison bonasus</i>	Bovidae	60	NA	LC	NA
Steppe bison	<i>Bison priscus</i>	Bovidae	NA	NA	NA	NA
Woodland	<i>Bison schoetensacki</i>	Bovidae	NA	NA	NA	NA
Gayal	<i>Bos frontalis</i>	Bovidae	58	3	LC	Sca
Gaur	<i>Bos gaurus</i>	Bovidae	58	NA	VU	NA
Yak	<i>Bos grunniens</i>	Bovidae	60	2.83	LC	Chr
Zebu cattle	<i>Bos indicus</i>	Bovidae	60	2.67	LC	Chr
Banteng	<i>Bos javanicus</i>	Bovidae	60	NA	EN	NA
Wild yak	<i>Bos mutus</i>	Bovidae	60	2.76	VU	Sca
Aurochs	<i>Bos primigenius</i>	Bovidae	60	NA	EX	NA
Kouprey	<i>Bos sauveli</i>	Bovidae	NA	NA	CE	NA
Cattle	<i>Bos taurus</i>	Bovidae	60	2.71	LC	Chr
Nilgai	<i>Boselaphus tragocamelus</i>	Bovidae	46	NA	LC	NA
Water buffalo	<i>Bubalus bubalis</i>	Bovidae	48	2.65	NA	Sca
Carabao	<i>Bubalus carabanensis</i>	Bovidae	NA	NA	NA	NA
Lowland anoa	<i>Bubalus depressicornis</i>	Bovidae	48	NA	EN	NA
Tamaraw	<i>Bubalus mindorensis</i>	Bovidae	46	NA	CE	NA
Mountain anoa	<i>Bubalus quarlesi</i>	Bovidae	46	NA	EN	NA
Kting voar	<i>Pseudonovibos spiralis</i>	Bovidae	NA	NA	NA	NA
Saola	<i>Pseudoryx nghetinhensis</i>	Bovidae	50	NA	CE	NA
African buffalo	<i>Syncerus caffer</i>	Bovidae	52	2.65	NT	Sca
Four-horned antelope	<i>Tetracerus quadricornis</i>	Bovidae	38	NA	VU	NA
Nyala	<i>Tragelaphus angasii</i>	Bovidae	55	NA	LC	NA
Mountain nyala	<i>Tragelaphus buxtoni</i>	Bovidae	NA	3.27	EN	Sca
Giant eland	<i>Tragelaphus derbianus</i>	Bovidae	31/32	NA	VU	NA

Bongo	<i>Tragelaphus eurycerus</i>	Bovidae	33/34	2.96	NT	Sca
Lesser kudu	<i>Tragelaphus imberbis</i>	Bovidae	38	4.04	NT	Sca
Common eland	<i>Tragelaphus oryx</i>	Bovidae	32/31	2.84	LC	Sca
Bushbuck	<i>Tragelaphus scriptus</i>	Bovidae	31/32	2.87	LC	Sca
Sitatunga	<i>Tragelaphus spekii</i>	Bovidae	30	3.72	LC	Sca
Greater kudu	<i>Tragelaphus strepsiceros</i>	Bovidae	32	2.87	LC	Sca
Barbary sheep	<i>Ammotragus lervia</i>	Bovidae	58	2.65	VU	Sca
Arabian tahr	<i>Arabitragus jayakari</i>	Bovidae	NA	NA	EN	NA
Woodland muskox	<i>Bootherium bombifrons</i>	Bovidae	NA	NA	EX	NA
Takin	<i>Budorcas taxicolor</i>	Bovidae	52	NA	VU	NA
Wild goat	<i>Capra aegagrus</i>	Bovidae	60	2.83	VU	Chr
West caucasica tur	<i>Capra caucasica</i>	Bovidae	NA	NA	EN	NA
Markhor	<i>Capra falconeri</i>	Bovidae	60	NA	NT	NA
Goat	<i>Capra hircus</i>	Bovidae	60	3.04	NA	Chr
Alpine ibex	<i>Capra ibex</i>	Bovidae	60	2.7	LC	Sca
Nubian ibex	<i>Capra nubiana</i>	Bovidae	NA	NA	VU	NA
Iberian wild goat	<i>Capra pyrenaica</i>	Bovidae	60	NA	LC	NA
Siberian ibex	<i>Capra sibirica</i>	Bovidae	60	2.73	LC	Sca
Walia	<i>Capra walie</i>	Bovidae	NA	NA	EN	NA
Japanese serou	<i>Capricornis crispus</i>	Bovidae	50	NA	LC	NA
Chinese serow	<i>Capricornis milneedwardsii</i>	Bovidae	NA	NA	NT	NA
Sumatran serow	<i>Capricornis sumatraensis</i>	Bovidae	48	NA	VU	NA
Taiwan serow	<i>Capricornis swinhoei</i>	Bovidae	50	NA	LC	NA
Himalayan serow	<i>Capricornis thar</i>	Bovidae	NA	NA	NT	NA
Nilgiri tahr	<i>Hemitragus hylocrius</i>	Bovidae	58	2.71	EN	Sca
Himalayan tahr	<i>Hemitragus jemlahicus</i>	Bovidae	48	NA	NT	NA
Balearic island cave goat	<i>Myotragus balearicus</i>	Bovidae	NA	NA	EX	NA
Red goral	<i>Naemorhedus baileyi</i>	Bovidae	54/56	NA	VU	NA
Long-tailed goral	<i>Naemorhedus caudatus</i>	Bovidae	56	NA	VU	NA
Himalayan goral	<i>Naemorhedus goral</i>	Bovidae	56	NA	NT	NA
Chinese goral	<i>Naemorhedus griseus</i>	Bovidae	55/56	NA	VU	NA
Mountain goat	<i>Oreamnos americanus</i>	Bovidae	42	NA	LC	NA
Harrington's mountain goat	<i>Oreamnos harringtoni</i>	Bovidae	NA	NA	NA	NA
Muskox	<i>Ovibos moschatus</i>	Bovidae	48	NA	LC	NA
Argali	<i>Ovis ammon</i>	Bovidae	56	2.66	NT	Sca
Sheep	<i>Ovis aries</i>	Bovidae	54	2.77	NA	Chr
Bighorn sheep	<i>Ovis canadensis</i>	Bovidae	54	2.86	LC	Sca
Dall sheep	<i>Ovis dalli</i>	Bovidae	54	NA	LC	NA
Snow sheep	<i>Ovis nivicola</i>	Bovidae	52	NA	LC	NA
Asian mouflon	<i>Ovis orientalis</i>	Bovidae	54	NA	VU	NA

Bharal/Blue sheep	<i>Pseudois nayaur</i>	Bovidae	56	2.58	LC	Sca
Dwarf bharal	<i>Pseudois schaeferi</i>	Bovidae	NA	NA	EN	NA
Pyrenean chamois	<i>Rupicapra pyrenaica</i>	Bovidae	58	NA	LC	NA
Chamois	<i>Rupicapra rupicapra</i>	Bovidae	58	NA	LC	NA
Aders's duiker	<i>Cephalophus adersi</i>	Bovidae	NA	NA	VU	NA
Peter's duiker	<i>Cephalophus callipygus</i>	Bovidae	NA	NA	LC	NA
Bay duiker	<i>Cephalophus dorsalis</i>	Bovidae	60	NA	NT	NA
Harvey's duiker	<i>Cephalophus harveyi</i>	Bovidae	60	2.82	LC	Sca
Jentink's duiker	<i>Cephalophus jentinki</i>	Bovidae	NA	NA	EN	NA
White-bellied duiker	<i>Cephalophus leucogaster</i>	Bovidae	NA	NA	NT	NA
Natal duiker	<i>Cephalophus natalensis</i>	Bovidae	60	NA	LC	NA
Black duiker	<i>Cephalophus niger</i>	Bovidae	60	NA	LC	NA
Black-fronted duiker	<i>Cephalophus nigrifrons</i>	Bovidae	NA	NA	LC	NA
Ogilby's duiker	<i>Cephalophus ogilbyi</i>	Bovidae	NA	NA	LC	NA
Red-flanked duiker	<i>Cephalophus rufilatus</i>	Bovidae	60	NA	LC	NA
Yellow-backed duiker	<i>Cephalophus silvicultor</i>	Bovidae	60	NA	NT	NA
Abbott's duiker	<i>Cephalophus spadix</i>	Bovidae	NA	NA	EN	NA
Weyns's duiker	<i>Cephalophus weynsi</i>	Bovidae	NA	NA	LC	NA
Zebra duiker	<i>Cephalophus zebra</i>	Bovidae	58	NA	VU	NA
Maxwell's duiker	<i>Philantomba maxwellii</i>	Bovidae	60	3.14	LC	Sca
Blue duiker	<i>Philantomba monticola</i>	Bovidae	NA	NA	LC	NA
Walter's duiker	<i>Philantomba walteri</i>	Bovidae	NA	NA	NA	NA
Common duiker	<i>Sylvicapra grimmia</i>	Bovidae	NA	3.14	LC	Sca
Addax	<i>Addax nasomaculatus</i>	Bovidae	58	NA	CE	NA
Roan antelope	<i>Hippotragus equinus</i>	Bovidae	60	NA	LC	NA
Blue antelope	<i>Hippotragus leucophaeus</i>	Bovidae	NA	NA	EX	NA
Sable antelope	<i>Hippotragus niger</i>	Bovidae	60	2.59	LC	S-sca
Beisa oryx	<i>Oryx beisa</i>	Bovidae	58	NA	EN	NA
Scimitar-horned oryx	<i>Oryx dammah</i>	Bovidae	56/58	NA	EXW	NA
Gemsbok	<i>Oryx gazella</i>	Bovidae	56	2.74	LC	C-sca
Arabian oryx	<i>Oryx leucoryx</i>	Bovidae	57/58	NA	VU	NA
Common rhebok	<i>Pelea capreolus</i>	Bovidae	56	NA	NT	NA
Waterbuck	<i>Kobus ellipsiprymnus</i>	Bovidae	50/52	2.89	LC	Sca
Kob	<i>Kobus kob</i>	Bovidae	50	NA	LC	NA
Lechwe	<i>Kobus leche</i>	Bovidae	48	NA	NT	NA
Nile lechwe	<i>Kobus megaceros</i>	Bovidae	52	NA	EN	NA
Puku	<i>Kobus vardonii</i>	Bovidae	NA	NA	NT	NA
Southern reedbuck	<i>Redunca arundinum</i>	Bovidae	NA	NA	LC	NA
Mountain reedbuck	<i>Redunca fulvorufula</i>	Bovidae	56	NA	EN	NA
Bohor reedbuck	<i>Redunca redunca</i>	Bovidae	NA	2.72	LC	Sca

Chital	<i>Axis axis</i>	Cervidae	66	NA	LC	NA
Bawean deer	<i>Axis kuhlii</i>	Cervidae	NA	NA	CE	NA
Hog deer	<i>Axis porcinus</i>	Cervidae	68	2.67	EN	S-sca
Calamian deer	<i>Cervus calamianensis</i>	Cervidae	NA	NA	LC	NA
Red deer	<i>Cervus elaphus</i>	Cervidae	68	3.44	LC	Chr
Cantral Asian red deer	<i>Cervus hanglu</i>	Cervidae	NA	NA	LC	NA
Sika deer	<i>Cervus nippon</i>	Cervidae	66	NA	LC	NA
Fallow deer	<i>Dama dama</i>	Cervidae	68	NA	LC	NA
Mesopotamian fallow deer	<i>Dama mesopotamica</i>	Cervidae	68	NA	EN	NA
Milu	<i>Elaphurus davidianus</i>	Cervidae	68	2.58	EXW	Sca
Irish elk	<i>Megaloceros giganteus</i>	Cervidae	NA	NA	NA	NA
White-lipper deer	<i>Przewalskium albirostris</i>	Cervidae	66	2.69	VU	Sca
Barasingha	<i>Rucervus duvaucelii</i>	Cervidae	56	NA	VU	NA
Eld's deer	<i>Rucervus eldii</i>	Cervidae	58	NA	EN	NA
Schomburgk's deer	<i>Rucervus schomburgki</i>	Cervidae	NA	NA	EX	NA
Visayan spotted deer	<i>Rusa alfredi</i>	Cervidae	NA	NA	EN	NA
Philippine sambar	<i>Rusa marianna</i>	Cervidae	64/65	NA	VU	NA
Java rusa	<i>Rusa timorensis</i>	Cervidae	60	NA	VU	NA
Sambar	<i>Rusa unicolor</i>	Cervidae	60	NA	VU	NA
Chinese water deer	<i>Hydropotes inermis</i>	Cervidae	70	2.53	VU	Sca
Tufted deer	<i>Elaphodus cephalophus</i>	Cervidae	46/48	NA	NT	NA
Bornean yellow muntjac	<i>Muntiacus atherodes</i>	Cervidae	NA	NA	NT	NA
Black muntjac	<i>Muntiacus crinifrons</i>	Cervidae	8/9	2.68	VU	Sca
Fea's muntjac	<i>Muntiacus feae</i>	Cervidae	14/13	NA	NA	NA
Gogshanensis muntjac	<i>Muntiacus gongshanensis</i>	Cervidae	8/9	NA	NA	NA
Indian muntjac	<i>Muntiacus muntjak</i>	Cervidae	6/7	2.57	LC	Chr
Leaf deer	<i>Muntiacus putaoensis</i>	Cervidae	NA	NA	NA	NA
Chinese muntjac	<i>Muntiacus reevesi</i>	Cervidae	46	2.6	LC	S-sca
Roosevelt muntjac	<i>Muntiacus rooseveltorum</i>	Cervidae	6	NA	NA	NA
Annamite muntjac	<i>Muntiacus truongsongensis</i>	Cervidae	NA	NA	NA	NA
Giant muntjac	<i>Muntiacus vuquangensis</i>	Cervidae	NA	NA	CE	NA
Eurasian elk	<i>Alces alces</i>	Cervidae	68	2.74	LC	Sca
American moose	<i>Alces americanus</i>	Cervidae	NA	NA	NA	NA
Marsh deer	<i>Blastocerus dichotomus</i>	Cervidae	66	NA	VU	NA
Roe deer	<i>Capreolus capreolus</i>	Cervidae	70	2.78	LC	Sca
Easter roe deer	<i>Capreolus pygargus</i>	Cervidae	70	NA	LC	NA
Peruvian guemal	<i>Hippocamelus antisensis</i>	Cervidae	NA	NA	VU	NA
Chilean guemal	<i>Hippocamelus bisulcus</i>	Cervidae	70	NA	EN	NA
Red brocket	<i>Mazama americana</i>	Cervidae	42/53	NA	NA	NA

Sao Paulo bororo	<i>Mazama bororo</i>	Cervidae	34	NA	VU	NA
Merida brocket	<i>Mazama bricenii</i>	Cervidae	NA	NA	VU	NA
Dwarf brocket	<i>Mazama chunyi</i>	Cervidae	NA	NA	VU	NA
Gray brocket	<i>Mazama gouazoubira</i>	Cervidae	68/70	NA	LC	NA
Pygmy brocket	<i>Mazama nana</i>	Cervidae	36/40	NA	VU	NA
Amazonian brown brocket	<i>Mazama nemorivaga</i>	Cervidae	NA	NA	LC	NA
Ecuador red brocket	<i>Mazama rufina</i>	Cervidae	NA	NA	VU	NA
Mule deer	<i>Odocoileus hemionus</i>	Cervidae	70	2.34	LC	Sca
White-tailed deer	<i>Odocoileus virginianus</i>	Cervidae	70	2.38	LC	S-sca
Pampas deer	<i>Ozotoceros bezoarticus</i>	Cervidae	68	NA	NT	NA
Northern pudu	<i>Pudu mephistophiles</i>	Cervidae	NA	NA	NA	NA
Pudu	<i>Pudu puda</i>	Cervidae	70	NA	NT	NA
Reindeer	<i>Rangifer tarandus</i>	Cervidae	70	2.54	VU	Sca
Giraffe	<i>Giraffa camelopardalis</i>	Giraffidae	30	2.44	VU	C-sca
South African giraffe	<i>Giraffa giraffa</i>	Giraffidae	30	NA	NA	NA
Giraffe	<i>Giraffa reticulata</i>	Giraffidae	30	NA	NA	NA
Giraffe	<i>Giraffa tippelskirchi</i>	Giraffidae	30	2.49	NA	Chr
Okapi	<i>Okapia johnstoni</i>	Giraffidae	46/45	2.89	EN	S-sca
Anhui musk deer	<i>Moschus anhuiensis</i>	Moschidae	NA	NA	EN	NA
Forest musk deer	<i>Moschus berezovskii</i>	Moschidae	58	2.73	EN	S-sca
Alpine musk deer	<i>Moschus chrysogaster</i>	Moschidae	NA	4.97	EN	Sca
Kashmir muskdeer	<i>Moschus cupreus</i>	Moschidae	NA	NA	EN	NA
Black musk deer	<i>Moschus fuscus</i>	Moschidae	NA	NA	EN	NA
Himalayan musk deer	<i>Moschus leucogaster</i>	Moschidae	NA	NA	EN	NA
Siberian musk deer	<i>Moschus moschiferus</i>	Moschidae	58	3.07	VU	Sca
Water chevrotain	<i>Hyemoschus aquaticus</i>	Tragulidae	NA	NA	LC	NA
Indian chevrotain	<i>Moschiola indica</i>	Tragulidae	NA	NA	LC	NA
White-spotted chevrotain	<i>Moschiola meminna</i>	Tragulidae	NA	NA	LC	NA
Java mouse-deer	<i>Tragulus javanicus</i>	Tragulidae	32	NA	NA	Chr
Lesser mouse-deer	<i>Tragulus kanchil</i>	Tragulidae	32	3.05	LC	Sca
Greater Oriental chevrotain	<i>Tragulus napu</i>	Tragulidae	32	NA	LC	NA
Williamson's chevrotain	<i>Tragulus williamsoni</i>	Tragulidae	NA	NA	NA	NA

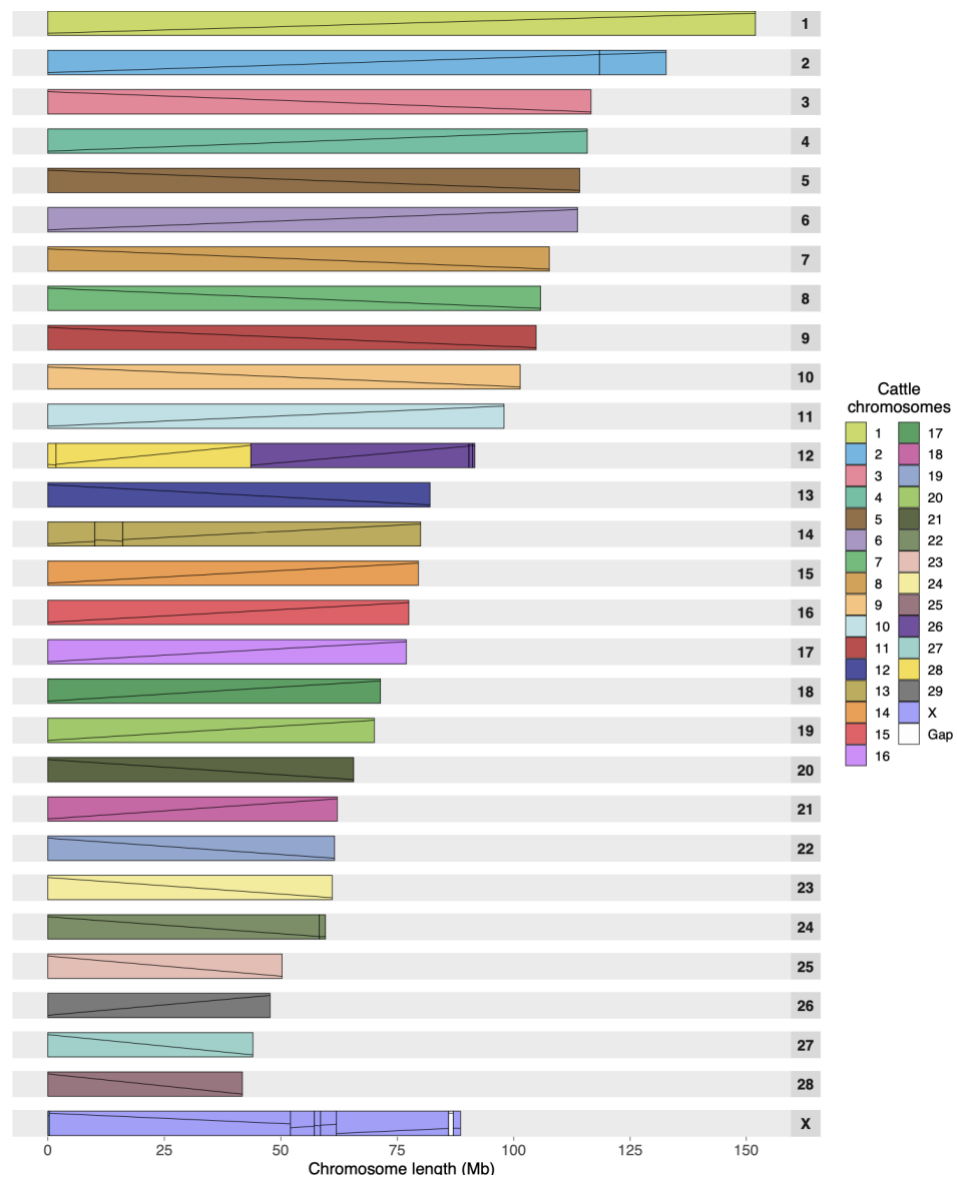


Figure S1. Ideogram comparing BOCE with cattle chromosomes. Ancestral chromosomes are colour-coded using cattle chromosomes.

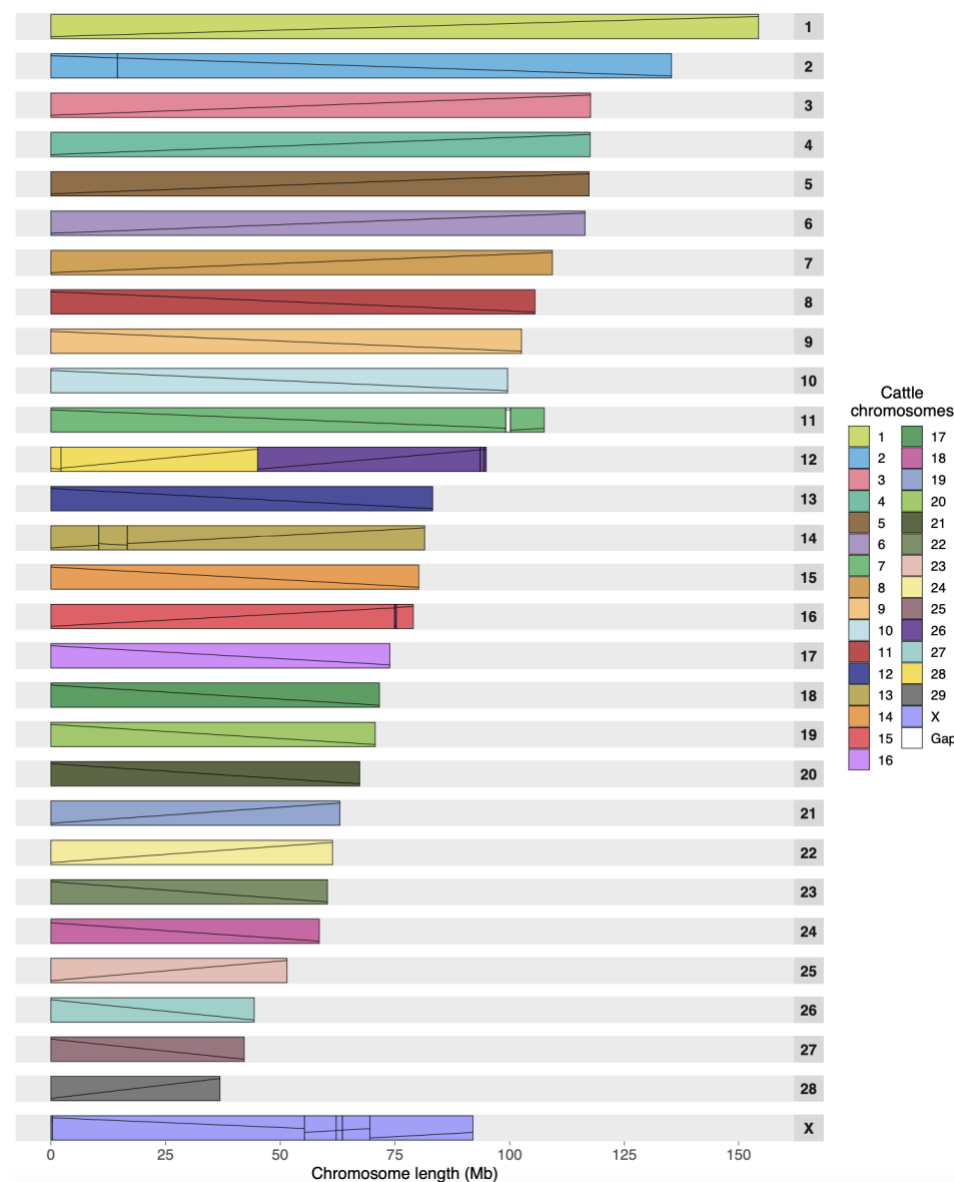


Figure S2. Ideogram comparing BOMO with cattle chromosomes. Ancestral chromosomes are colour-coded using cattle chromosomes.

Table S2. Genes found in Tasmanian devil well-defined EBRs expanded up- and downstream until reach 50 Kb. Genome location indicates where the gene is located in Tasmanian devil genome, following the structure: Chromosome: start coordinate - end coordinate. N/A means not available.

Genome location	Gene name	Ensembl IDs	Type of gene
1: 246704964 - 246735249	N/A	ENSSHAG00000028339	Protein coding
1: 346276743 - 346386902	N/A	ENSSHAG00000000633	Protein coding
2: 168173722 - 168225605	N/A	ENSSHAG00000022136	Protein coding
2: 491603869 - 491604309	N/A	ENSSHAG00000004591	Protein coding
3: 229537199 - 229591483	N/A	ENSSHAG00000002711	Protein coding
3: 427583259 - 427609246	N/A	ENSSHAG00000014382	Protein coding
3: 427647153 - 427703097	N/A	ENSSHAG00000013973	Protein coding
3: 530224792 - 530258035	N/A	ENSSHAG00000003884	Protein coding
3: 530290224 - 530321186	N/A	ENSSHAG00000000434	Protein coding
4: 117139891 - 117269720	N/A	ENSSHAG00000009793	Protein coding
4: 392584337 - 392607917	N/A	ENSSHAG00000022205	Protein coding
4: 392631755 - 392763626	N/A	ENSSHAG00000021067	Protein coding
5: 68949481 - 69028090	N/A	ENSSHAG00000007145	Protein coding

Table S3. Genes found in Tasmanian devil not well-defined EBRs. Genome location indicates where the gene is located in Tasmanian devil genome, following the structure: Chromosome: start coordinate - end coordinate. N/A means not available.

Genome location	Gene name	Ensembl IDs	Type of gene
1: 349775366 - 349780456	N/A	ENSSHAG00000026957	Protein coding
3: 189258971 - 189393142	N/A	ENSSHAG00000000850	Protein coding
3: 233911168 - 234045175	N/A	ENSSHAG000000008796	Protein coding
3: 234119923 - 234149762	N/A	ENSSHAG00000010304	Protein coding
5: 2327635 - 2328743	N/A	ENSSHAG00000029015	Protein coding
5: 57641955 - 57688960	N/A	ENSSHAG00000007931	Protein coding
6: 233835636 - 233842252	N/A	ENSSHAG00000026132	Protein coding
6: 233844111 - 233847239	N/A	ENSSHAG00000026544	Protein coding
6: 148584707 - 148643622	N/A	ENSSHAG00000021174	Protein coding
X: 79582316 - 79643770	N/A	ENSSHAG00000026190	Protein coding
X: 79693788 - 79715327	N/A	ENSSHAG00000001331	Protein coding

Table S4. Classification of HSBs in African elephant.

Chromosome	Start	End	Classification
1	16,237	35,683,919	Fission + Fusion
1	35,683,919	236,189,003	Fusion + Collinear
2	133,453	7,044,311	Fission + Inversion
2	7,044,311	8,044,627	Fission + Inversion
2	8,547,838	32,136,952	Fission + Collinear
2	32,252,149	57,951,989	Fission + Inversion
2	60,991,767	73,823,636	Fission + Inversion
2	74,430,870	187,721,299	Fission + Collinear
2	191,585,620	195,797,255	Fission + Collinear
2	195,797,255	196,303,304	Fission + Inversion
2	196,310,952	204,336,147	Fission + Collinear
3	1,549,027	18,201,227	Fission + Fusion
3	25,321,698	25,850,296	Fusion + Inversion
3	26,149,352	26,697,441	Fusion + Inversion
3	27,852,423	28,686,827	Fusion + Collinear
3	29,598,280	31,947,210	Fission + Inversion
3	32,134,917	86,342,620	Fusion + Collinear
3	86,353,146	109,311,191	Fusion + Inversion
3	109,674,999	231,977,993	Fusion + Collinear
4	107,005	15,647,569	Fission + Inversion
4	16,911,327	189,233,634	Fission + Collinear
5	-	135,860,597	Fusion + Collinear
5	136,196,365	162,338,110	Fission + Fusion
6	27,827	126,550,135	Collinear
6	126,550,135	138,821,298	Inversion
7	3	117,608,024	Fission + Collinear
8	83,401	120,487,352	Fusion + Collinear
8	120,564,205	132,087,530	Fission + Fusion
9	98	21,384,137	Fission + Inversion
9	22,168,733	22,843,349	Fission + Collinear
9	23,449,786	112,956,954	Fission + Inversion
10	10,147	5,280,341	Collinear
10	5,331,628	10,709,192	Inversion
10	11,553,848	96,421,332	Collinear
10	96,421,332	96,728,078	Inversion
10	96,733,561	109,222,807	Collinear
10	111,792,997	119,812,541	Inversion
11	37,853	72,048,915	Fission + Fusion

11	72,048,915	78,154,021	Fission + Fusion
11	78,154,021	80,695,907	Fission + Fusion
12	312,883	46,398,607	Fission + Fusion
12	46,398,607	52,049,514	Fission + Fusion
12	52,049,514	87,726,135	Fusion + Collinear
12	87,734,379	103,950,477	Fusion + Inversion
13	3,982	28,069,529	Fission + Fusion
13	28,292,559	89,182,736	Fusion + Inversion
13	89,439,513	97,605,695	Fusion + Collinear
13	97,876,176	107,851,238	Fusion + Inversion
14	62,182	36,581,479	Fission + Fusion
14	36,663,960	68,881,911	Fission + Fusion
14	68,881,911	85,385,163	Fission + Fusion
15	16,141	5,205,107	Fission + Inversion
15	5,232,803	6,738,908	Fission + Collinear
15	7,202,373	92,357,247	Fission + Collinear
16	570,401	27,203,528	Fission + Fusion
16	27,219,292	27,521,797	Fission + Fusion
16	27,656,707	75,914,372	Fission + Fusion
16	75,914,372	80,536,365	Fission + Fusion
17	571	59,993,475	Fission + Fusion
17	59,993,475	91,560,434	Fission + Fusion
18	3,597	82,688,476	Fusion + Collinear
18	82,688,476	83,021,036	Fission + Fusion
18	83,231,934	83,831,754	Fission + Fusion
19	754,329	29,373,876	Fission + Collinear
19	29,373,876	64,430,908	Fission + Collinear
20	2,697	11,100,256	Fission + Fusion
20	11,335,978	49,537,097	Fission + Fusion
20	52,810,827	76,256,137	Fission + Fusion
21	24,560	58,719,085	Fission + Fusion
21	59,188,946	78,238,945	Fission + Fusion
22	9	12,512,852	Fission + Inversion
22	12,512,852	44,703,683	Fission + Collinear
22	44,703,683	71,661,433	Fission + Collinear
23	7,375	63,807,622	Collinear
24	168	24,167,559	Fission + Inversion
24	24,167,559	28,443,243	Fission + Collinear
25	1,254	43,590,696	Fission + Fusion
25	43,958,174	56,207,045	Fission + Fusion
25	56,969,504	58,089,047	Fission + Fusion
25	58,232,572	60,222,137	Fission + Fusion
25	60,262,820	78,619,778	Fission + Fusion

26	4	16,042,484	Fission + Fusion
26	16,042,484	33,616,572	Fission + Fusion
26	33,616,572	55,588,453	Fission + Fusion
26	55,588,453	75,824,685	Fission + Fusion
27	-	37,203,799	Fission + Fusion
27	37,203,799	47,134,923	Fission + Fusion
X	9,008	70,317,911	Collinear
X	70,510,779	71,327,583	Inversion
X	71,350,950	103,444,847	Collinear
X	103,444,847	107,607,452	Inversion
X	107,607,452	172,689,940	Inversion

Table S5. Genes showing significant P-values in Group 1 after running RERConverge. Genes in green are all genes shared by Group 1 and Group 2.

Gene ID	P-value adj.	Gene ID	P-value adj.	Gene ID	P-value adj.
ADD2	0.2580538	LRRC41	0.2962082	FRMD5	0.3329923
AMOTL1	0.2580538	LSM1	0.2962082	GATA5	0.3329923
ATL3	0.2580538	NDUFB10	0.2962082	GSR	0.3329923
C3H1orf94	0.2580538	NEDD9	0.2962082	HDHD5	0.3329923
DEPDC1	0.2580538	NUDT12	0.2962082	IQSEC2	0.3329923
DNA2	0.2580538	P3H2	0.2962082	KLF6	0.3329923
EHMT1	0.2580538	PEAK1	0.2962082	MED24	0.3329923
EIF2B5	0.2580538	PHETA2	0.2962082	METTL26	0.3329923
FAM160B2	0.2580538	PKD2	0.2962082	MPI	0.3329923
FBXO15	0.2580538	PLCD3	0.2962082	MRPL19	0.3329923
GPKOW	0.2580538	PRKCD	0.2962082	MRPL54	0.3329923
IL12RB2	0.2580538	PRSS12	0.2962082	NCEH1	0.3329923
JMJD1C	0.2580538	PTPN11	0.2962082	NFE2L3	0.3329923
MROH9	0.2580538	RASSF4	0.2962082	PPP1R32	0.3329923
NAA25	0.2580538	RBM33	0.2962082	PRDM1	0.3329923
NAIF1	0.2580538	RCN2	0.2962082	RMND5B	0.3329923
NKAPD1	0.2580538	RNF103	0.2962082	RRP7A	0.3329923
PCIF1	0.2580538	SCRN3	0.2962082	SBK1	0.3329923
PLCL1	0.2580538	SH3TC1	0.2962082	SCARA3	0.3329923
RNPEP	0.2580538	SHCBP1L	0.2962082	SEMA6D	0.3329923
RTKN2	0.2580538	SLC39A3	0.2962082	SFRS18	0.3329923
SAAL1	0.2580538	SNAPC4	0.2962082	SIPA1L1	0.3329923
SAMD4B	0.2580538	SNED1	0.2962082	SLC26A3	0.3329923
SLC16A13	0.2580538	SNX25	0.2962082	SLMAP	0.3329923
SLC16A9	0.2580538	STAC	0.2962082	ST8SIA5	0.3329923
SLC8B1	0.2580538	TEKT4	0.2962082	TDRD1	0.3329923
SPACA9	0.2580538	TFR2	0.2962082	TMEM101	0.3329923

SREBF2	0.2580538	TRAF3	0.2962082	TRIM42	0.3329923
STAU1	0.2580538	UNC5A	0.2962082	TSPAN7	0.3329923
TCTN2	0.2580538	ZBTB18	0.2962082	TTI1	0.3329923
ZFYVE1	0.2580538	ZMIZ2	0.2962082	VTN	0.3329923
HYDIN	0.2604492	ZW10	0.2962082	ZCCHC2	0.3329923
MEX3D	0.2604492	C28H1orf198	0.3028731	ZMYND19	0.3329923
MKRN2	0.2604492	C3H1orf56	0.3028731	ZNF503	0.3329923
RASEF	0.2604492	CA6	0.3028731	ACAD9	0.3370996
SIX1	0.2604492	CAPN10	0.3028731	ADARB2	0.3370996
ZNF703	0.2604492	CASD1	0.3028731	AMDHD1	0.3370996
DGKE	0.2630758	DNAAF5	0.3028731	ATP1B2	0.3370996
DNTT	0.2630758	DNPEP	0.3028731	C11H2orf68	0.3370996
GLI2	0.2630758	DPP3	0.3028731	C19H17orf49	0.3370996
LAMP3	0.2630758	EXOSC4	0.3028731	CDH13	0.3370996
PCNX1	0.2630758	FIGN	0.3028731	CHRNA2	0.3370996
PDHX	0.2630758	GCGR	0.3028731	CLN5	0.3370996
RIN2	0.2630758	GRINA	0.3028731	CLUAP1	0.3370996
ANKRD10	0.266129	HAAO	0.3028731	CNNM2	0.3370996
BCAR3	0.266129	IQCA1	0.3028731	CNPY4	0.3370996
BCL9L	0.266129	KCNK4	0.3028731	CWF19L1	0.3370996
BNC1	0.266129	KDM4B	0.3028731	EIPR1	0.3370996
CCDC148	0.266129	KLC4	0.3028731	FOXB2	0.3370996
CDH17	0.266129	LTBP3	0.3028731	FOXN1	0.3370996
CHST13	0.266129	PRPF39	0.3028731	GGA3	0.3370996
COL6A3	0.266129	R3HCC1	0.3028731	GRIP2	0.3370996
CYP27A1	0.266129	REXO1	0.3028731	HGS	0.3370996
DMRTC2	0.266129	RGS4	0.3028731	HSCB	0.3370996
DPH1	0.266129	SLC2A4RG	0.3028731	ILKAP	0.3370996
EGFLAM	0.266129	SLCO4A1	0.3028731	IMP4	0.3370996
FAXC	0.266129	STK17A	0.3028731	KCNB2	0.3370996
FGF1	0.266129	TCHP	0.3028731	KDM5A	0.3370996
GOLGA2	0.266129	TMF1	0.3028731	LPCAT4	0.3370996
KAT7	0.266129	TRPM6	0.3028731	LRP2	0.3370996
KCNK9	0.266129	URGCP	0.3028731	LSMEM2	0.3370996
KIAA1109	0.266129	AGXT2	0.3088533	MCOLN3	0.3370996
KIAA1324L	0.266129	ALDOB	0.3088533	NDFIP1	0.3370996
LDLRAP1	0.266129	BCORL1	0.3088533	NEK11	0.3370996
LGALS4	0.266129	CD3EAP	0.3088533	NEURL1	0.3370996
LSG1	0.266129	CDK5RAP2	0.3088533	NR1D1	0.3370996
MACO1	0.266129	CDON	0.3088533	OPLAH	0.3370996
MRC1	0.266129	COL20A1	0.3088533	OSBPL7	0.3370996
NAGK	0.266129	DDX20	0.3088533	PDZRN3	0.3370996
NUP107	0.266129	DKK4	0.3088533	PHLDB2	0.3370996

PHF7	0.266129	DLX3	0.3088533	PHYHIP	0.3370996
PIAS4	0.266129	EFHC1	0.3088533	PLPPR3	0.3370996
PIP4P1	0.266129	FAM69A	0.3088533	PPCDC	0.3370996
RALGAPB	0.266129	GGT7	0.3088533	PTPN12	0.3370996
RBM45	0.266129	GPAT2	0.3088533	PYGO2	0.3370996
RNF10	0.266129	ID3	0.3088533	RADIL	0.3370996
RNF180	0.266129	INO80	0.3088533	RNF169	0.3370996
RNF208	0.266129	LGALS1	0.3088533	SEC62	0.3370996
SKOR1	0.266129	LRRK2	0.3088533	SLC30A5	0.3370996
SP6	0.266129	MUTYH	0.3088533	SOX8	0.3370996
SPACA1	0.266129	NAV1	0.3088533	SPATC1	0.3370996
STAT1	0.266129	NCF1	0.3088533	SPICE1	0.3370996
SYNRG	0.266129	NRG1	0.3088533	USP40	0.3370996
TNXB	0.266129	PCDH11X	0.3088533	USP42	0.3370996
TRIM36	0.266129	PID1	0.3088533	VRK1	0.3370996
UBA6	0.266129	RNF123	0.3088533	WASHC4	0.3370996
ZNF521	0.266129	SFN	0.3088533	ABHD16B	0.3494354
ZSWIM4	0.266129	SPATA2	0.3088533	ACSS2	0.3494354
ANKRD2	0.2829914	SYTL1	0.3088533	ACVR1C	0.3494354
ARAP1	0.2829914	TTLL8	0.3088533	ADAM17	0.3494354
ATXN7L2	0.2829914	WNT11	0.3088533	ALG6	0.3494354
CCDC173	0.2829914	YTHDC1	0.3088533	ASH1L	0.3494354
ISM1	0.2829914	ZBTB41	0.3088533	BFSP2	0.3494354
KCNQ2	0.2829914	ZNF438	0.3088533	C7H19orf70	0.3494354
MSH6	0.2829914	ABL1	0.3188047	CDIPT	0.3494354
MYPN	0.2829914	AK4	0.3188047	CEP76	0.3494354
PUM3	0.2829914	AKIRIN1	0.3188047	CLIP2	0.3494354
SOX9	0.2829914	BAHCC1	0.3188047	COL15A1	0.3494354
ABCC10	0.2856472	CGREF1	0.3188047	DMP1	0.3494354
ADAM33	0.2856472	CHERP	0.3188047	DOPEY1	0.3494354
AHR	0.2856472	COG3	0.3188047	EGR1	0.3494354
ATL1	0.2856472	DCAF12	0.3188047	FGF22	0.3494354
BRD1	0.2856472	DNAH2	0.3188047	IFT172	0.3494354
CDC7	0.2856472	FAM83E	0.3188047	IGFBP3	0.3494354
CDHR1	0.2856472	FASTKD1	0.3188047	KCNS2	0.3494354
CLEC11A	0.2856472	FHAD1	0.3188047	KCTD20	0.3494354
COL8A1	0.2856472	FUT11	0.3188047	KIAA2026	0.3494354
COLEC12	0.2856472	GATAD2B	0.3188047	MINK1	0.3494354
DGKZ	0.2856472	GRXCR2	0.3188047	MRPS14	0.3494354
DHX37	0.2856472	HSD11B2	0.3188047	MS4A1	0.3494354
EDAR	0.2856472	IQCH	0.3188047	MSH3	0.3494354
FAM57A	0.2856472	KIAA1024	0.3188047	NUP155	0.3494354
FKBP14	0.2856472	LCTL	0.3188047	OVCH2	0.3494354

FLRT3	0.2856472	MSH4	0.3188047	PDIA3	0.3494354
FLVCR1	0.2856472	MTERF4	0.3188047	PLOD3	0.3494354
GFOD1	0.2856472	MTSS1L	0.3188047	RGS16	0.3494354
GRN	0.2856472	NKX2-2	0.3188047	SLC5A9	0.3494354
IGSF3	0.2856472	PLAT	0.3188047	SOX30	0.3494354
MUT	0.2856472	PLCL2	0.3188047	SPHKAP	0.3494354
NOL12	0.2856472	S1PR1	0.3188047	SUSD2	0.3494354
PLAG1	0.2856472	SLC1A2	0.3188047	SUSD5	0.3494354
RBM44	0.2856472	SMC1B	0.3188047	TMEM131	0.3494354
SLC7A5	0.2856472	TATDN3	0.3188047	TMEM45B	0.3494354
SSC5D	0.2856472	TCTN1	0.3188047	TMEM94	0.3494354
STARD6	0.2856472	UPK2	0.3188047	TOM1	0.3494354
STC1	0.2856472	USP3	0.3188047	TRAPPC10	0.3494354
TM6SF1	0.2856472	ZBBX	0.3188047	TRIM69	0.3494354
TMEM245	0.2856472	ABCB8	0.3281838	TRPS1	0.3494354
TMPRSS7	0.2856472	ABCC5	0.3281838	USP32	0.3494354
TRMT6	0.2856472	ANAPC1	0.3281838	XPR1	0.3494354
TTC21A	0.2856472	APAF1	0.3281838	ZNF277	0.3494354
TUFT1	0.2856472	ASXL1	0.3281838	ZZEF1	0.3494354
UBAP1	0.2856472	B4GALNT1	0.3281838	ACVR1	0.3658906
WDR66	0.2856472	BAIAP2L2	0.3281838	ALPK2	0.3658906
XPNPEP3	0.2856472	CCNT2	0.3281838	ALPK3	0.3658906
ADCYAP1	0.2865843	COG1	0.3281838	APCDD1	0.3658906
AP5Z1	0.2865843	CREB3L3	0.3281838	ARID4B	0.3658906
CCDC146	0.2865843	CYP27B1	0.3281838	ASAH1	0.3658906
CFAP58	0.2865843	DOT1L	0.3281838	ASPHD2	0.3658906
DNAJB2	0.2865843	ELFN2	0.3281838	ASRGL1	0.3658906
DPP9	0.2865843	GPCPD1	0.3281838	C15H11orf16	0.3658906
EPN3	0.2865843	GPR12	0.3281838	C16H1orf115	0.3658906
FIP1L1	0.2865843	GUCY2C	0.3281838	COPS3	0.3658906
HOXD12	0.2865843	HMGXB4	0.3281838	DAB2	0.3658906
KCNJ10	0.2865843	IKZF4	0.3281838	DEPDC5	0.3658906
LRRC59	0.2865843	KIF4A	0.3281838	FAAP100	0.3658906
MAP3K3	0.2865843	MFSB8	0.3281838	GANC	0.3658906
MAP3K4	0.2865843	MRPL32	0.3281838	GP1BA	0.3658906
MRM3	0.2865843	PAPD7	0.3281838	GSX2	0.3658906
NSMF	0.2865843	PLEKHA8	0.3281838	GUCD1	0.3658906
PEX2	0.2865843	PPEF1	0.3281838	HSPB7	0.3658906
RBM47	0.2865843	PPM1J	0.3281838	LCAT	0.3658906
SDCCAG8	0.2865843	PRRX1	0.3281838	LIPT2	0.3658906
SLC2A8	0.2865843	SPRYD4	0.3281838	LRRC43	0.3658906
TBL2	0.2865843	STOX2	0.3281838	LTA4H	0.3658906
TMEM110	0.2865843	TBX5	0.3281838	MAN2A2	0.3658906

TRUB1	0.2865843	TDRKH	0.3281838	MGST1	0.3658906
ZNF652	0.2865843	TNIP2	0.3281838	MUL1	0.3658906
ALG5	0.2962082	TNRC6B	0.3281838	NCSTN	0.3658906
ATXN7L1	0.2962082	TPD52L2	0.3281838	NECAP2	0.3658906
BEND4	0.2962082	UBAP2L	0.3281838	NME7	0.3658906
DDX54	0.2962082	XPO5	0.3281838	NPFFR2	0.3658906
DENND4A	0.2962082	YEATS2	0.3281838	PARP11	0.3658906
DHX34	0.2962082	ZNF629	0.3281838	PDE10A	0.3658906
EEFSEC	0.2962082	ABHD5	0.3329923	PLA2G3	0.3658906
EMILIN2	0.2962082	ARPP21	0.3329923	PLEKHG6	0.3658906
FAM227A	0.2962082	BBS1	0.3329923	PRSS53	0.3658906
GATA2	0.2962082	C16H1orf174	0.3329923	RAD54L2	0.3658906
GRAP2	0.2962082	CCDC18	0.3329923	SEMA4C	0.3658906
GUSB	0.2962082	CCDC39	0.3329923	SLC35B2	0.3658906
HACD3	0.2962082	CLASRP	0.3329923	SLC35F6	0.3658906
HEATR5A	0.2962082	CLSPN	0.3329923	STX3	0.3658906
KIF13A	0.2962082	CPN1	0.3329923	TMEM170A	0.3658906
KLHL8	0.2962082	DCN	0.3329923	VWA5B1	0.3658906
KYAT3	0.2962082	DHX33	0.3329923	WDR12	0.3658906
L3MBTL3	0.2962082	DNAH12	0.3329923	ZBPB2	0.3658906
LOC527796	0.2962082				
