

Contribution of structural variation to adaptive evolution of mammalian genomes

Lorena Derežanin

Publikationsbasierte Dissertation

zur Erlangung des akademischen Grades
"doctor rerum naturalium" (Dr. rer. nat.)
in der Wissenschaftsdisziplin "Evolutionsbiologie"

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
Institut für Biochemie und Biologie
der Universität Potsdam

Disputationsdatum: 27.02.2023
Hauptbetreuer: Prof. Dr. Jörns Fickel (University of Potsdam)
Weitere Gutachter: Prof. Dr. Ralph Tiedemann (University of Potsdam) and Assoc. Prof.
Paulo Célio Alves (University of Porto, Portugal)

This work is protected by copyright and/or related rights. You are free to use this work in any way that is permitted by the copyright and related rights legislation that applies to your use. For other uses you need to obtain permission from the rights-holder(s).
<https://rightsstatements.org/page/InC/1.0/?language=en>

Published online on the
Publication Server of the University of Potsdam:
<https://doi.org/10.25932/publishup-59144>
<https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-591443>

Declaration of authorship

I, Lorena Derežanin, hereby declare that the thesis has not previously been submitted to any other university and it has been prepared independently and exclusively with the specified resources.

This thesis is based on the following manuscripts:

1. Derežanin L, Blažytė A, Dobrynin P, Duchêne DA, Grau JH, Jeon S, Kliver S, Koepfli K-P, Meneghini D, Preick M, Tomarovsky A, Totikov A, Fickel J, & Förster DW (2022) “**Multiple types of genomic variation contribute to adaptive traits in the mustelid subfamily Guloninae**”, *Molecular Ecology* 31, 2898 - 2919. <https://doi.org/10.1111/mec.16443>

I conceived the study, conducted genome assembly, gene family evolution and structural variation analyses, interpreted the data, and wrote the manuscript.

2. Derežanin L*, Safanova Y*, Kliver S, Fonsere C, Serres-Armero A, Tomarovsky A, Totikov A, Etherington G, Haerty W, Di Palma F, Perelman PL, Beklemisheva V, Serdyukova N, Graphodatsky A, Melo-Ferreira J, Marinari P, Marques-Bonet T, Fickel J, Förster DW, Koepfli K-P (2022) “**Comparative analyses inform the genomic consequences of the population bottleneck in the endangered black-footed ferret**” (manuscript prepared for submission to *Current Biology*)
*These authors contributed equally.

My contribution to this study encompasses structural variation analysis (presented in this thesis), interpretation of the results, writing, revising, and editing of the manuscript.

3. Palma-Vera SE, Reyer H, Langhammer M, Reinsch N, Derežanin L, Fickel J, Qanbari S, Weitzel J, Franzenburg S, Hemmrich-Stanisak G, Schön J (2021) “**Genomic characterization of world’s longest selection experiment in mouse reveals the complexity of polygenic traits**”. *BMC Biology* 20, 52, 10.1186/s12915-022-01248-9

I conducted structural variation analysis, interpretation of the results and assisted in writing and revising the manuscript.

4. Totikov A, Tomarovsky A, Prokopov D, Yakupova A, Bulyonkova T, Derežanin L, Rasskazov D, Wolfsberger WW, Koepfli K-P, Oleksyk TK, Kliver S (2021) “**Chromosome-Level Genome Assemblies Expand Capabilities of Genomics for Conservation Biology**”, *Genes*, 12(9), 1336, 10.3390/genes12091336

I conceptualized the project idea together with Sergei Kliver and assisted with writing, revising, and editing the manuscript.

Acknowledgements

The work presented in this thesis is brought to life by the unstinting effort of many people. I wish to thank them all for the help and support along my scientific journey.

I am foremost grateful to my supervisors, Prof. Dr. Jörns Fickel and Dr. Daniel Förster, for the opportunity to work on exciting projects in a field I was previously not all that familiar with. Thank you for your generous support and guidance throughout the years of my development as a scientist. I immensely appreciate our brainstorming sessions about new ideas and approaches, as well as our nerdy discussions about sci-fi books, movies and other similar topics. Your enthusiasm for science, high expectations and great attention to detail have motivated me to think better, write better and strive to be a better researcher. Thank you for being great mentors. I still owe you that poster, though! :)

Furthermore, I would like to thank all my collaborators, who helped shape our research studies with their valuable contributions and feedback. In particular, I would like to acknowledge Dr. Klaus-Peter Koepfli and Sergei Kliver for their unsparing support and insightful advice, especially in the final months of my thesis preparation. Additionally, I thank Dr. Koepfli for the opportunity to visit and work at the Smithsonian Conservation Biology Institute and my supervisors for the outstanding support for this voyage.

Likewise, I want to thank my colleagues from the Department of Evolutionary Genetics at the IZW. Thank you for your help in the lab (especially Suse and Anke), thoughtful discussions during our group meetings and for being a pretty awesome bunch of people. Of course, I will mostly miss our meticulously organized theme parties on the Institute premises. Please, continue this quirky tradition with the highest standards possible.

To friends, fellow PhD students and postdocs - Paula, Cecília, Kseniia, Liam, Shannon, Leon, Sónia, Saba, Michał, John, Riddhi, and many more, thank you for all the hangouts, game nights and gigs. You are making my (PhD) life way more cheerful and lively. Stay gold!

A special appreciation to my family and friends from Croatia who wholeheartedly supported me in this endeavour. I am grateful to you for listening to neverending monologues about my projects. Thank you for showing interest, nodding along and overall being a great source of stress relief and laughter. I hope to spend many such moments with you all!

Lastly, the biggest thank you goes to Zoltán, my husband and a supplier of the best code hacking tips and tricks. I am forever grateful for your patience, advice, our late-night discussions and shared joys and sorrows. Thank you for being my beloved home office mate during the long days of the pandemic and beyond. You being here got me through this, thank you for always caring. I excitedly look forward to the next big adventure ahead of us. Drago mi je da te imam!

Table of contents

Summary	8
Zusammenfassung	10
General Introduction	12
Adaptive evolution of mammalian genomes	12
The role of structural variation in shaping biological diversity	14
History of structural variant identification	15
Types of structural variants	16
Adaptive variation in protein-coding regions	20
Methods for structural variant characterization	22
Reference genome improvements and conservation genomics	27
Aims of study	29
Chapter I	32
Multiple types of genomic variation contribute to adaptive traits in the mustelid subfamily Guloninae	32
Chapter II	56
Comparative analyses inform the genomic consequences of the population bottleneck in the endangered black-footed ferret	56
Chapter III	68
Genomic characterization of the world’s longest selection experiment in mouse reveals the complexity of polygenic traits	68
Chapter IV	90
Chromosome-Level Genome Assemblies Expand Capabilities of Genomics for Conservation Biology	90
General discussion	109
Aims and importance of this dissertation	109
References	122
Appendix A.	136
Appendix B.	164

List of abbreviations

bp - base pairs
CDS - coding sequence
CNV - copy number variation
DEL - deletion
DUP - duplication
FDR - false discovery rate
GEMs - gel beads in emulsion
GWAS - genome-wide association studies
INS - insertion
INV - inversion
kbp - kilobase pairs
Mbp - megabase pairs
NGS - next-generation sequencing
PE - paired-end
QTL - quantitative trait loci
ROH - runs of homozygosity
SD - segmental duplication
SMRT - Single Molecule, Real-Time (sequencing)
SNP - single nucleotide polymorphism
SV - structural variation
TE - transposable element
WGS - whole-genome sequencing

Summary

Following the extinction of dinosaurs, the great adaptive radiation of mammals occurred, giving rise to an astonishing ecological and phenotypic diversity of mammalian species. Even closely related species often inhabit vastly different habitats, where they encounter diverse environmental challenges and are exposed to different evolutionary pressures. As a response, mammals evolved various adaptive phenotypes over time, such as morphological, physiological and behavioural ones. Mammalian genomes vary in their content and structure and this variation represents the molecular mechanism for the long-term evolution of phenotypic variation. However, understanding this molecular basis of adaptive phenotypic variation is usually not straightforward.

The recent development of sequencing technologies and bioinformatics tools has enabled a better insight into mammalian genomes. Through these advances, it was acknowledged that mammalian genomes differ more, both within and between species, as a consequence of structural variation compared to single-nucleotide differences. Structural variant types investigated in this thesis - such as **deletion, duplication, inversion and insertion**, represent a change in the structure of the genome, impacting the size, copy number, orientation and content of DNA sequences. Unlike short variants, structural variants can span multiple genes. They can alter gene dosage, and cause notable gene expression differences and subsequently phenotypic differences. Thus, they can lead to a more dramatic effect on the fitness (reproductive success) of individuals, local adaptation of populations and speciation.

In this thesis, I investigated and evaluated the potential functional effect of structural variations on the genomes of mustelid species. To detect the genomic regions associated with phenotypic variation I assembled the first reference genome of the tayra (*Eira barbara*) relying on linked-read sequencing technology to achieve a high level of genome completeness important for reliable structural variant discovery. I then set up a bioinformatics pipeline to conduct a comparative genomic analysis and explore variation between mustelid species living in different environments. I found numerous genes associated with species-specific phenotypes related to diet, body condition and reproduction among others, to be impacted by structural variants.

Furthermore, I investigated the effects of artificial selection on structural variants in mice selected for high fertility, increased body mass and high endurance. Through selective breeding of each mouse line, the desired phenotypes have spread within these populations, while maintaining structural variants specific to each line. In comparison to the control line, the litter size has doubled in the fertility lines, individuals in the high body mass lines have become considerably larger, and mice selected for treadmill performance covered substantially more distance. Structural variants were found in higher numbers in these

trait-selected lines than in the control line when compared to the mouse reference genome. Moreover, we have found twice as many structural variants spanning protein-coding genes (specific to each line) in trait-selected lines. Several of these variants affect genes associated with selected phenotypic traits. These results imply that structural variation does indeed contribute to the evolution of the selected phenotypes and is heritable.

Finally, I suggest a set of critical metrics of genomic data that should be considered for a stringent structural variation analysis as comparative genomic studies strongly rely on the contiguity and completeness of genome assemblies. Because most of the available data used to represent reference genomes of mammalian species is generated using short-read sequencing technologies, we may have incomplete knowledge of genomic features. Therefore, a cautious structural variation analysis is required to minimize the effect of technical constraints.

The impact of structural variants on the adaptive evolution of mammalian genomes is slowly gaining more focus but it is still incorporated in only a small number of population studies. In my thesis, I advocate the inclusion of structural variants in studies of genomic diversity for a more comprehensive insight into genomic variation within and between species, and its effect on adaptive evolution.

Zusammenfassung

Nach dem Aussterben der Dinosaurier kam es zu einer großen adaptiven Radiation der Säugetiere, die eine erstaunliche ökologische und phänotypische Vielfalt von Säugetierarten hervorbrachte. Selbst eng verwandte Arten bewohnen oft sehr unterschiedliche Lebensräume, in denen sie verschiedenen Umwelteinflüssen und evolutionärem Druck ausgesetzt sind. Als Reaktion darauf haben Säugetiere im Laufe der Zeit verschiedene adaptive Phänotypen entwickelt, z. B. morphologische, physiologische und verhaltensbezogene. Die Genome von Säugetieren variieren in ihrem Inhalt und ihrer Struktur, und diese Variation stellt den molekularen Mechanismus für die langfristige Evolution der phänotypischen Variation dar. Das Verständnis dieser molekularen Grundlage der adaptiven phänotypischen Variation ist jedoch meist nicht trivial.

Die jüngste Entwicklung von Sequenzierungstechnologien und Bioinformatik-Tools hat einen besseren Einblick in die Genome von Säugetieren ermöglicht. Durch diese Fortschritte wurde erkannt, dass sich die Genome von Säugetieren sowohl innerhalb als auch zwischen den Arten stärker durch strukturelle Variationen als durch Unterschiede zwischen einzelnen Nukleotiden unterscheiden. Variantenarten, die in dieser Arbeit untersucht werden - wie **Deletion, Duplikation, Inversion und Insertion** - stellen eine Veränderung der Genomstruktur dar, die sich auf die Größe, die Kopienzahl, die Richtung und den Inhalt der DNA-Sequenzen auswirken. Im Gegensatz zu kurzen Varianten können strukturelle Varianten mehrere Gene umfassen. Sie können die Genkopien verändern und bemerkenswerte Unterschiede in der Genexpression und in der Folge phänotypische Unterschiede hervorrufen. Dadurch können sie dramatischere Auswirkungen auf die Fitness (den Fortpflanzungserfolg) von Individuen, die lokale Anpassung von Populationen und die Artbildung haben.

In dieser Arbeit untersuchte und bewertete ich die potenziellen funktionellen Auswirkungen von strukturellen Variationen auf die Genome von Mustelidenarten. Weil für die zuverlässige Entdeckung struktureller Varianten ein hohes Maß an Genomvollständigkeit wichtig ist, habe ich das erste Referenzgenom der Tayra (*Eira barbara*) mit Hilfe der Linked-Read-Sequenzierungstechnologie zusammengestellt, um die mit der phänotypischen Variation verbundenen Genomregionen zu ermitteln. Anschließend habe ich eine Bioinformatik-Pipeline aufgesetzt, um eine vergleichende Genomanalyse durchzuführen und die Variationen zwischen den in unterschiedlichen Umgebungen lebenden Mustelidenarten zu untersuchen. Ich fand heraus, dass zahlreiche Gene, die mit artspezifischen Phänotypen in Verbindung stehen, durch strukturelle Variationen beeinflusst werden. Diese Phänotypen stehen u.a. in Zusammenhang mit Ernährung, Körperzustand und Fortpflanzung.

Darüber hinaus untersuchte ich die Auswirkungen der künstlichen Selektion auf strukturelle Variationen bei Mäusen, die auf hohe Fruchtbarkeit, erhöhte Körpermasse und hohe

Ausdauer selektiert wurden. Durch selektive Züchtung jeder Mauslinie haben sich die gewünschten Phänotypen innerhalb dieser Populationen durchgesetzt, wobei die für jede Linie spezifischen strukturellen Variationen erhalten blieben. Im Vergleich zur Kontrolllinie hat sich die Wurfgröße in den Linien selektiert auf Fruchtbarkeit verdoppelt, die Individuen in den Linien mit hoher Körpermasse sind erheblich größer geworden, und die auf Laufbandleistung selektierten Mäuse haben wesentlich mehr Strecke zurückgelegt. Im Vergleich zum Referenzgenom der Maus wurden in diesen nach Merkmalen selektierten Linien mehr strukturelle Variationen gefunden als in der Kontrolllinie. Darüber hinaus fanden wir doppelt so viele strukturelle Variationen, die proteinkodierende Gene überspannen (spezifisch für jede Linie), in nach Merkmalen selektierten Linien. Mehrere dieser Varianten betreffen Gene, die mit ausgewählten phänotypischen Merkmalen in Verbindung stehen. Diese Ergebnisse deuten darauf hin, dass strukturelle Variationen tatsächlich zur Evolution der ausgewählten Phänotypen beiträgt und vererbbar ist.

Abschließend schlage ich eine Sammlung von maßgeblichen Metriken für Genomdaten vor, die für eine strenge Analyse der strukturellen Variation berücksichtigt werden sollten, da vergleichende Genomstudien in hohem Maße von der Kontiguität und Vollständigkeit der Genomensembles abhängen. Weil die meisten der verfügbaren Daten, die verwendet wurden, um Referenzgenome von Säugetierarten zu repräsentieren mit Short-Read-Sequenzierungstechnologien erzeugt wurden, verfügen wir möglicherweise nur über unvollständige Kenntnisse der genomischen Merkmale. Daher ist eine vorsichtige Analyse der strukturellen Variationen erforderlich, um die Auswirkungen technischer Beschränkungen zu minimieren.

Der Einfluss struktureller Variationen auf die adaptive Evolution von Säugetiergenomen rückt langsam immer mehr in den Mittelpunkt, wird aber immer noch nur in wenigen Populationsstudien berücksichtigt. In meiner Dissertation befürworte ich die Einbeziehung struktureller Variationen in Studien zur genomischen Diversität, um einen umfassenderen Einblick in die genomische Variation innerhalb und zwischen den Arten und ihre Auswirkungen auf die adaptive Evolution zu erhalten.

General Introduction

Adaptive evolution of mammalian genomes

Genomes vary in the content and structure of their DNA sequence both within the same species and between species. Genome variation is caused by mutation, gene flow, and recombination and is influenced by evolutionary processes - natural selection and genetic drift. These factors contribute to the distribution of variants along the chromosomes of individuals within and between populations. Mutations are changes in the DNA sequences that occur mostly randomly, leading to genetic diversity among individuals. In mammals, each individual inherits half of its genome from the mother and half from the father. Therefore, all processes influencing who the parents are and which parts of their DNA sequence they transmit to their offspring influence the genetic diversity of the population as well. This is the case of gene flow - an exchange of genes between populations caused by the migration of individuals of the same species and non-random mating. Finally, the recombination of DNA fragments of maternal and paternal chromosomes via sexual reproduction leads to new genetic variants.

Genetic drift and natural selection are two evolutionary processes with an important role in shaping genome diversity. Genetic drift represents the change in the frequency of an existing gene variant in a population due to random chance. It may cause some gene variants to disappear from a population completely and thereby reduce genetic variation. Moreover, drift can randomly cause initially rare gene variants to become much more frequent within a population. Natural selection, on the other hand, is a process where the increase in the frequency of a gene variant in a population is tied to a trait that is also influenced by the environment, known as a phenotype, which provides the individuals harbouring this trait variant with a reproductive advantage.

Genetic variation represents the raw material for long-term evolutionary change. It translates into diverse phenotypes through molecular networks and metabolic pathways. Mammals differ greatly in their phenotypic traits, such as morphological, physiological and behavioural ones. Phenotypic differences shape the survival and reproductive potential of individuals in the environment they inhabit. Over time, these differences can lead to an increase in the frequency of the particular features of the genomes, collectively known as a genotype, that translate to successful phenotypes – those that promote the reproductive success of individuals, and to a decrease in the frequency of those that translate to less performant phenotypes. Such a change in the frequency of genotypes of individuals in their environment occurs via natural selection. In general, the effect of phenotypic variation on fitness (reproductive success) is considered beneficial, deleterious or neutral. A phenotype may be adaptive in a given environment, but not in another, due to differing environmental

factors. Thus, understanding genetic variants facilitating species' adaptation to their environments is one of the key objectives in evolutionary biology. Yet, disentangling which variants are relevant for which adaptive phenotypic differences remains a challenge.

The extreme adaptive radiation of mammalian species occurred following the extinction of dinosaurs, an event that created an ecological opportunity for the exploitation of previously unavailable resources (Stroud and Losos 2016). Subsequently, mammals occupied different niches with a multitude of biotic and abiotic factors that require both short-term ecological and long-term evolutionary responses. To understand adaptive evolution, we need to look into the molecular basis of heritable variation in traits coupled with environmental factors affecting them is the principal focus of the study of evolution. Therefore, the aim of my research is to identify specific genomic regions responsible for the species' differences in morphological, physiological and behavioural traits and examine the adaptive evolution through which these responses arise and persist.

Numerous sequenced genomes have made it possible to apply comparative genomic methods to associate genomic variation with phenotypic differences within and between species. Variation in mammalian genomes ranges from single-nucleotide differences to large chromosomal rearrangements (Iafrate et al. 2004; Alkan, Coe, and Eichler 2011). Previously, single nucleotide polymorphisms (SNPs) were thought to account for the majority of variation among species and thus have become the most studied type of variants (Wellenreuther et al. 2019). In recent years, it has been acknowledged that mammalian genomes differ more, both on intra- and interspecies levels, as a consequence of structural variation compared to variation at a single base-pair level (Chain and Feulner 2014; Radke and Lee 2015; Chakraborty et al. 2019). For example, the average genomic variation between two humans is 0.1% in terms of single nucleotide variants, and when SVs are taken into account, this increases to 1.5% (Pang et al. 2010; Mahmoud et al. 2019). The resolution of structural variants was accelerated by the improvement of existing short-read sequencing technologies and the development of long-read technologies (Balachandran and Beck 2020), along with advances in analytical methods for improved SV identification.

The role of structural variation in shaping biological diversity

Unlike monogenic traits, such as the flower colour of the pea plant (Ellis et al. 2011), the majority of heritable traits is polygenic, involving the effect of multiple genes. These complex traits include disease susceptibility, e.g. diabetes; agriculturally important traits, such as milk yields of dairy livestock; and traits that affect the fitness of wild species, e.g. litter sizes and reproductive strategies (seasonal vs. aseasonal mating) (Goddard and Hayes 2009).

Considerable effort was put into characterizing SNPs and short insertions and deletions (indels) linked to certain traits by employing quantitative trait loci (QTL) mapping and genome-wide association studies (GWAS) (Chakraborty et al. 2019). However, for most traits, both QTL and GWAS managed to explain only a small fraction (5 - 10%) of trait heritability (Frazer et al. 2009; Eichler et al. 2010; Liu et al. 2010).

A growing number of studies support the hypothesis that chromosomal rearrangements accumulated over time contribute to phenotypic variation in complex traits and that they may be of notable relevance in both adaptation and speciation (Conrad and Hurler 2007; Hall and Quinlan 2012; Bickhart and Liu 2014; Fan et al. 2019). Structural variation (SV) intersects with genes more often than SNPs (Pang et al. 2010; Catanach et al. 2019; Chiang et al. 2017), can span multiple genes or gene blocks and likely have a larger impact on fitness (Hämälä et al. 2021). Identifying genes associated with fitness-related traits, affected by structural variation, represents an important aspect of conservation genomics. For example, uncovering deleterious structural variants affecting trait-related genes leading to low fitness could serve to better inform conservation management decisions and mitigate the impact of these variants with a minimal loss of genome-wide diversity (Wold et al. 2021).

The available data suggests that structural variants are under strong selection in both wild and domesticated species. SVs affecting dietary preferences have been detected, such as gene copy number variation (CNV) facilitating the shift to a primarily carnivorous diet in polar bears from a more omnivorous diet of brown bears during the divergence of these species (Rinker et al. 2019). A higher copy number of a gene encoding pancreatic amylase is associated with adaptation to a diet rich in starch during dog domestication (Axelsson et al. 2013; Arendt et al. 2014). Furthermore, a number of SVs have been associated with the selection of favourable production traits in livestock (Zhao et al. 2016; Liu et al. 2019).

The impact of SVs on fitness may be direct - leading to the disruption of coding regions or regulatory elements; or indirect - causing suppression of recombination (Mérot et al. 2020). Suppression of recombination by SVs, particularly inversions, might play an important role in local adaptation and subsequently speciation, as inversions can shelter beneficial alleles

of multiple genes from gene flow by suppressing the formation of crossovers within chromosomal heterozygotes (Giner-Delgado et al. 2019).

History of structural variant identification

The discovery of structural variants dates back to the 1920s before scientists had knowledge about DNA and its function. Structural variants were first discovered when Alfred Sturtevant compared chromosome maps of closely related fruit flies, *D. melanogaster* and *D. simulans*. Sturtevant noticed that the structure of chromosomes was similar in both species, except for a large section of the third chromosome, where the segment was inverted (Sturtevant 1921). Following this discovery, Sturtevant performed experiments using *D. melanogaster* mutants with inversions present in different chromosomes and concluded that in heterozygotes, inversions are suppressing recombination in these genomic regions (Sturtevant and Mather 1938; Dobzhansky and Sturtevant 1938).

In the same decade, Barbara McClintock identified transposable elements (TEs) in maize, genomic rearrangements commonly referred to as “jumping genes” (McClintock 1931; McClintock 1950). TEs can move from one genomic location and insert themselves into another during replication. If inserted in a coding sequence (CDS), they can lead to its disruption and subsequently affect its gene products. Nevertheless, TEs don’t necessarily have deleterious effects and may cause “exon shuffling”, alteration in gene sequence that could give rise to novel proteins (Moran, DeBerardinis, and Kazazian 1999).

In recent decades, with the development of molecular genetics and novel markers, including microsatellites, and the more widespread single nucleotide polymorphism (SNP) markers, focus shifted from large chromosomal rearrangements toward shorter variants (Elshire et al. 2011). As SNP genotyping emerged as a scalable and affordable high-throughput method, generating results comparable across different laboratories, the growing number of studies centred around single base-pair variation (Kim and Misra 2007).

Still, associating genotypic to phenotypic variation represents one of the most challenging tasks in genomics. The application of SNP genotyping assays within genome-wide association studies (GWAS), suggests thousands of associations between SNPs and complex traits in humans (Wood et al. 2014). Despite multiple studies confirming notable phenotypic effects of SNPs, this type of variation is unlikely solely responsible for a broad spectrum of phenotypic diversity within and among species (Hoekstra et al. 2006; Shastry 2009; Young et al. 2019) and is only explaining a small amount of heritability of a trait (Eichler et al. 2010). Fifty years ago, Ohno (1970) proposed that gene expression can be significantly altered by copy number variation (CNV), a subtype of SVs, including deletions and duplications of whole gene sequences, subsequently affecting phenotype and evolutionary trajectories of species (Ha, Kim, and Chen 2009).

Types of structural variants

Structural variation represents a change in the structure of genomic regions, impacting the size, copy number, location, orientation and content of a DNA sequence (Fan et al. 2014; Bickhart and Liu 2014). Originally, the term *structural variant* was used to define a genomic region that differs between individuals, either in copy number (deletion, duplication, insertion), orientation (inversion), or chromosomal location (translocation), concerning regions greater than 1000 bp in length (Alkan, Coe, and Eichler 2011). Following the improvement of sequencing resolution, the definition was expanded to variants greater than 50 bp (Kosugi et al. 2019), while smaller elements are referred to as indels (insertions or deletions). Another way of characterizing SVs is based on whether they are unbalanced, i.e. showing loss or gain of genomic information as do CNVs, or balanced, i.e. with no change in net genomic content - as in inversions and translocations (Escaramís, Docampo, and Rabionet 2015). In my dissertation, I focused on four types of chromosomal rearrangements - **uplications, deletions, inversions and insertions** (Figure 1.), involving a change in copy number and orientation of the genomic segments.

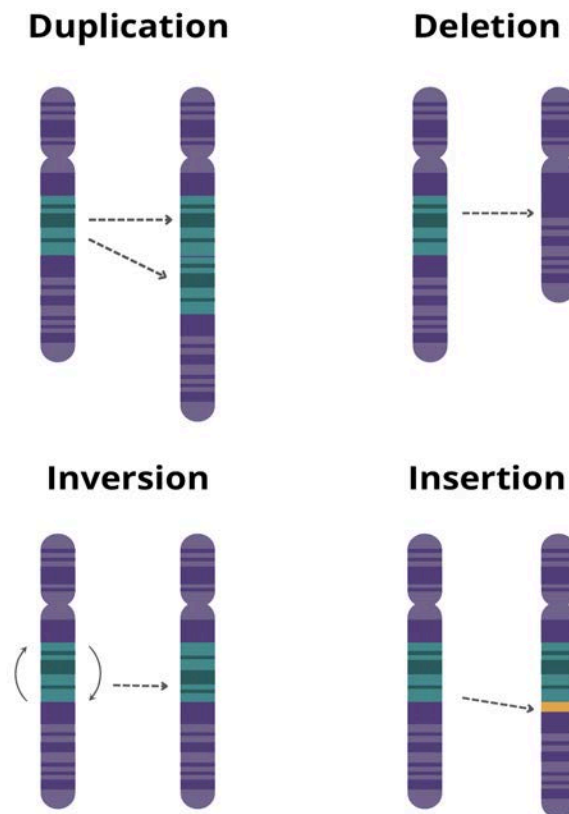


Figure 1. Types of SVs investigated in this study: duplication, deletion, inversion and insertion.

Copy number variation (CNV)

This type of structural variation accounts for the largest fraction of detected SV and involves unbalanced genomic rearrangements such as duplication (DUP), deletion (DEL) or insertion (INS), which lead to a decrease or increase in genomic content (Redon et al. 2006). CNVs encompass more polymorphic nucleotides than SNPs by an order of magnitude (Conrad and Hurler 2007, Conrad et al. 2010).

Although the study of gene duplications (Ohno, 1970) represents a turning point in the research of genome evolution, the first segmental gene duplication was identified in the 1920s by Sturtevant (Sturtevant 1925). It was an X chromosome-linked *Bar* mutation in *D. melanogaster*. This duplication causes hemizygous male and homozygous female flies to develop smaller, elongated eyes compared with round eyes in the wild-type (Wolfner and Miller 2016). Moreover, a tandem triplication of the *Bar* region, resulting from an unequal crossover, led to an even more severely affected phenotype, with a further reduction of the eyes (Sturtevant 1925). Until recently, gene duplication and its effect on gene dosage and species diversification represented one of the best-documented types of copy number variants (Dhar, Bergmiller, and Wagner 2014; Mérot et al. 2020). However, the adaptive potential of deletions and insertions, along with their deleterious effect, is gaining more focus in the last decades, following advancements in sequencing technologies and more rigorous comparative genomics analyses.

Transposable elements (TEs), also known as transposons, represent one type of insertion abundantly present in eukaryotic genomes (Bourque et al. 2018). They can be balanced or unbalanced, appear in different sizes and insert randomly across the genome, often leading to the disruption of genes (McClintock 1950), and comprise approximately one-fourth of the SVs differentially present in human genomes (Gardner et al. 2017). There is a growing interest in the adaptive potential of transposable elements and their influence on genome evolution and genetic differences in individuals (Bourque et al. 2018). These insertions can trigger a broad range of molecular variations in a population with potentially severe fitness and phenotypic consequences for individuals (Schrader and Schmitz 2019). One of the classic examples is industrial melanism, the adaptive response of peppered moths to environmental changes during the Industrial Revolution in Britain. Influenced by coal pollution and bird predation, the replacement of the common pale form with a black form occurred. The mutation underlying industrial melanism was the insertion of a large, tandemly repeated, transposable element into the first intron of the *cortex* - a previously unknown gene (Van't Hof et al. 2016).

Origin and formation of CNVs

CNV formation occurs during recombination and replication events, with higher *de novo* locus-specific mutation rates compared to SNPs (Turner et al. 2008). There are four mechanisms associated with the formation of CNVs: nonallelic homologous recombination (NAHR), and nonhomologous end-joining (NHEJ), both occurring during the recombination stage, retrotransposition (Kazazian and Moran 1998; Kidd et al. 2008; Xing et al. 2009), and fork stalling and template switching (FoSTeS) (Zhang et al. 2009). The latter is a replication-based mechanism suggested to explain more complex genomic rearrangements. Sequencing of variant breakpoint regions supported the finding that a fraction of complex CNVs occur by a mechanism consistent with FoSTeS (Perry et al. 2008).

1. Nonallelic homologous recombination (NAHR)

NAHR occurs during mitosis and meiosis, leading to duplication, deletion or inversion. It is caused by sequence alignment and crossover between two paralogous, nonallelic sequences showing high similarity. In cases where these nonallelic sequences are located on the same chromosome and in direct orientation, duplication and/or deletion can occur, while inversions emerge if the genomic region is flanked by inverted repeats (Stankiewicz and Lupski 2002). NAHR taking place between repeats on different chromosomes can lead to chromosomal translocation (Samonte and Eichler 2002). Repeats that facilitate NAHR are primarily low copy repeats or segmental duplications (SDs) of >10kb in length and 95 - 97% sequence similarity (Shaw and Lupski 2004). Different types of SDs are occasionally grouped together, with some tandem subunits, or ones with reverse orientation, forming a more complex SD.

In addition to SDs, retrotransposons such as L1 elements (Han et al. 2008), *Alu* (Babcock et al. 2003), or homologous pseudogenes (Steinmann et al. 2007; Kim et al. 2008) can also trigger the NAHR event. If NAHR occurs in meiosis, it results in unequal recombination and leads to genomic rearrangements that can be neutral polymorphisms or could give rise to *de novo*, or inherited genomic disorders (Turner et al. 2008). When NAHR takes place during mitosis, it leads to somatic mosaicism, characterized by populations of somatic cells with SVs (Quinlan and Hall 2012).

2. Nonhomologous end-joining (NHEJ)

This mechanism evolved as a primary way to repair DNA double-strand breaks in eukaryotic cells such as those caused by ionizing radiation. There are four stages of the NHEJ process: detection of double-strand break, molecular bridging of both broken DNA ends, modification of the ends, and the final ligation (Weterings and van Gent 2004). NHEJ

does not require breakpoints with high similarity and often leads to short indel events at these SV breakpoints (Lee, Carvalho, and Lupski 2007). Genomic regions where NHEJ takes place often overlap repetitive elements such as LTR, LINE, and Alu. NHEJ is also considered to be the mechanism important for rejoining translocated chromosomes in cancers (Gu, Zhang, and Lupski 2008).

3. Fork stalling and template switching (FoSTeS)

Most of the complex SVs are formed following this mechanism during DNA replication. It happens when the DNA replication fork stalls at one position, the lagging strand gets detached from the DNA template, translocates to the 3' end at the homologous region, and restarts the replication process with another fork (Lee, Carvalho, and Lupski 2007). Upon transferring, the joining point rather than a breakpoint is created, as one DNA segment is joined with another. Switching to a downstream fork (forward invasion) would result in a deletion, whereas switching to an upstream fork (backward invasion) yields a duplication.

4. Retrotransposition

Retrotransposition represents the insertion of a DNA sequence mediated by an RNA intermediate (Boeke et al. 1985). In this process, an RNA copy of the original retrotransposon is generated and then reverse-transcribed back into the genome by reverse transcriptase. Several studies showed a correlation between retrotransposons such as *Alu* and the breakpoints of segmental duplications in human and primate genomes. They suggest a role in the expansion of gene-rich segmental duplications and subsequently origin of other types of SVs. (Bailey, Liu, and Eichler 2003; Goidts et al. 2006; Lee et al. 2008; Cao et al. 2020).

Inversions and translocations

Inversions (INV) are generally characterized as copy number neutral rearrangements that affect the order and orientation, but not the gain or loss of the genomic segment (Sharp, Cheng, and Eichler 2006). Inversions are represented as segments of chromosomes in the opposite orientation in comparison to the reference genome. This renders them hard to detect with common CNV detection methods, especially ones relying on the read depth analysis (Chaisson et al. 2019). Previously, inversions have mostly been identified through karyotyping in cytogenetic studies. More recently, inversions are identified at a higher resolution with sequence-based methods (Sanders et al. 2016). Still, polymorphic inversions are challenging to detect, because they are often flanked by segmental duplications that can

span > 1 Mbp in length and cannot be fully overlapped by short-read sequencing methods (Alkan et al. 2009; Kidd et al. 2010). Within the human genome, it was found that the majority of the polymorphic inversion breakpoints indeed mapped to regions of segmental duplication (Tuzun et al. 2005). Inversions can spread in a population by reducing recombination between alleles that independently increase fitness (Lyon 2003; Stefansson et al. 2005; Hoffmann and Rieseberg 2008).

Translocations (TRA) occur when there is an exchange of genomic content between two chromosomes (interchromosomal) or distal regions within the same chromosome (intrachromosomal). Most of the translocations result in a copy number neutral rearrangement (balanced), although some can alter the genomic content (unbalanced) (Balachandran and Beck 2020) such as translocation leading to Down's syndrome (Bornstein et al. 2010). In this study, translocations were excluded from the analysis due to the difficulty in detecting them reliably, especially in the mustelid species with different karyotypes.

Adaptive variation in protein-coding regions

Gene duplications and losses

Gene duplications exist in large numbers in mammalian genomes and contribute to many differences in phenotypes between species. They act as a source of genetic material from which new functions may arise over time (Han et al. 2009; Lynch and Force 2000).

Duplicated genes or paralogs usually undergo two different processes: long-term maintenance in the genome or loss (Ohno 1970). For many years, loss of function or pseudogenization was thought to be the most common outcome for one of the gene copies that gets silenced by degenerative mutation. However, a high degree of retention of functional gene duplicates (20 - 50 %) across eukaryotic genomes shows otherwise (Han et al. 2009). Thus, at least a subset of duplicated genes goes through subfunctionalization, with paralogs dividing their ancestral function into its parts or subfunctions among themselves; or neofunctionalization, where one paralog retains its ancestral function, and the other takes on a completely new function. An example of neofunctionalization has been detected in placental mammals, where the non-shivering thermogenesis arose via neofunctionalization of the UCP1 paralog (Mendes et al. 2020). Subfunctionalization of duplicates is a product of neutral evolution where the duplicates become fixed for complementary mutations where no new functions are formed (Lynch and Force 2000).

Another mechanism that contributes to adaptive evolution and phenotypic differences is gene loss, a partial or complete absence of a functional gene sequence encoding a protein.

Two of the mechanisms by which a gene loss occurs suddenly is through unequal crossing-over resulting in the physical removal of a gene, or through the insertion of a transposable element into the exon leading to disruption of gene function. In contrast, a more incremental process is gene loss via pseudogenization, where point mutations accumulate after an initial mutation that causes a loss of function (Albalat and Cañestro 2016). This initial mutation can be a nonsense mutation leading to the production of truncated proteins, short insertions or deletions causing a frameshift, or a missense mutation that by affecting splice sites, results in atypical, often nonfunctional transcripts.

Associations between gene losses and changes in mammalian phenotypes have been implied by several studies (Danchin, Gouret, and Pontarotti 2006; Guijarro-Clarke, Holland, and Paps 2020; Sharma and Hiller 2020; Yuan et al. 2021). Some of the direct roles gene losses have played in the evolution of new adaptations have been shown in aquatic mammals (Zhou et al. 2018; Yuan et al. 2021). For example, adaptations to the aquatic environment have been facilitated by a decrease in the copy number of genes that encode olfactory receptors (Hughes et al. 2018), and epidermis- and hair-related functions (Sharma et al. 2018; Espregueira Themudo et al. 2020).

Artificial selection and structural variation

Artificial selection is a process in which animals or plants with particular phenotypic traits are chosen for further breeding with the aim of enhancing and propagating these traits in future generations (Conner 2003). Systems where artificial selection can be induced, provide a valuable method to directly measure adaptive genetic responses in populations (Kessner and Novembre 2015). The aim of artificial selection experiments is to assess the genetic basis for complex traits by analyzing changes in allele frequency in selected populations. This approach has been used in a wide variety of organisms, from yeast (Ehrenreich et al. 2010) to mice (Copes et al. 2015) and livestock (Gomez-Raya et al. 2002).

The majority of the studies in mammals have focused on the differences expressed in morphological characteristics such as coat colouration, muscle growth, increased litter size and others. Moreover, the effect of artificial selection on copy number variation has been investigated in domesticated species (Lye and Purugganan 2019). Protein-coding genes related to metabolic activity and production traits have been shown to be affected by SVs during the selection of certain domesticated species. In Holstein cattle with high and low estimated breeding values of milk protein and fat percentage, copy-number-variable regions have been detected spanning genes associated with milk and fat composition (Gao et al. 2017). Similarly, CNV has been suggested to be involved with high fertility in goats. Duplication of genes encoding prolactin, a hormone that regulates lactation, ovarian

function and fetal growth and development, is observed in the high fertility group (Zhang et al. 2019).

Furthermore, in several artificially selected species, mainly livestock and pets, heterozygotes were found at high frequencies in populations, with the same variant having detrimental effects when present in a mutant homozygous state (Hedrick 2015). These heterozygotes are artificially selected for, as they exhibit higher relative fitness compared to both homozygotes (Hedrick 2012). For example, the taillessness in Manx cats shows marginal heterozygote advantage over the wild-type homozygote, but a greater heterozygote advantage over the mutant homozygote which is lethal (Adalsteinsson 1980). In sheep, two genes that affect female fecundity lead to higher fecundity when heterozygous, relative to the wild-type homozygote, but if homozygous mutant alleles are inherited, they lead to infertility (Gemmell and Slate 2006).

During the diversification and/or enhancement of selected traits, domesticated species develop local or population-specific adaptations to different environments or human preferences (Larson et al. 2014; Wang et al. 2014). Some of the adaptive traits that arose during this process may have evolved under ‘unconscious selection’, such as adaptation to a human diet rich in starch during dog domestication enabled by an increase in copy number of the gene encoding for pancreatic amylase (Axelsson et al. 2013; Arendt et al. 2014). Moreover, it has been demonstrated that CNVs have been implicated in the adaptation of Chinese indigenous cattle to high-altitude environments, with copy number variable genes related to hypoxia also showing strong selection signatures (Zhang et al. 2019). These findings suggest that both artificial and natural selection may have shaped the landscape of CNVs in genomes of domesticated animals, thereby contributing to adaptive evolution and breed differentiation.

Methods for structural variant characterization

Identification of structural variants has notably improved since the first findings of large rearrangements spanning more than several megabase pairs, discovered using light microscopy. Nowadays, *in silico* variant calling using NGS data enables the detection of most types and sizes of SVs. Despite further advancements in sequencing technologies and a growing number of open-source tools for SV characterization, reliable detection of the whole range of complex types and sizes of SVs is difficult to achieve relying on a single method at a reasonable cost. One of the challenges is the variable amount of repeats present in genomes. Thus, studies of primarily model or commercially important species have implemented a combination of sequencing and bioinformatic methods to reliably identify and validate SVs (Balachandran and Beck 2020).

Traditional identification techniques

One of the first SV identification methods used is karyotyping, a staining method used to identify and pair homologous chromosomes, and to determine the number and size of chromosomes. It forms a banding pattern along chromosomes, where the intensity of staining depends on the structure of the DNA region (Trask 2002). The heterochromatic, AT-rich regions are dyed more intensely than euchromatic, GC-rich genomic regions. Such stained chromosomes, observed in cells in the metaphase stage, are then compared and inspected for insertions, deletions, translocations or aneuploidies. Karyotyping is suited for the identification of large (> 3 Mbp) chromosomal rearrangements, due to its low resolution.

The fluorescence *in situ* hybridization (FISH) technique has a higher resolution than karyotyping and is capable of detecting chromosomal alterations > 100 kbp and longer with a low false discovery rate (FDR) (Cui, Shu, and Li 2016). It utilizes fluorescent probes that hybridize to complementary DNA segments on chromosomes. Following successful hybridization, metaphase cells are inspected for fluorescent signals (Hu et al. 2014). In recent years, a Cas9-mediated FISH method was developed for marking highly repetitive genomic regions (Deng et al. 2015).

Comparative genomic hybridization (CGH) is a hybridization method that anneals genomic DNA from two samples (test and reference, or tumour and control) to either long oligonucleotides or bacterial artificial chromosome (BAC) clones, to identify changes in copy number between them. The most commonly used form, array CGH (aCGH), uses an array of oligonucleotides for high throughput and high-resolution hybridization (Conrad et al. 2010). Although it cannot identify balanced SVs, aCGH can be scaled up to detect multiple CNVs at once, even at the resolution of 1 kb for the whole human genome (Wiszniewska et al. 2014).

The single-nucleotide polymorphism (SNP) arrays work in a similar way as previously noted aCGH but contain allele-specific oligos and support allele frequency measurement (Cooper et al. 2008). These arrays are predominantly used for SNP genotyping, but can also be applied for CNV detection (Wang et al. 2007). These SNP genotyping platforms cannot identify balanced SVs such as inversions and balanced translocations.

Second and third-generation sequencing technologies

Although hybridization methods are effective in CNV detection, they fail to quantify higher copy number gains and cannot detect SV breakpoints with base-pair accuracy. Development of second- or more widely referred to as next-generation sequencing (NGS) technologies enabled large-scale assessment of SVs, using whole-genome sequencing (WGS) and high-throughput sequencing data, primarily generated with Illumina sequencing platforms. This SV detection method is based either on mapping paired-end (PE) reads to the reference genome and identifying SVs based on the evidence supporting the called variant, or a direct comparison of the *de novo* genome assembly, aligned to the reference genome of the same or closely related species, to infer synteny between them (Balachandran and Beck 2020).

Short-read sequencing includes library preparation, where DNA is sheared into shorter fragments followed by insert size selection, and sequencing of both ends of the DNA fragment (Korbel et al. 2007). Reads are then either mapped to the reference genome or *de novo* assembled. To infer SVs from mapped reads, change in insert size and read orientation, along with the depth of coverage are assessed. Currently, the common approach for short-read SV detection is an analysis of mapped Illumina 150 bp PE reads. The mapping methods usually rely on four types of evidence for SV detection: read pairs, split reads, read depth, and contig assembly (Kosugi et al. 2019).

Still, short-read sequencing technologies have a common limitation - the inability to sequence long stretches of DNA, particularly repetitive ones, and to resolve complex SVs and phase haplotypes. To sequence a large stretch of DNA using NGS, the strands have to be fragmented and amplified. Unfortunately, these steps can introduce sequence gaps and biases into the library preparation. Also, short-read sequencing can fail to generate a sufficient overlap between the DNA fragments and sufficient coverage. Thus, sequencing highly complex and repetitive genomes can be challenging using these technologies, along with mapping issues, due to the non-unique read mapping in repetitive regions.

To amend some of these issues, while still employing widely used short-read sequencing, synthetic long-read methods have been developed, with the 10x Genomics ‘Linked-Reads’ approach being the most prominent and cost-effective. Linked reads provide long-range information from short-read sequencing data, using molecular barcodes to tag short reads that originate from the same long DNA fragment, making it possible to assemble larger neighbouring fragments back together and access previously inaccessible repetitive segments to a certain extent. Linked-reads rely on emulsion technology to partition long DNA molecules into GEMs (gel beads in emulsion). Each GEM contains fragments of a single long DNA molecule, with each fragment getting a unique barcode following the amplification process. After Illumina sequencing, barcodes are used for the identification of adjacent DNA fragments and reconstruction of the initial long DNA strand, while phasing

out haplotypes, where the maternal and paternal origin of genomic loci is inferred. This made the linked-read approach also favourable for structural variant detection (Marks et al. 2019; Karaođlanođlu et al. 2020).

In contrast to short-read sequencing, recently developed third-generation sequencing, long-read technologies have the capacity to sequence on average over 10 kb in one single read, thereby requiring fewer reads to cover the same genomic segment. There are two predominant long-read sequencing technologies, Pacific Biosciences' Single-Molecule Real-Time (SMRT) sequencing (Rhoads and Au 2015), and Oxford Nanopore Technologies' platform (Jain et al. 2016). Long reads originating from a single molecule eliminate amplification bias and generate longer DNA fragments to overlap repetitive regions for an improved genome assembly (Amarasinghe et al. 2020). Initially, a downside to long-read sequencing was the accuracy per read being lower than that of short-read sequencing. Recently, accuracy in both SMRT and nanopore sequencing has been notably improved, leading to the generation of highly accurate reads, with the raw base-called error rate claimed to have been reduced to < 1% for SMRT sequencers (Wenger et al. 2019) and < 5% for nanopore sequences (Jain et al. 2018).

The detection method used in this thesis

In my studies, I used an SV detection approach based on short-read sequencing data. Prior to the SV calling and annotation, this approach comprises several exhaustive steps of alignment preparation to retain uniformity among samples: adapter and quality trimming, read mapping to the reference genome, sorting and deduplication. Subsequently, filtered alignments underwent structural variant detection, as shown in Figure 2. To reduce the bias toward a specific SV type or size, I employed three SV calling tools - Manta (Chen et al. 2016), Whamg (Kronenberg et al. 2015) and Lumpy (Layer et al. 2014), in an ensemble approach, with each tool using different detection algorithm. Such a combination of algorithms, relying on information from paired and split reads, depth of coverage and contig assembly, sums up to a higher specificity and sensitivity rate, compared to applying only a single tool.

Subsequently, I genotyped and quality filtered all variant call sets, and further analyzed only scaffolds assigned to chromosomes. To reduce the number of potential false-positive events, I identified and removed SV calls overlapping gaps and high coverage regions in the reference genome. I used Survivor (Jeffares et al. 2017), an open-source tool to merge and compare SV call sets within and among samples, and retained SV calls detected by at least two detection tools per sample. Furthermore, I annotated a union of SV calls among all samples containing sample-specific and shared variants with Ensembl Variant Effect Predictor (McLaren et al. 2016) to identify variants affecting protein-coding genes.

Finally, I conducted a functional classification of genes through detailed literature and database search (OrthoDB v10, Kriventseva et al. 2019; Uniprot, The UniProt Consortium 2017; NCBI Entrez gene, Maglott et al. 2011), and Gene Ontology enrichment analysis of biological processes (Shiny GO, Ge, Jung, and Yao 2020) for SVs encompassing larger gene blocks (> 5 genes). As the precise effect of inversions overlapping large sets of genes is still challenging to determine, I inspected inversions affecting only up to 20 genes for significantly enriched biological processes, with no limitation imposed on deletions, duplications and insertions.

Structural variant detection pipeline

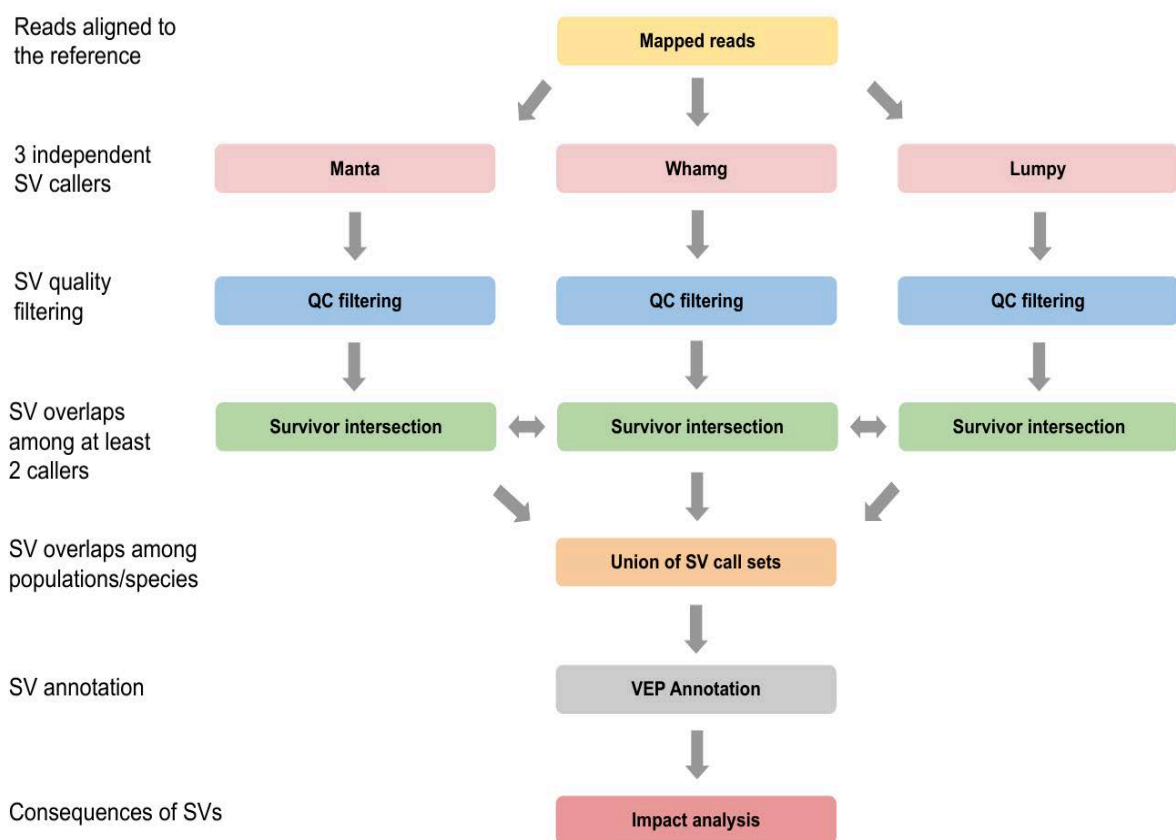


Figure 2. The schematic overview of seven steps of the structural variant detection pipeline with three SV calling algorithms (Manta, Whamg, Lumpy), a tool to sort and intersect SVs found in all species or individuals (Survivor), and Variant Effect Predictor (VEP), an Ensembl annotation tool.

Reference genome improvements and conservation genomics

Assessment of all types of variation among genomes of closely related species is affected by contiguity and completeness of genome assemblies (Gurevich et al. 2013). The majority of reference genomes of non-model species are represented with draft genome assemblies (Meltz Steinberg et al. 2017; Rhie et al. 2021). Incomplete genomic features such as genes, regulatory elements, repeats, presence of gaps and contig misassemblies, complicate the interpretation of results of comparative genomic analysis and could lead to over- and underestimations of certain attributes.

The recent advancements in high-throughput sequencing, development of linked-read (Weisenfeld et al. 2017) and long-read technologies (Murigneux et al. 2020), and chromosome conformation capture (Hi-C) methods (Lieberman-Aiden et al. 2009) enabled the generation of more contiguous assemblies built from scaffolds spanning the length of entire chromosomes.

The linked-read strategy, developed by 10x Genomics, generates synthetic long reads by incorporating information from high-molecular-weight DNA (>50 kb in length) through barcoding of its fragments sequenced with short-read technologies (Weisenfeld et al. 2017). The information retained in linked reads enables long stretches of the genome to be resolved more accurately with fewer gaps and haplotype phased, making the whole genome assembly process more straightforward and notably faster than the assembly from only the short-read sequencing (Ott et al. 2018; Armstrong et al. 2019). These characteristics and the lower input and cost compared to the true long-read technologies made the linked-read method popular for sequencing mammalian genomes (Armstrong et al. 2019; Etherington et al. 2019).

Previously mentioned long-read sequencing technologies, the Single Molecule, Real-Time (SMRT) sequencing, developed by Pacific Biosciences, and nanopore sequencing (Oxford Nanopore Technologies), generate read lengths in the range of 5 kbp to >100 kbp. Longer reads provide better resolution of repetitive elements and substantially reduce the computational complexity of genome assembly (Sohn and Nam 2018; Giani et al. 2020).

Furthermore, as chromosome-length assemblies can be generated with Hi-C scaffolding of existing draft assemblies, assembly improvements come at a smaller additional cost. Thus, the generation of more contiguous assemblies leads to improved analyses of genome-wide features. Such advancements have enabled the transition from conservation genetics to conservation genomics, with more focus on functional variation across whole genomes and their interaction with the environment.

Initiatives to generate high-quality reference genomes of non-model species have been on the rise in recent years. A community-driven effort has already yielded many chromosome-length genomes, crucial for more accurate comparative genomic analysis,

population studies and subsequently, conservation management decisions. To address these issues, the Earth BioGenome Project has an ambitious plan to sequence genomes of all eukaryotic species on our planet in the next ten years, and the Vertebrate Genomes Project (VGP) aims to generate genomes for ~70,000 species representing major orders and families of vertebrates (Rhie et al. 2021). In addition, the DNA Zoo is working on improving the contiguity of published draft genomes with Hi-C mapping, besides generating new ones, and has made available already more than 200 assemblies to be freely used for scientific research (Dudchenko et al. 2017). Similar initiatives have been established in Europe, such as the Darwin Tree of Life and European Reference Genome Atlas (ERGA), with the focus on sequencing all eukaryotic species in Great Britain and Ireland, and genomes of species at risk of extinction across Europe, respectively.

One of the more specialized projects, the *Martes* Genome Consortium, and the most relevant for my dissertation, aims to generate and analyze at least one reference genome of each species within the subfamily *Guloninae*, with four *Martes* genomes already assembled in collaboration with the DNA Zoo, and tayra genome generated as part of this thesis.

For these large and noteworthy initiatives to be executed successfully, it is important to collect and preserve tissue samples and genetic material in an optimal manner, in order to ensure high-quality genomes can be assembled. Furthermore, long-term storage of genetic material obtained from captive or free-ranging individuals will likely have a notable impact on future conservation decisions and the management of endangered populations. The EAZA Biobank serves such a purpose as the primary repository of samples supporting conservation research, and the Leibniz Institute for Zoo and Wildlife Research (IZW) represents one of the four European hubs.

Aims of study

Despite the growing interest in research of structural variation in model species, this type of variation in nonmodel species has been understudied. The goal of this thesis was to identify and evaluate the potential effect of structural variants on the genomes in wild mustelid species, along with the study of SV selection and inheritance in artificially selected mice lines. Here, I suggest that structural variants should be an integral part of genomic variation studies as they affect larger segments of the genome, have a stronger effect on the phenotype, and account for the majority of variation leading to trait differences between and within species compared to shorter variants.

The topics of my thesis chapters revolve around three general aims.

My first aim was to detect structural variants associated with trait-related genes and adaptive phenotypic differences among mustelids. In the last decade, it has been acknowledged that mammalian genomes differ more, both on the intra- and interspecies levels, as a consequence of structural variation compared to single-nucleotide variation.

- 1. Can genomic regions harbouring structural variants be associated with trait-related genes?**

Identifying variable regions of the genome associated with differences in species' morphological and physiological traits is an important aim of evolutionary studies and my thesis. Still, examining the adaptive evolution through which these traits arise and persist in a population is usually not straightforward in wild species. Thus, I inspected the dynamics of structural variants in mice lines artificially selected for several phenotypic traits.

- 2. Does the artificial selection of phenotypic traits result in an accumulation of SVs in trait-related genes?**

As the majority of reference genomes of mammalian species are represented with draft genome assemblies, this may lead to incomplete knowledge of genomic features. Due to this issue, more contiguous and complete genome sequences, and a cautious and stringent variation analysis is needed to avoid over- and underestimating certain attributes. Furthermore, I suggest the critical metrics of genomic data to consider prior to structural variation identification.

- 3. How do assembly quality and discovery methods impact variation detection? Which metrics are important for SV detectability?**

To achieve these aims, in chapter I, I generated a *de novo* genome assembly of the tayra (*Eira barbara*) and conducted a comparative genomic analysis with published genomes of three closely related mustelids: wolverine (*Gulo gulo*), sable (*Martes zibellina*) and domestic ferret (*Mustela putorius furo*). I examined three different types of variation displaying adaptive potential, from signals of positive selection in single-copy genes, expansions in gene families, to structural variation.

Chapter II encompasses structural variation analysis as part of the conservation genomics study of genomic consequences of the population bottleneck in the endangered black-footed ferret (*Mustela nigripes*) in comparison with closely related mustelid species. This is one of the most endangered mammals in North America, which exhibits low genomic diversity and long stretches of runs of homozygosity due to inbreeding depression.

In chapter III, I performed structural variation analysis as part of the study of candidate genes associated with complex traits (high fertility, high fat and high protein mass, and endurance) selected for in five mice lines over more than 100 generations.

Chapter IV focuses on the benefit of chromosome-length genome assemblies in conservation genomic studies, enabling a comprehensive and accurate assessment of genetic diversity in endangered species.

Chapter I

Multiple types of genomic variation contribute to adaptive traits in the mustelid subfamily *Guloninae*

Lorena Derežanin^{1*}, Asta Blažytė², Pavel Dobrynin³, David A. Duchêne⁴, José Horacio Grau⁵, Sungwon Jeon^{2,#}, Sergei Kliver⁶, Klaus-Peter Koepfli^{3,7,8}, Dorina Meneghini¹, Michaela Preick⁹, Andrey Tomarovsky^{3,6,10}, Azamat Totikov^{3,6,10}, Jörns Fickel^{1,9}, Daniel W. Förster¹

1 Leibniz Institute for Zoo and Wildlife Research (IZW), Alfred Kowalke Straße 17, 10315 Berlin, Germany

2 Department of Biomedical Engineering, College of Information and Biotechnology, Ulsan National Institute of Science and Technology (UNIST), Ulsan, 44919, Republic of Korea

3 Computer Technologies Laboratory, ITMO University, 49 Kronverkskiy Pr., 197101 Saint Petersburg, Russia

4 Center for Evolutionary Hologenomics, The GLOBE Institute, Faculty of Health and Medical Sciences, University of Copenhagen, Øster Farimagsgade 5, 1353 Copenhagen, Denmark

5 amedes Genetics, amedes Medizinische Dienstleistungen GmbH, Jägerstr. 61, 10117 Berlin, Germany

6 Institute of Molecular and Cellular Biology, SB RAS, 8/2 Acad. Lavrentiev Ave., Novosibirsk 630090, Russia

7 Smithsonian-Mason School of Conservation, 1500 Remount Road, Front Royal, VA 22630, USA

8 Smithsonian Conservation Biology Institute, Center for Species Survival, National Zoological Park, 1500 Remount Road, Front Royal, VA 22630, USA

9 Institute for Biochemistry and Biology, Faculty of Mathematics and Natural Sciences, University of Potsdam, Karl-Liebknecht-Str. 24-25, 14476 Potsdam OT Golm, Germany



10 Novosibirsk State University, 1 Pirogova str., Novosibirsk, 630090, Russia

current address: Clinomics Inc., Ulsan, 44919, Republic of Korea

**Corresponding author*

Published in *Molecular Ecology*, 2022, DOI: 10.1111/mec.16443

Multiple types of genomic variation contribute to adaptive traits in the mustelid subfamily Guloninae

Lorena Derežanin¹  | Asta Blažytė² | Pavel Dobrynin³  | David A. Duchêne⁴ | José Horacio Grau⁵ | Sungwon Jeon² | Sergei Kliver⁶ | Klaus-Peter Koepfli^{3,7,8} | Dorina Meneghini¹ | Michaela Preick⁹ | Andrey Tomarovsky^{3,6,10} | Azamat Totikov^{3,6,10} | Jörns Fickel^{1,9} | Daniel W. Förster¹

¹Leibniz Institute for Zoo and Wildlife Research (IZW), Berlin, Germany

²Department of Biomedical Engineering, College of Information and Biotechnology, Ulsan National Institute of Science and Technology (UNIST), Ulsan, Korea

³Computer Technologies Laboratory, ITMO University, Saint Petersburg, Russia

⁴Center for Evolutionary Hologenomics, The GLOBE Institute, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

⁵Amedes Genetics, amedes Medizinische Dienstleistungen GmbH, Berlin, Germany

⁶Institute of Molecular and Cellular Biology, SB RAS, Novosibirsk, Russia

⁷Smithsonian-Mason School of Conservation, Front Royal, Virginia, USA

⁸Center for Species Survival, Smithsonian Conservation Biology Institute, National Zoological Park, Front Royal, Virginia, USA

⁹Institute for Biochemistry and Biology, Faculty of Mathematics and Natural Sciences, University of Potsdam, Potsdam OT Golm, Germany

¹⁰Novosibirsk State University, Novosibirsk, Russia

Correspondence

Lorena Derežanin, Leibniz Institute for Zoo and Wildlife Research (IZW), Alfred Kowalke Straße 17, 10315 Berlin, Germany.
Email: lorenaderezanin@gmail.com

Present address

Sungwon Jeon, Clinomics Inc., Ulsan, Korea

Funding information

Russian Foundation for Basic Research, Grant/Award Number: 0-04-00808; Carlsbergfondet, Grant/Award Number: CF18-0223

Handling Editor: Andrew DeWoody

Abstract

Species of the mustelid subfamily Guloninae inhabit diverse habitats on multiple continents, and occupy a variety of ecological niches. They differ in feeding ecologies, reproductive strategies and morphological adaptations. To identify candidate loci associated with adaptations to their respective environments, we generated a *de novo* assembly of the tayra (*Eira barbara*), the earliest diverging species in the subfamily, and compared this with the genomes available for the wolverine (*Gulo gulo*) and the sable (*Martes zibellina*). Our comparative genomic analyses included searching for signs of positive selection, examining changes in gene family sizes and searching for species-specific structural variants. Among candidate loci associated with phenotypic traits, we observed many related to diet, body condition and reproduction. For example, for the tayra, which has an atypical gulonine reproductive strategy of aseasonal breeding, we observed species-specific changes in many pregnancy-related genes. For the wolverine, a circumpolar hypercarnivore that must cope with seasonal food scarcity, we observed many changes in genes associated with diet and body condition. All types of genomic variation examined (single nucleotide polymorphisms, gene family expansions, structural variants) contributed substantially to the identification of candidate

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 The Authors. *Molecular Ecology* published by John Wiley & Sons Ltd.

loci. This argues strongly for consideration of variation other than single nucleotide polymorphisms in comparative genomics studies aiming to identify loci of adaptive significance.

KEYWORDS

adaptation, gene family evolution, genomics, mustelids, positive selection, structural variation

1 | INTRODUCTION

The Mustelidae are the most ecologically and taxonomically diverse family within the mammalian order Carnivora, representing a remarkable example of adaptive radiation among mammals that is rich with recent speciation events (Koepfli et al., 2008; Liu et al., 2020). Closely related mustelid species often inhabit vastly different ecosystems, where they experience diverse environmental challenges and are thus exposed to different evolutionary pressures. This is particularly pronounced in the mustelid subfamily Guloninae, within which species occupy a variety of ecological niches, ranging from scansorial (adapted to climbing) omnivores in the neotropics to terrestrial hypercarnivores in circumpolar regions. Members of the Guloninae display a range of behavioural and physiological

adaptations associated with environment-specific resource availability, and consequently differ markedly in feeding ecology, reproductive strategy and morphology (Heldstab et al., 2018; Zhou et al., 2011). Here, we focus on tayra, wolverine and sable (Figure 1), for which genomic resources are now available.

The tayra (*Eira barbara*) is a predominantly diurnal, solitary species that inhabits tropical and subtropical forests of Central and South America, ranging from Mexico to northern Argentina (Wilson & Mittermeier, 2009). It is a scansorial, opportunistic omnivore, feeding on fruits, small mammals, birds, reptiles, invertebrates and carrion. Caching of unripe fruit for later consumption has been observed (Soley & Alvarado-Díaz, 2011). Unlike other gulonine species, which are characterized by seasonal breeding and embryonic diapause, the tayra is an aseasonal polyoestrous

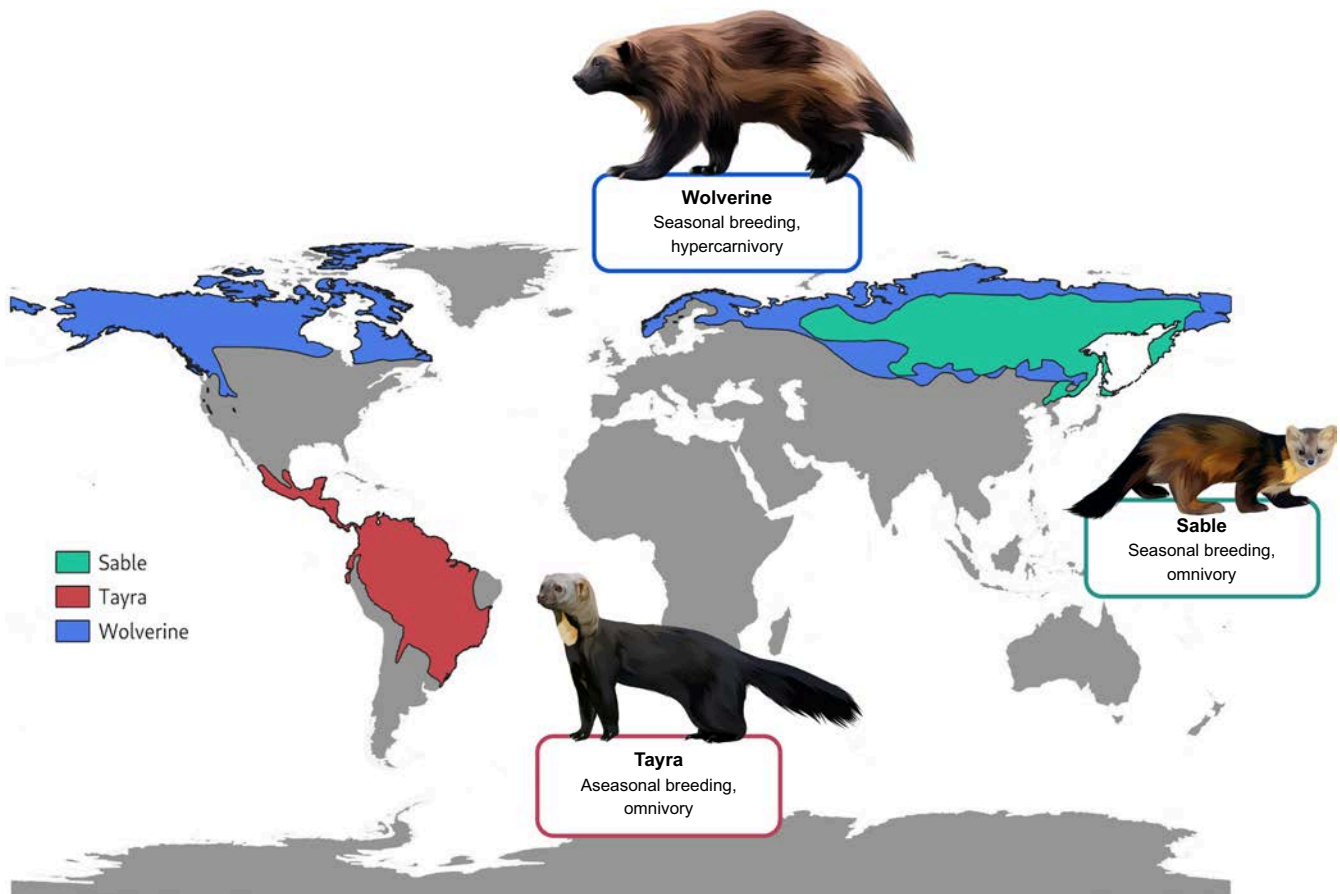


FIGURE 1 Distribution and species-specific traits of the tayra (*Eira barbara*), wolverine (*Gulo gulo*) and sable (*Martes zibellina*). Vector graphics of species are created based on royalty-free images (Source: Shutterstock)

breeder and does not exhibit delayed implantation (Proulx & Aubry, 2017), which may be due to the less prominent seasonality and fluctuation in food availability in neotropical habitats (Heldstab et al., 2018).

The largest terrestrial mustelid, the wolverine (*Gulo gulo*), is a circumpolar species, inhabiting alpine and boreal zones across North America and Eurasia (Ekblom et al., 2018). The wolverine is an opportunistic predator and facultative scavenger, either feeding on carrion or actively hunting medium- to large-sized mammals, such as roe deer, wild sheep and occasionally moose (Pasitschniak-Arts & Larivière, 1995). Morphological and behavioural adaptations such as dense fur, plantigrade locomotion facilitating movement through deep snow, and food caching enable wolverines to survive in cold habitats with limited food resources (Copeland & Kucera, 1997). In addition, wolverines occupy large home ranges, and display territoriality, seasonal breeding and delayed implantation, traits indicating an adaptive response necessary for survival in scarce resource environments (Inman et al., 2012).

The sable (*Martes zibellina*) is distributed in the taiga and deciduous forests of north central and northeastern Eurasia. The sable is solitary and omnivorous, relying on hearing and olfaction to locate prey, even under snow cover during winter months (Liu et al., 2020; Monakhov, 2011). Unlike wolverines, seasonal changes do not cause dramatic fluctuations in resource availability for sables as they are able to exploit a wider variety of food sources, and are adapted to tolerate short-term food scarcity (Mustonen et al., 2006). Their diet consists of small mammals, birds, nuts and berries, and in some instances food caching during the winter period has been reported (Monakhov, 2011). Similar to wolverines and many other species of mustelids, sables have a well-defined reproductive season and exhibit delayed blastocyst implantation (Proulx & Aubry, 2017).

To date, only a few studies have investigated adaptive variation in mustelids using comparative genomics (Abduriyim et al., 2019; Beichman et al., 2019; Liu et al., 2020; Miranda et al., 2021). Here, we generated a highly contiguous genome assembly of the tayra, an early diverging gulonine (Koepfli et al., 2008; Law et al., 2018), and compared it to previously published genomes of the wolverine and sable to identify the genetic basis underlying the adaptations to the diverse environments inhabited by these species.

In addition to identifying genes under positive selection, we investigated gene family evolution and structural variants (SVs), as these types of variants represent a significant source of intra- and interspecific genomic differentiation, affecting more nucleotides than single-nucleotide polymorphisms (SNPs) (Catanach et al., 2019). Gene copy number variation and large SVs can be associated with an adaptive response to new ecological circumstances (Rinker et al., 2019), and are thus an important source of genomic novelty to consider when studying adaptive divergence among species (Hecker et al., 2019). We focused on candidate loci linked to species-specific traits associated with response to environmental challenges, such as resource availability in the respective habitats of our study species.

2 | MATERIALS AND METHODS

2.1 | Sequencing, genome assembly and alignment

Whole blood from a captive (second-generation) male tayra was collected by the veterinary staff of the "Wildkatzenzentrum Felidae," Barnim (Germany), during a routine medical checkup. High-molecular-weight (HMW) genomic DNA extraction was performed using the Qiagen MagAttract HMW DNA Kit, following the manufacturer's protocol. We used 1 ng of DNA and the Chromium Genome Reagents Kits Version 2 and the 10x Genomics Chromium Controller instrument with a microfluidic chip for library preparation. Sequencing was carried out on an Illumina NovaSeq 6000 with 300 cycles on an S1 lane.

We generated a *de novo* genome assembly using the 10x Genomics Supernova assembler version 2.1.1 (Weisenfeld et al., 2017) with default parameters (assembly metrics given in Table 1; Table S1). The assemblies of tayra (this study, JAHRI000000000), wolverine (Ekblom et al., 2018; GCA_900006375.2), sable (Liu et al., 2020; GCA_012583365.1) and domestic ferret (MusPutFur1.0_HiC; Dudchenko et al., 2017, 2018; Peng et al., 2014) were assessed for gene completeness with BUSCO version 4.1.2 using the mammalian lineage data set mammalia_odb10 (Simão et al., 2015). To accurately identify repeat families, we used REPEATMODELER version 2 (Flynn et al., 2020) with the "-LTRstruct" option, followed by REPEATMASKER version 4.1.2 (Smit, 2004) to identify and mask the modelled repeats in the tayra genome assembly.

2.2 | Demographic reconstruction

Trimmed reads of all three gulonine were mapped to their respective genomes in local mode with BOWTIE2 version 2.3.5.1 (Langmead & Salzberg, 2012), and analysis of demographic history was performed with PSMC version 0.6.5 (Li & Durbin, 2011) using the following parameters (repeated 100 times for bootstrapping): `psmc -N25 -t15 -r5 -b -p '4+25*2+4+6' -o round-${ARRAY_TASK_ID}.psmc ${name}.split.psmcfa`. Results for each genome were plotted with `psmc_plot.pl`, and the mutation rate was set to 1e-08 substitutions per site per generation (Cahill et al., 2016; Dobrynin et al., 2015). Generation times were set to 7.4 years for tayra, 5.7 years for sable (Pacifiçi et al., 2013) and 6 years for wolverine (Ekblom et al., 2018).

2.3 | Reference-based scaffolding

Using the domestic ferret genome as a reference, we generated pseudochromosome assemblies for tayra, wolverine and sable, to visualize heterozygosity along chromosomes. Scaffolding was performed using RAGOO version 1.1 (Alonge et al., 2019). The X chromosome in the domestic ferret assembly was identified via whole genome alignment to the domestic cat (*Felis catus*) *Felis_catus_9.0* assembly (Buckley et al., 2020) and ZooFISH data available from

TABLE 1 Comparison of genome assembly metrics among four mustelid species

	<i>Tayra (Eira barbara)</i>	Domestic ferret (<i>Mustela putorius furo</i>)	Sable (<i>Martes zibellina</i>)	Wolverine (<i>Gulo gulo</i>)
Assembly accession/reference	JAHRI0000000000 (this study)	MusPutFur1.0_HiC (Dudchenko et al., 2017, 2018; Peng et al., 2014) GCF_000215625.1 (Dudchenko et al., 2017, 2018; Peng et al., 2014)	GCA_012583365.1 (Liu et al., 2020)	GCA_900006375.2 (Ekblom et al., 2018)
Sequencing/assembly approach	Illumina + 10× Genomics/Supernova	Illumina/ALLPATHS-LG + Hi-C scaffolding	Illumina/SOAPDENOV2	Illumina/SOAPDENOV2
Raw coverage (x)	75.6	162	114.5	76
Contig N50 (kb)	289.9	44.7	41.7	3.6
Scaffold N50 (Mb)	42.0	145.3	5.2	0.2
Number of scaffolds	14,579	7428	15,814	47,417
Total genome length (Gb)	2.44	2.40	2.42	2.42

the *Atlas of Mammalian Chromosomes* (Cavagna et al., 2000; O'Brien et al., 2020). Whole genome alignment was performed using LAST version 971 (Frith & Kawaguchi, 2015).

Variant calling followed by quality filtration was performed using the BCFTOOLS pipeline version 1.10 (Poplin et al., 2018). Low-quality variants were removed (BCFTOOLS filter, "QUAL < 20.0 || (FORMAT/SP > 60.0 || FORMAT/DP < 5.0 || FORMAT/GQ < 20.0)"). In each sample, positions with coverage lower or higher than 50%–250% of the whole genome median value were removed. Of the remaining positions only those common to all samples were retained. Finally, SNPs with uncalled genotypes in any sample and variants with the same genotypes for all samples were removed. For visualization, heterozygous SNPs were counted in nonoverlapping sliding windows of 1 Mbp (counts scaled to SNPs per kbp). Indels were not included due to the low quality of calls from short reads. SNP density plots were created using the MACE package (<https://github.com/mahajrod/MACE>).

2.4 | Phylogenomic data preparation, analysis and dating

We performed sequence alignments and filtering of excessively divergent segments in each of 6020 coding genomic regions of single-copy orthologues shared across eight species of carnivores, using the software MACSE version 2 (Ranwez et al., 2011). Our taxon set included domestic cat (*Felis catus*), domestic dog (*Canis familiaris*), northern elephant seal (*Mirounga angustirostris*) and walrus (*Odobenus rosmarus*), in addition to four mustelid species. To extract the most reliable signal from these coding data, we excluded whole alignments that were excessively divergent, contained excessive missing data or violated basic substitution model assumptions (further details in the Supporting Information). This led to a phylogenomic data set with 2457 gene regions comprising over 3.2 million nucleotide sites. Gene trees were estimated from

gene regions by first selecting the best substitution model from the GTR+F+I+R family (Kalyaanamoorthy et al., 2017), and calculating approximate likelihood-ratio test (aLRT) branch supports (Anisimova & Gascuel, 2006), as implemented in IQ-TREE version 2 (Minh et al., 2020a,b).

Species tree estimates were performed using (i) concatenated sequence alignments for maximum-likelihood inference using IQ-TREE version 2, and (ii) gene trees for inference under the multispecies coalescent using the summary coalescent method in ASTRAL-III (Zhang et al., 2018). The maximum-likelihood estimate of the species tree was accompanied by aLRT branch supports, while summary coalescent inference was accompanied by local posterior probabilities (Sayyari & Mirarab, 2016). The decisiveness of the data regarding the phylogenetic signals was examined using gene- and site-concordance factors, calculated in IQ-TREE version 2 (Minh, Schmidt, et al., 2020).

Bayesian molecular dating analysis was performed using MCMCtree in PAML version 4.8 (Yang, 2007). To minimize the violation of the time-tree prior (Angelis & Dos Reis, 2015) and the negative impact of gene tree discordance on rate estimates (Mendes & Hahn, 2016), we only included genomic regions with gene trees concordant with the species tree, and assumed the reconstructed species tree from ASTRAL-III (see Supporting Information for further details). This led to a data set for molecular dating that included 992 single-copy orthologous gene regions, comprising 0.53 million sites. The data were partitioned by codon positions, each modelled under individual GTR+I substitution models. We used an uncorrelated gamma prior on rates across lineages and a birth–death prior for divergence times. Fossil calibrations are listed in the Supporting Information. The posterior distribution was sampled every 1×10^3 Markov chain Monte Carlo (MCMC) steps over 1×10^7 steps, after a burn-in phase of 1×10^6 steps. We verified convergence to the stationary distribution by comparing the results from two independent runs, and confirming that the effective sample sizes for all parameters were above 1,000 using the R package coda (Plummer et al., 2006).

2.5 | Positive selection on single-copy orthologues

To investigate genes under positive selection, the coding sequences (CDS) corresponding to 1:1 orthologues were aligned for the eight aforementioned carnivoran species. Multiple sequence alignments (MSAs) were constructed with PRANK version 120716 (Löytynoja, 2014), and 17 MSAs were removed due to short alignment length. The CODEML module in the PAML version 4.5 package was used to estimate the ratio of nonsynonymous to synonymous substitutions, also called d_N/d_S or ω (Yang, 2007). We applied the one-ratio model to estimate the general selective pressure acting among all species, allowing only a single d_N/d_S ratio for all branches. A free-ratio model was also used to estimate the d_N/d_S ratio of each branch. Furthermore, the CODEML branch-site test for positive selection was performed on 6003 orthologue alignments for three separate foreground branches: *Eira barbara*, *Gulo gulo* and *Martes zibellina* (Zhang et al., 2005). Statistical significance was assessed using likelihood ratio tests (LRTs) with a conservative 10% false discovery rate (FDR) criterion (Nielsen et al., 2005). Orthologues with a free-ratio >2 in the branch model were considered for further analysis of signatures of positive selection.

To account for differences in genome assembly quality, we evaluated the alignments of selected orthologues based on the transitive consistency score (TCS), an extension to the T-Coffee scoring scheme used to determine the most accurate positions in MSAs (Chang et al., 2014). Additionally, alignments were visually inspected for potential low-scoring MSA portions.

2.6 | Gene family evolution

To investigate changes in gene family sizes, we constructed a matrix containing 7838 orthologues present as either complete "single-copy," complete "duplicated" or "missing," identified using the BUSCO genome assembly completeness assessment of all eight carnivoran genomes. Orthologues were retained if they were detected in at least four species (including *Felis catus* as an outgroup) to obtain meaningful likelihood scores for the global birth and death (λ) parameter.

We applied a probabilistic global birth and death rate model of CAFE version 4.2.1. (Han et al., 2013) to analyse gene gains ("birth") and losses ("death") accounting for phylogenetic history. First, we estimated the error distribution in our data set, as genome assembly and annotation errors can result in biased estimates of the average rate of change (λ), potentially leading to an overestimation of λ . Following the error distribution modelling, we ran the CAFE analysis guided by the ultrametric tree estimated earlier, calculating a single λ parameter for the whole species tree. The CAFE results were summarized (Table S4A) with the python script *cafetutorial_report_analysis.py* (<https://github.com/hahnlab/CAFE>).

We examined differences between duplicates arising through gene family expansion, to determine how these paralogues differed and if a signal of selection could be detected. Pairwise codon-aware sequence alignment of paralogues was performed with DIALIGN-TX

version 1.0.2 (Subramanian et al., 2008). Ratios of nonsynonymous to synonymous substitution rates were estimated using KAKS_CALCULATOR version 2.0 (Zhang et al., 2006; details are given in the Supporting Information). Paralogues with identical nucleotide sequences were considered to be recent duplications ("NAs" in Table S5B).

2.7 | Structural variation

To avoid reference genome bias, preprocessed reads from the three Guloninae were aligned to the domestic ferret (*Mustela putorius furo*) genome with BOWTIE2 version 2.3.5.1 (Langmead & Salzberg, 2012) (details given in Supporting Information). Duplicated reads were removed with PICARD TOOLKIT version 2.23 (MarkDuplicates, Broad Institute, 2019). Trimmed tayra reads were downsampled to $\sim 38\times$ with SEQTK version 1.3 (<https://github.com/lh3/seqtk>) prior to mapping to maintain uniformity among libraries and to avoid bias in variant calling.

We applied an ensemble approach for SV calling, encompassing three SV callers: MANTA version 1.6.0 (Chen et al., 2016), WHAMG version 1.7.0 (Kronenberg et al., 2015) and LUMPY version 0.2.13 (Layer et al., 2014). SV calls originating from reads mapping in low-complexity regions and with poor mapping quality were removed from all three call sets. We retained MANTA calls with paired-read (PR) and split-read (SR) support of $PR \geq 3$ and $SR \geq 3$, respectively. To reduce the number of false positive calls, the WHAMG call set was filtered for potential translocation events, as WHAMG flags but does not specifically call translocations. We further removed calls with a low number of reads supporting the variant (PR, SR) from the WHAMG ($A < 10$) and the LUMPY call set ($SU < 10$). All SV call sets were filtered based on genotype quality ($GQ \geq 30$). WHAMG and LUMPY SV call sets were genotyped with SVTYPER version 0.7.1 (Chiang et al., 2015) prior to filtering. Only scaffolds assigned to chromosomes were included in further analyses. SURVIVOR version 1.0.7 (Jeffares et al., 2017) was used to merge and compare SV call sets within and among samples. The union set of SV calls among the three gulonine species containing species-specific and shared variants was annotated, using LIFTOFF version 1.5.1 (Shumate & Salzberg, 2020), for preparation of reference genome annotation, and Ensembl Variant Effect Predictor version 101.0 (McLaren et al., 2016) for identifying variants affecting protein-coding genes. Gene ontology analysis was performed with SHINY GO (Ge et al., 2020) with an $FDR < 0.05$ for each SV type (excluding inversions) overlapping multiple protein-coding genes (more than five genes).

2.8 | Candidate loci

The functional and biological roles of positively selected genes, loci affected by changes in gene family size, and structural variants, were explored using literature sources and online databases, including OrthoDB version 10 (Kriventseva et al., 2019), Uniprot (The UniProt Consortium, 2017) and NCBI Entrez Gene (Maglott et al., 2011).

Gene descriptions, GO biological processes, functions and relevant citations are provided in the supporting tables (see below). Gene Ontology enrichment analysis was performed with SHINY GO version 0.65 (Ge et al., 2020), for gene sets obtained from previously mentioned analyses (positive selection on single genes, PSG; gene family evolution, GF; and SV) for the three gulonine species. Gene sets were inspected for significant enrichment of biological processes with the following parameters: best matching species, top 10 pathways, and FDR p -value cutoff.05. SHINY GO version 0.65 is based on a database derived from Ensembl Release 103.

3 | RESULTS

3.1 | Genome assembly

We generated a highly contiguous reference genome assembly for the tayra (*Eira barbara*). Extracted genomic DNA had an average molecular size of 50.75 kb and was sequenced to ~76-fold coverage (Table S1). The final assembly showed a total length of ~2.44 Gb (excluding scaffolds shorter than 5 kb), with a contig N50 of 290 kb, scaffold N50 of 42.1 Mb, and identity in 95% of all positions in an alignment with the domestic ferret genome (Figure S1). The tayra assembly has higher contiguity than the Illumina-only-based assemblies of both wolverine and sable, but it is more fragmented than the chromosome-length domestic ferret (*Mustela putorius furo*) assembly (Table 1; Figure S2A) that we used as a reference genome for some analyses. The haploid tayra genome of ~2.4 Gb is contained in 162 scaffolds (>100 kb) with 40 scaffolds having a length above 50 Mb (Figure S2A).

The tayra assembly has high gene completeness as assessed with BUSCO version 4.1.2 using 9226 conserved mammalian orthologues in total, 8540 (92.5%) complete benchmarking Universal Single-Copy Orthologs (BUSCOs) were identified, encompassing 8492 (92.0%) of complete and single-copy, and 48 (0.5%) complete and duplicated orthologues. Additionally, 104 (1.1%) orthologues were fragmented and 582 (6.4%) were missing. As measured by this metric, the tayra genome has higher gene completeness than the published genomes of wolverine, sable or domestic ferret (Figure S2B).

3.2 | Repetitive elements

The repeat landscape of the tayra assembly contains ~0.85 Gb of repetitive elements (Table S2). L1 type LINE elements are the most abundant, constituting 23% of the tayra genome. L1 elements also show signs of recent proliferation in comparison to DNA transposons and LTR retroelements (Figure S3). Endogenous retroviruses constitute 3.8% of the tayra genome and can be classified as Gammaretroviruses and Betaretroviruses.

The overall repeat landscape of the tayra genome assembly is comparable to other carnivore genomes (Liu et al., 2020; Peng et al., 2018). It is similar to that of the sable genome, differing mostly in

the number of L1 LINE elements, which have been recently proliferating and accumulating within the tayra genome more than in other Guloninae genomes. The diversity of endogenous retroviruses is similar to that of other mustelids. Although endogenous delta-retroviruses have been described from a broad range of mammal genomes, including several smaller carnivores such as mongoose (family Herpestidae) and the fossa (*Cryptoprocta ferox*) (Hron et al., 2019), no delta-retroviruses were found in the genome of tayra.

3.3 | Demographic reconstruction

Reconstruction of historical demography for tayra, wolverine and sable, using the Pairwise Sequentially Markovian Coalescent model (PSMC; Li & Durbin, 2011) revealed different trends in effective population sizes (N_e) (Figure S4). While the trajectories for all three species involve multiple declines and rebounds in N_e , the timing, duration and magnitude of these differ. In tayras, there are three extensive declines beginning around 4.5 million years ago (Ma) (>35% reduction in N_e), 500 thousand years ago (ka) (>30%) and 80 ka (>80%), resulting in a recent N_e of ~14,000 individuals. In wolverines, a sharp decline 1 Ma (>45%) is followed by a plateau in N_e and subsequent decline beginning ~400 ka (>30%), followed by a moderate rebound and final decline beginning 40 ka (>80%), resulting in a recent N_e of ~2,500 individuals. In sable, N_e gradually declines until ~500 ka (>40%), followed by a moderate rebound and sharp decline around 200 ka (>40%). This is followed by an extensive rebound beginning 100 ka and subsequent sharp declines 50 ka (>30%) and 15 ka (>50%), resulting in a recent N_e of ~23,000 individuals. Thus, all three species exhibit complex historical trends in N_e , with little overlap in timing or magnitude.

Consistent with Ekblom et al. (2018), we observe low recent N_e in wolverines. However, our reconstructed trajectory of historical N_e for the wolverine differs from that of Ekblom et al. (2018), probably owing to different PSMC parameters. That notwithstanding, the timing and relative magnitude of the final decline is in broad agreement in the two studies.

3.4 | Nucleotide diversity

The tayra, sable and wolverine assemblies were generated using different approaches and differ significantly in contiguity (Table 1). To compare nucleotide diversity among the Guloninae, we generated pseudochromosome assemblies for each species using the chromosome-length assembly of the domestic ferret as a reference. The domestic ferret has more chromosomes than the other species ($2n = 40$ vs. $2n = 38$), and the same number (20) of pseudochromosomes (Lewin et al., 2019) were obtained after scaffolding in each case. For each assembly, we identified the X chromosome (labelled as *ps_chrX*) and arranged pseudoautosomes (labelled as *ps_aut1* – *ps_aut19*) according to the length of the original scaffolds in the domestic ferret reference. This allowed us to verify the sex of the

animals using a coverage-based approach (Figure S5), which confirmed morphological sexing for the tayra (male) and wolverine (female) individuals. While the sable individual is referred to as a male (Liu et al., 2020), our analysis suggests it is a female (further details given in Supporting Information).

We counted heterozygous SNPs in 1-Mbp stacking windows for all three species and scaled it to SNPs per kbp (Figure 2). Median values for tayra, sable and wolverine were 1.89, 1.44 and 0.28 SNPs per kbp, respectively, the last being in agreement with previous findings (Ekblom et al., 2018). All scaffolds of ≥ 1 Mbp in pseudochromosome assemblies were taken into account. Exclusion of ps_chrX resulted in slight increases of medians to 1.93, 1.47 and 0.29, respectively (Table S3, Figure S6). Regardless of whether ps_chrX was included or excluded, the genome-wide diversities among the three species were significantly different ($p < 0.001$, Mann-Whitney test).

3.5 | Phylogenomics and molecular dating

We reconstructed the phylogenetic relationships of the following carnivoran species: domestic cat, domestic dog, walrus, northern elephant seal, domestic ferret, wolverine, sable and tayra. Phylogenomic analyses using concatenation and summary coalescent methods led to an identical resolution of the relationships among mustelid taxa. Within Guloninae, the wolverine and sable were placed as sisters, to the exclusion of the tayra (Figure 3). Branch supports were maximal across all branches using both aLRT and local posterior probabilities. Similarly, concordance factors for genes and sites (gCF, sCF) were high across branches and consistently more than twice as high as the values of the discordance factors (gDF, sDF). The lowest concordance factors were those in support of the resolution of *Gulo* and *Martes* as sisters (gCF = 64.52, sCF = 54.38). However, the discordance factors were less than half these values (gDF < 14, sDF < 24), suggesting substantial decisiveness across genes and sites for this resolution.

Divergence time estimates across mustelids were largely in agreement with previous findings (Koepfli et al., 2008; Law et al., 2018; Li et al., 2014; Sato et al., 2012), placing the split between *Mustela* and Guloninae at 11.2 Ma (highest posterior density interval (HPDI) between 13.1 and 9.5 Ma), and the split between *Eira* and the *Gulo*-*Martes* group at 7.5 Ma (HPDI between 9 and 6.1 Ma). The split between *Gulo* and *Martes* was dated at 5.9 Ma (HPDI between 7.4 and 4.7 Ma).

3.6 | Positive selection on single-copy orthologues

In the three Guloninae species, we found sites under positive selection ($5 > d_N/d_S > 1$; Barnett et al., 2020) in 55 single-copy orthologues

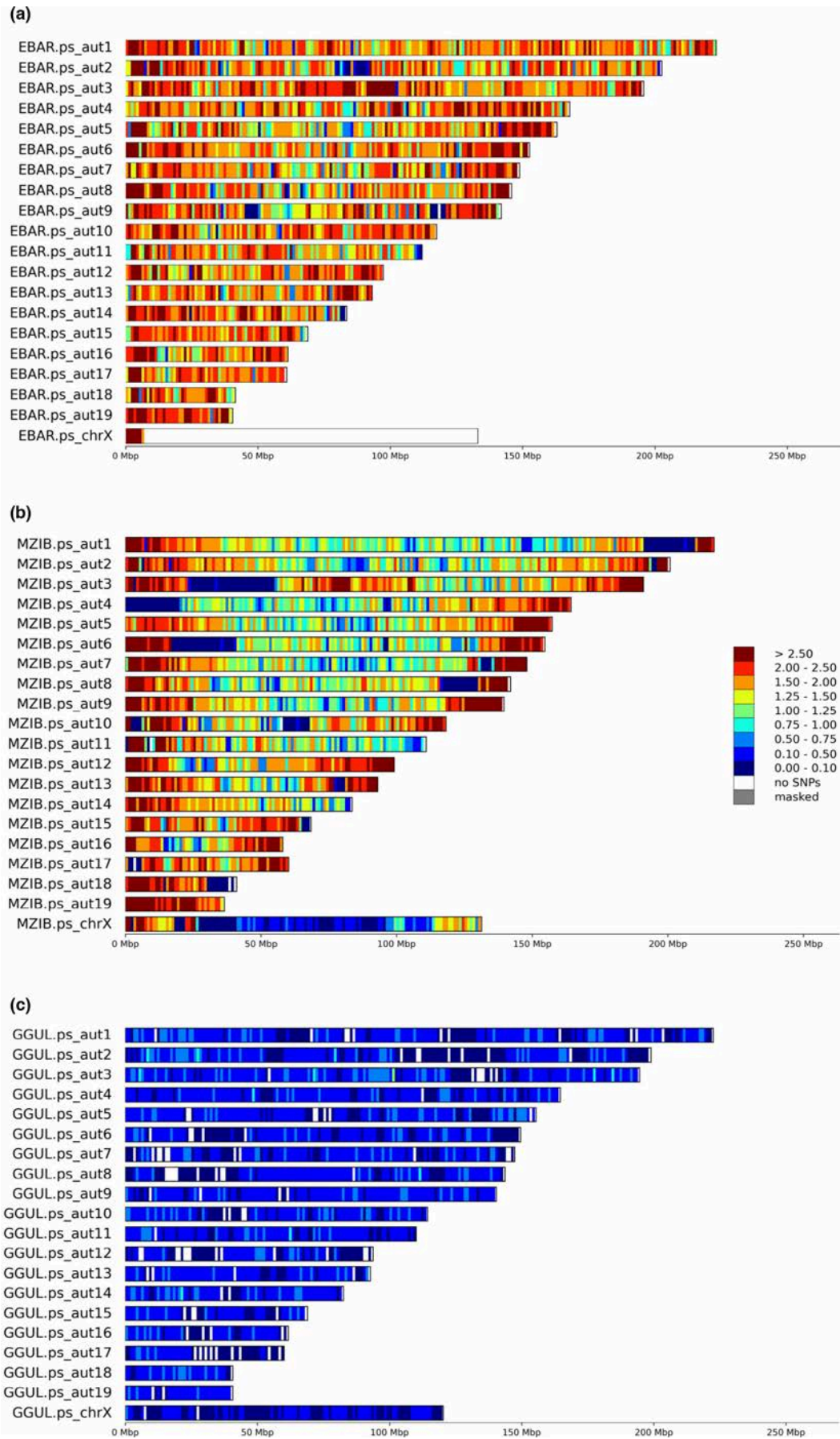
that were highly significant (free-ratio > 2). Of these 55 positively selected genes (PSGs), 15 were observed in tayra, 22 in wolverine and 18 in sable (Figure 4a,b). Gene names, descriptions and functions are given in Table S4.

Among the 15 PSGs we detected in tayra, five are associated with reproduction (*NSMCE1*, *ETV2*, *SPATA25*, *MUC15* and *PIH1D2*) with functions involving spermatogenesis, placenta and embryo development, and blood vessel morphogenesis. Among the remaining 10 PSGs, *HSPB6* is involved in vasodilation and muscle contraction, *DERA* is associated with environmental stressors, including exposure to toxins, and uricase (*UOX*) is a liver enzyme involved in purine catabolism and regulation of fructose metabolism. Three PSGs (*IP6K3*, *MAGIX* and *FAM149B1*) are found to be associated with the nervous system, synapse formation and structural plasticity, as well as motor skills and coordination. Three further PSGs (*DUSP19*, *TNLG2B* and *LRR46*) are related to the immune system and *HEMK1* regulates methylation processes. Gene enrichment analysis revealed an overrepresentation of genes in gene ontology (GO) categories associated with reproduction (Table S4A; GO:0046483, "Heterocycle metabolic process," $p = .022$) and metabolism/energy conversion (Table S4A, multiple pathways, $p < .03$).

We detected 22 PSGs in wolverine, including six genes associated with energy production and conversion. Among them, *ATP6V0B*, *KMO* and *SLC16A4* are primarily involved in insulin level regulation, and the metabolism of carbohydrates and fatty acids. Three PSGs (*OIP5*, *ZADH2* and *MTPAP*) are specifically associated with adipose tissue formation and intramuscular fat deposition. Additionally, we found three PSGs (*NBR1*, *TMEM38B*, *PPP1R18*) involved in selective autophagy as a response to nutrient deprivation along with bone mass and density regulation, and resorption. We also detected PSGs (*DAB1*, *OPA1* and *CTNS*) linked to cognition, brain development and vision. Several PSGs (*BNIPL*, *IL18BP*, *CRNN*) were associated with the immune system; three others (*ANAPC7*, *RNF212B*, *IZUMO3*) are involved in reproduction processes and *USB1* and *CLCN4* have a role in basal cell cycle processes. For the remaining two, *CEP95* and *FAM185A*, it was not possible to associate a specific phenotypic trait. No overrepresentation of GO categories was detected.

Among the 18 PSGs detected in sable, three (*PRRT2*, *ATL2*, *SELENOI*) are associated with locomotion and coordination, and *USP53* is associated with sensory perception and the nervous system. Two PSGs, *VEGFC* and *RASA1*, are associated with blood vessel formation, three (*TTC4*, *ZBP1*, *CD247*) with the immune system and three (*IQUB*, *UBQLNL*, *MEIKIN*) with reproduction. Several PSGs (*EEF2KMT*, *DEUP1*, *ECD*, *IQCK*) are associated with cell cycle processes, while *ZC2HC1C* and *CCDC17* could not be associated with a particular biological process. Gene enrichment analysis revealed an overrepresentation of GO categories associated with ear morphogenesis (Table S4A, multiple pathways, $p < .05$).

FIGURE 2 Heterozygosity density among pseudochromosomes for (a) tayra, (b) sable and (c) wolverine. Heterozygous SNPs were counted in stacking windows of 1 Mbp and scaled to SNPs per kbp. Tayra is a male individual and thus heterozygous SNP density is underestimated (due to only one X chromosome), while sable and wolverine are females and therefore likely to be representative of true SNP density



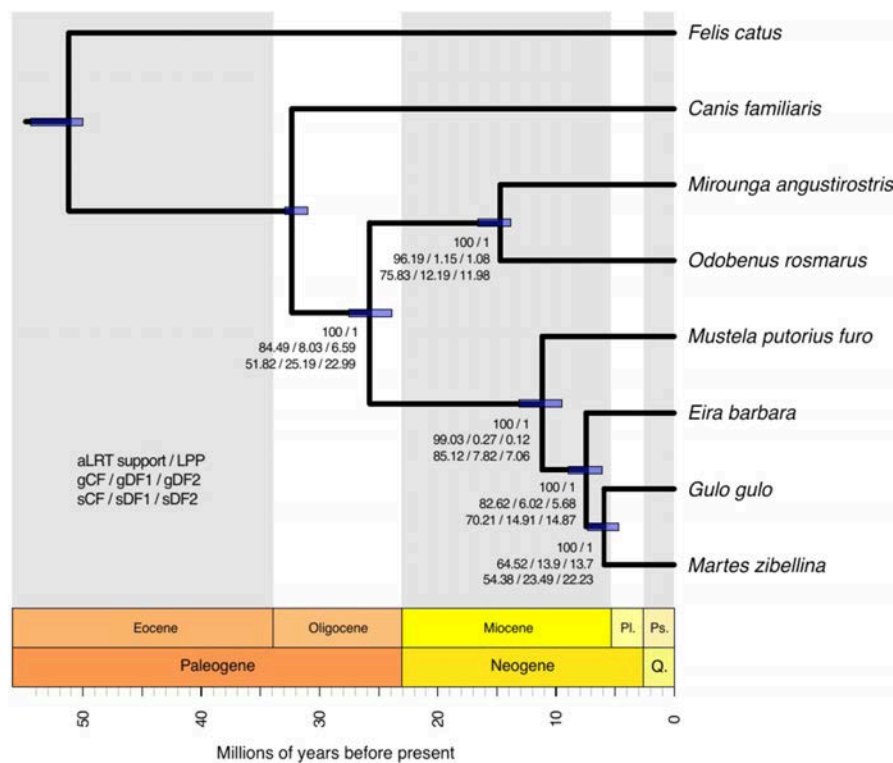


FIGURE 3 Phylogenetic tree and divergence times of Guloninae and five other carnivorans. The mean age of each node is shown, with 95% confidence intervals depicted as purple bars. The gene and site concordance (gCF, sCF) and discordance (gDF, sDF) factors are given. While the concordance factors refer to the portions of the data in agreement with the tree shown, each of the two discordance factors (DF1 and DF2) refer to the support for each of the two other possible alternative quartet resolutions for each branch. Also included are the tree branch supports as calculated using approximate likelihood-ratio tests (aLRT) and local posterior probabilities (LPP)

3.7 | Gene family expansions and contractions

Adaptive divergence between species may also be caused by changes in gene family sizes that occur during genome evolution and are due to gains (expansions) or losses (contractions) of genes or groups of genes (Olson, 1999; Tigano et al., 2020). All species analysed displayed more gene family contractions than expansions, with the wolverine having the highest contraction rate. This is probably an artefact resulting from the fragmented genome assembly of this species (Figure S7). All identified expansions were in the form of gene duplications, with one putative triplication detected in tayra (Table S5B).

Tayra and sable had similar numbers of gene family expansions and contractions (Figure S7): 34 expansions and 169 contractions in tayra, and 33 expansions and 162 contractions in sable. The less contiguous wolverine genome contained seven expansions and 649 contractions (Table S5A,B). Due to the stochastic nature of gene losses and the potential inflation of estimates resulting from different genome assembly contiguities, we focus here on gains of gene copies.

Expanded gene families in the tayra genome are associated with reproduction, metabolism, the nervous and immune system, and cell cycle, among others (Figure 4c,d; gene names, descriptions and functions are given in Table S5B). Of the three reproduction-related genes, *SLC38A2* regulates supply of nutrients for fetal growth through the placenta during the peri-implantation period. The second, *HSD17B10*, is associated with regulation of pregnancy-sustaining steroid hormones, and *RBP2* is involved in retinol binding and vitamin A metabolism, necessary for oogenesis and embryogenesis, as well

as vision. Four genes (*PDHB*, *SH3GLB1*, *SLC35A1*, *N6AMT1*) are involved in metabolic processes, with *N6AMT1* specifically associated with modulation of arsenic-induced toxicity. Four genes (*ATP6V1D*, *DBX2*, *SLC38A1*, *MAPKAPK5*) are associated with cerebral cortex development, synapse formation, visual perception and learning processes. *ANKRD13A* is also associated with vision, more specifically with lens fibre generation and vitamin A metabolism. The olfactory receptor gene *TAAR5* is involved in behavioural responses in mammals, and was duplicated in both tayra and sable. We detected one putative triplication of *FKBP3*, a gene associated with immunoregulation, predominantly of T-cell proliferation. Four additional genes are associated with the innate immune system (*TUFM*, *UBXN6*, *SPON2*, *SERPINB1*). Two genes (*MRPS14* and *MRPS23*) are involved in energy conversion. Two genes (*ATF4* and *ARDI2*) are associated with the cardiovascular system and development, respectively. The rest of the genes are involved in processes related to the cell cycle, and *PRR11* could not be associated with a particular biological process. Among duplications in the tayra, seven were recent duplications, 16 are under relaxed selection and 10 under purifying selection (Table S5B). Gene enrichment analysis revealed an overrepresentation of genes in GO categories associated with metabolism/energy conversion (Table S5C, multiple pathways, $p < .02$), the cardiovascular system (multiple pathways, $p < .02$), the immune system (multiple pathways, $p < .02$) and cell cycle processes (multiple pathways, $p < .02$).

In the wolverine, two duplicated genes are related to the nervous system: *GFRA4* is implicated in motor neuron development and *KCNS1* in regulating mechanical and thermal pain sensitivity. *MTM1* is associated with positive regulation of skeletal muscle tissue growth and *MON1B* is implicated in the immune response to viral

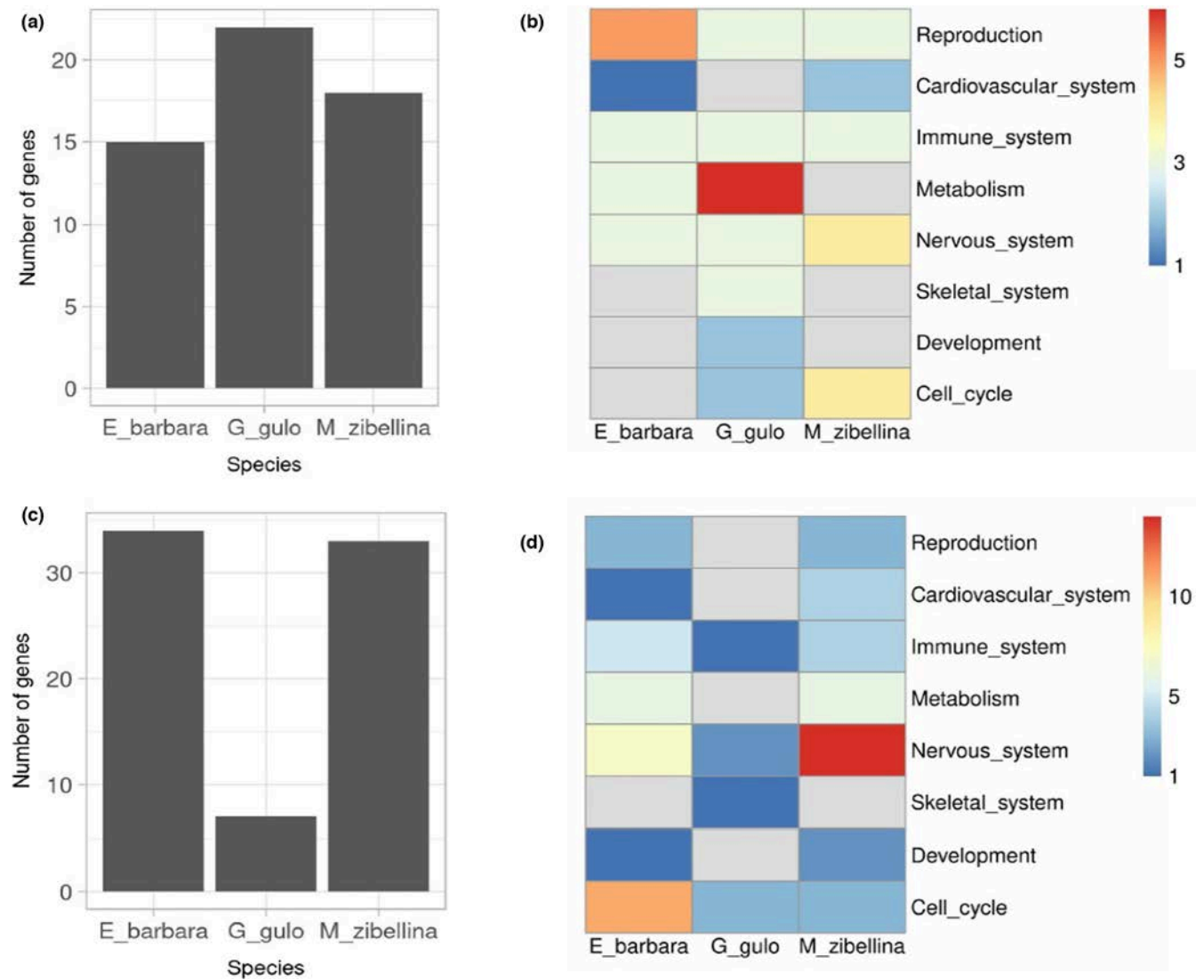


FIGURE 4 Number of candidate genes and their functional groups. Genes identified from analyses of (a, b) positive selection on single genes (PSG), and (c, d) gene family expansions. Heatmap scale represents the number of genes. Heatmap cells in grey indicate no observations for a given variable

infection. Three duplicated genes are associated with cell cycle processes. Among duplications in the wolverine, one is a recent duplication, one is under relaxed selection and two are under purifying selection (Table S5B). No overrepresentation of GO categories was detected.

In the sable, expansions involve gene families associated with the nervous system, metabolism, angiogenesis, hair follicle development and the immune system, among others (Table S5B). Fourteen genes are associated with the nervous system. Among them, six (*PPA1*, *THOC6*, *SHISAL2A*, *SHISAL1*, *FICD*, *DUSP8*) are involved in neuronal development, two (*SYNGR3*, *TM4SF20*) are associated with locomotion, and two (*MFSD5* and *BICDL2*) with energy regulation and secretion. *TAAR5* is associated with olfaction, *FBXL3* with regulation of the circadian clock and *TIMM10* with hearing. *CD93* is associated with regulation of inflammation in the central nervous system. Six genes are involved with metabolism and energy conversion (*ASB6*,

CRYZL1, *SLC25A10*, *CLDN20*, *RNF186*, *BORCS6*). Three genes (*GPS2*, *BRICD5*, *HCFC1R1*) are associated with the immune system. Two genes, *CDC42* and *TCHHL1*, are implicated in hair-follicle development. Additionally, *CDC42*, a gene coding for a cell division control protein, is also involved in angiogenesis and haematopoiesis, alongside *SLC25A39*, *TNFRSF12A* and *LXN*. Three genes (*MRPL38*, *RPP30*, *TBL3*) are associated with basal cell cycle processes and two genes (*SEPT12*, *CDK2*) with gametogenesis.

Among duplications in the sable, seven were recent duplications, nine are under relaxed selection, 15 are under purifying selection and two are under positive selection (Table S5B). Gene enrichment analysis revealed an overrepresentation of genes in GO categories associated with metabolism/energy conversion (Table S5C; GO:0006839, "Mitochondrial transport," $p = .018$), the nervous system (multiple pathways, $p < .03$) and cell cycle processes (multiple pathways, $p < .03$).

3.8 | Structural variation

SVs modify the structure of chromosomes and can affect gene syntax, repertoire, copy number and/or composition (e.g. gain or loss of exons), create linkage-blocks and modify gene expression (Chiang et al., 2017; Mérot et al., 2020), leading to complex variation in phenotypes and genetic diseases (Weischenfeldt et al., 2013). We investigated four types of SVs (deletions, duplications, insertions, inversions) in the three Guloninae relative to the domestic ferret genome.

We identified the highest number of species-specific SVs in sable (22,979), followed by tayra (8907) and wolverine (264) (Figure 5a). The most abundant SVs detected in all three species are deletions (>50 bp), ranging from 183 species-specific deletions in wolverine to 21,713 in sable. Duplications were the least frequent SV type among the three species (Figure S8A). For all three species, the majority of SVs are located in intergenic regions (>80%), with a smaller proportion found in genic regions, completely or partially overlapping protein-coding genes (untranslated regions, exons, introns). According to Variant Effect Predictor (VEP) classification, SVs impacting genic regions are classified either as high-impact variants or modifiers (McLaren et al., 2016) with putative consequences on gene transcription ranging from transcript truncation to transcript ablation or amplification. The highest number of species-specific genic SVs was detected in tayra, with 330 (3.70% of species-specific SVs), followed by 156 (0.68%) in sable and 53 (20.08%) in wolverine (Figure S8B). Other than the well-documented impact of inversions on intra- and interspecific gene flow (Porubsky et al., 2020; Wellenreuther & Bernatchez, 2018), determining the impact of inversions overlapping large sets of genes is still challenging, as the largest effect is likely to be restricted to genes near SV breakpoints. Therefore, we restricted our examination of gene function to loci affected by deletions, duplications and insertions (Figure 5b; gene names, descriptions and functions are given in Table S6).

In the tayra genome, we observed 14 duplications spanning a combined length of 2.92 Mb, putatively affecting 24 protein-coding genes. Duplicated genes and gene blocks are associated with reproduction, olfaction, metabolism and energy conversion. This included *RNASEH2B*, a gene involved in *in utero* embryo development, and two genes involved in spermatogenesis, *DIAPH3* and *PCNX1*, with the latter an example of a complex SV involving heterozygous duplication and deletion of an exon (SV ~2 kb in length). We detected 212 deletions in the tayra genome in relation to the domestic ferret reference, comprising a total length of 2.08 Mb, and affecting 247 genes, which are associated with reproduction, metabolism/energy conversion, the nervous system and cell cycle processes, among other functional categories (Table S6). Genes involved in placenta development and *in utero* embryogenesis include *HSF1*, *RSPO2* and *DNMT3A*. Additionally, we detected *NLRP1* and *NLRP8*, both associated with pre-implantation development, and highly expressed in oocytes. One short insertion was observed in *LIX1L*, a gene associated with anatomical structure morphogenesis. No overrepresentation of GO categories was detected.

In the wolverine genome, no duplications overlapped genic regions. However, 47 deletions spanning a combined length of 229 kb are putatively associated with transcript truncation or ablation in 48 genes. The majority of affected genes are associated with metabolism/energy conversion, development and basic cell cycle processes. These include *GLUD1*, a gene involved in amino-acid-induced insulin secretion, also found to be affected by a shorter deletion in sable, and *NSDHL*, a gene regulating cholesterol biosynthesis. Additionally, we detected deletions affecting *PARVA*, a gene associated with angiogenesis and smooth muscle cell chemotaxis, and *DNAJC7*, involved in positive regulation of ATPase activity and regulation of cellular response to heat. We also detected one insertion in a gene of unknown function. No overrepresentation of GO categories was detected.

In the sable genome, we detected 11 duplications spanning a combined length of 324 kb, overlapping 16 genes associated with sensory perception, development, the cell cycle and the immune system. The 130 detected deletions (combined length of 408 kb) overlap 125 protein-coding genes associated with reproduction, the immune system, development, metabolism, sensory perception and the cell cycle. Deletions were identified in two genes involved in keratinocyte differentiation, *PPHLN1*, also affected in wolverine, and *IVL*, associated with hair follicle development. Additionally, two short insertions were found in *NCOA4* and *YIPF5*, genes associated with mitochondrial iron homeostasis and protein transport, respectively. Gene enrichment analysis revealed an overrepresentation of genes in GO categories associated with cellular responses to xenobiotic compounds (Table S6A; multiple pathways, $p < .05$).

4 | DISCUSSION

Here, we present a highly contiguous genome for the tayra (*Eira barbara*). Contiguity of the assembly and its gene completeness are similar to or higher than those of other carnivoran species using the same sequencing approach (Armstrong et al., 2019; Etherington et al., 2020; Kim et al., 2020), confirming the utility of linked reads for assembly of mammal genomes.

Phylogenomic relationships among the mustelids resulted in a tree topology and divergence time estimates in agreement with previous studies using fewer loci (Koepfli et al., 2008; Koepfli et al., 2018; Li et al., 2014; Sato et al., 2012). We estimated the split between *Mustela* and Guloninae occurred 11.2 Ma (HPDI 13.1–9.5 Ma), followed by the split between *Eira* and the *Gulo-Martes* group 7.5 Ma (HPDI 9–6.1 Ma), and the split between *Gulo* and *Martes* at 5.9 Ma (HPDI 7.4–4.7 Ma).

Perhaps unsurprisingly, we observed different historical trends in effective population size among the three Guloninae. They differ markedly in ecology, and it is not unexpected that climatic and environmental changes (affecting, for example, habitat, ecological competition, prey and pathogens) also differentially impacted tayra, wolverine and sable populations. Consistent with previous

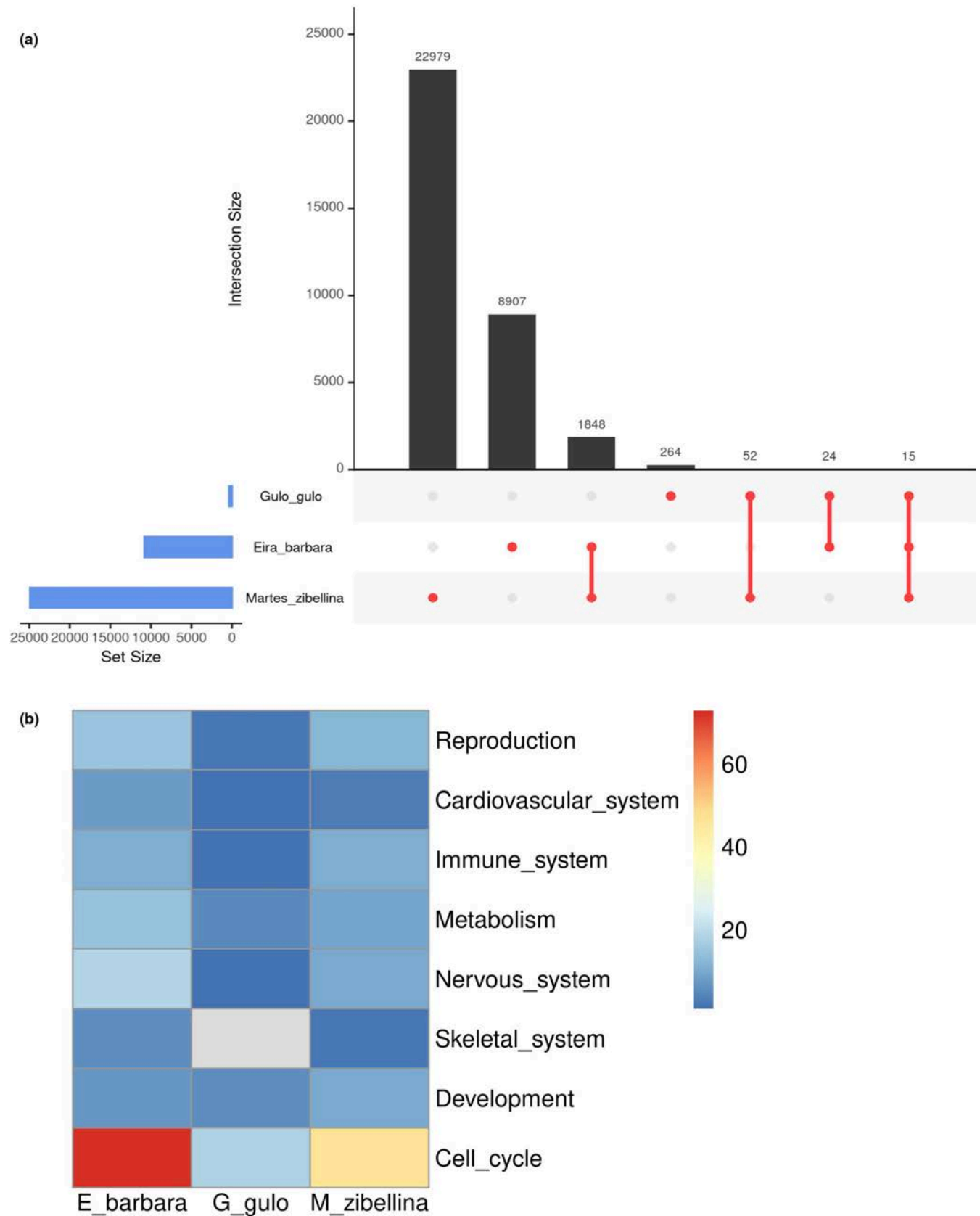


FIGURE 5 Structural variants detected in gulonine species. (a) Shared and species-specific structural variants detected in wolverine (*Gulo gulo*), tayra (*Eira barbara*) and sable (*Martes zibellina*). (b) Functional groups of genes affected by species-specific structural variants in three gulonine species (SV types: DEL, DUP, INS). Heatmap scale represents the number of genes. Heatmap cells in grey indicate no observations for a given variable

work (Ekblom et al., 2018), we observed low recent effective population size in wolverines, and a concomitant low genome-wide heterozygosity.

Contrary to the findings by Weissensteiner et al. (2020) in corvids, we did not observe a positive relationship between variation at the nucleotide level (heterozygous SNPs) and variation at the structural level (heterozygous SVs) within the gulonine species. The tayra displayed the highest nucleotide diversity (1.89 SNPs per kbp), but only the second highest amount of heterozygous SVs (2,543, 23.6% of the total SVs, Figure S9). The sable had the second highest nucleotide diversity (1.44 SNPs per kbp), but the highest number of heterozygous SVs (14,823, 59.5% of the total). The wolverine displayed the lowest variation for both (0.28 SNPs per kbp, and 153 or 43.1% heterozygous SVs in total). It is known that SV calling using short-read data can miss a large number of SVs (Ebert et al., 2021). The fact that we did not detect a positive correlation between variation at the nucleotide and structural level, as would be expected if diversity of SNPs and SVs are correlated with population size, may result from our SV analysis relying on short-read data only. Weissensteiner et al. (2020), who did report a positive correlation between SNP and SV diversity, performed long-read-based SV typing.

Assessment of variation among genome assemblies of closely related species is also strongly impacted by the contiguity and completeness of the analysed assemblies (Gurevich et al., 2013; Totikov et al., 2021). This needs to be accounted for when examining variation among discontinuous genome assemblies. Here, the low contiguity of the wolverine assembly has probably impacted the number of PSGs and gene family expansions/contractions detected. Additionally, the use of multiple, short insert size libraries sequenced at low coverage for the wolverine (Ekblom et al., 2018) has probably resulted in decreased SV detectability. We would thus argue that future comparative genomics studies of Guloninae may benefit from improving the contiguity and completeness of the wolverine genome.

4.1 | Adaptive genomic variation

Among positively selected genes, gene family expansions and coding regions impacted by SVs, we found numerous candidate loci that may be associated with species-specific traits in Guloninae.

For example, the tayra has an atypical reproductive strategy among Guloninae, namely aseasonal breeding. Among the 23 genes associated with reproduction in tayra (Figure 6a), 10 were pregnancy-related (two PSGs, two GF, six SVs), which may be linked to this species' reproductive strategy. In the hypercarnivorous wolverine, we did not observe any candidate loci associated with carbohydrate metabolism ("omnivorous diet," Figure 6b), while several were detected in the omnivorous tayra (one PSG, two GF, three SVs). However, we did observe seven genes (six PSGs, one GF) associated with body condition in wolverines, which may reflect this species' adaptive response to unfavourable environmental conditions in its circumpolar habitat. We discuss candidate loci in the context of the three species' ecology in more detail below.

We note that in two analyses (PSGs and gene family evolution), we only considered variation in single-copy orthologues, not in the entire gene repertoire of these species. Thus, our results are probably only an incomplete reflection of the genes involved in these traits.

4.2 | Seasonal breeders in the north palaeartic: wolverine and sable

Obligate embryonic diapause or delayed implantation of the blastocyst is a widespread reproductive strategy among seasonally breeding mustelids and other carnivorans. For example, wolverines and sables delay implantation for several months (Mead, 1981; Svishcheva & Kashtanov, 2011). Conspecific encounters are rare (Inman et al., 2012; Kashtanov et al., 2015), and thus induced ovulation during encounters is advantageous (Larivière & Ferguson, 2003). Previous studies in mink showed that increased levels of vascular endothelial growth factors (VEGFs) and their receptors correlate with the implantation process (Lopes et al., 2003). *VEGFC*, primarily associated with angiogenesis and regulation of permeability of blood vessels during embryogenesis, was positively selected in sable, suggesting its possible involvement in embryo implantation regulation in this species. In wolverine, we detected signals of positive selection in *ANAPC7*, a gene involved in progesterone-mediated oocyte maturation and release from cell arrest prior to fertilization (Papin et al., 2004; Reis et al., 2006), that may have a role in increasing progesterone secretion and renewed embryonic development, as observed in skunks and mink (Mead, 1989).

Changes in testicular activity and spermatogenesis also correlate strongly with season (mink: Blottner et al., 2006, lynx: Jewgenow et al., 2006). In the wolverine, positively selected genes involved in spermatogenesis included *IZUMO3*, essential for gamete fusion during fertilization (Ellerman et al., 2009), and *RNF212B*, critical for crossing over in gametes (Reynolds et al., 2013). In sable, candidate genes involved in spermatogenesis were *UBQLNL* and *SEPT12*, with the latter also being duplicated. Furthermore, *MEIKIN* and *CDK2*, both involved in meiosis, show signals of positive selection and rapid evolution through gene family expansion, respectively.

Seasonal breeding in many mammals is largely under photoperiod regulation, suggesting that the circadian system plays an important role in this reproductive strategy. *FBXL3*, associated with maintenance of circadian clock oscillation in mammals (Shi et al., 2013; Siepka et al., 2007), is duplicated in sable.

4.3 | Aseasonal breeder in the neotropic: tayra

In tropical regions, reproduction does not depend on season as environmental conditions are relatively stable throughout the year (McNutt et al., 2019). Tayras are aseasonal breeders with multiple oestrous cycles per year (Proulx & Aubry, 2017) and do not exhibit embryonic diapause (Poglayen-Neuwall et al., 1989). The tayra

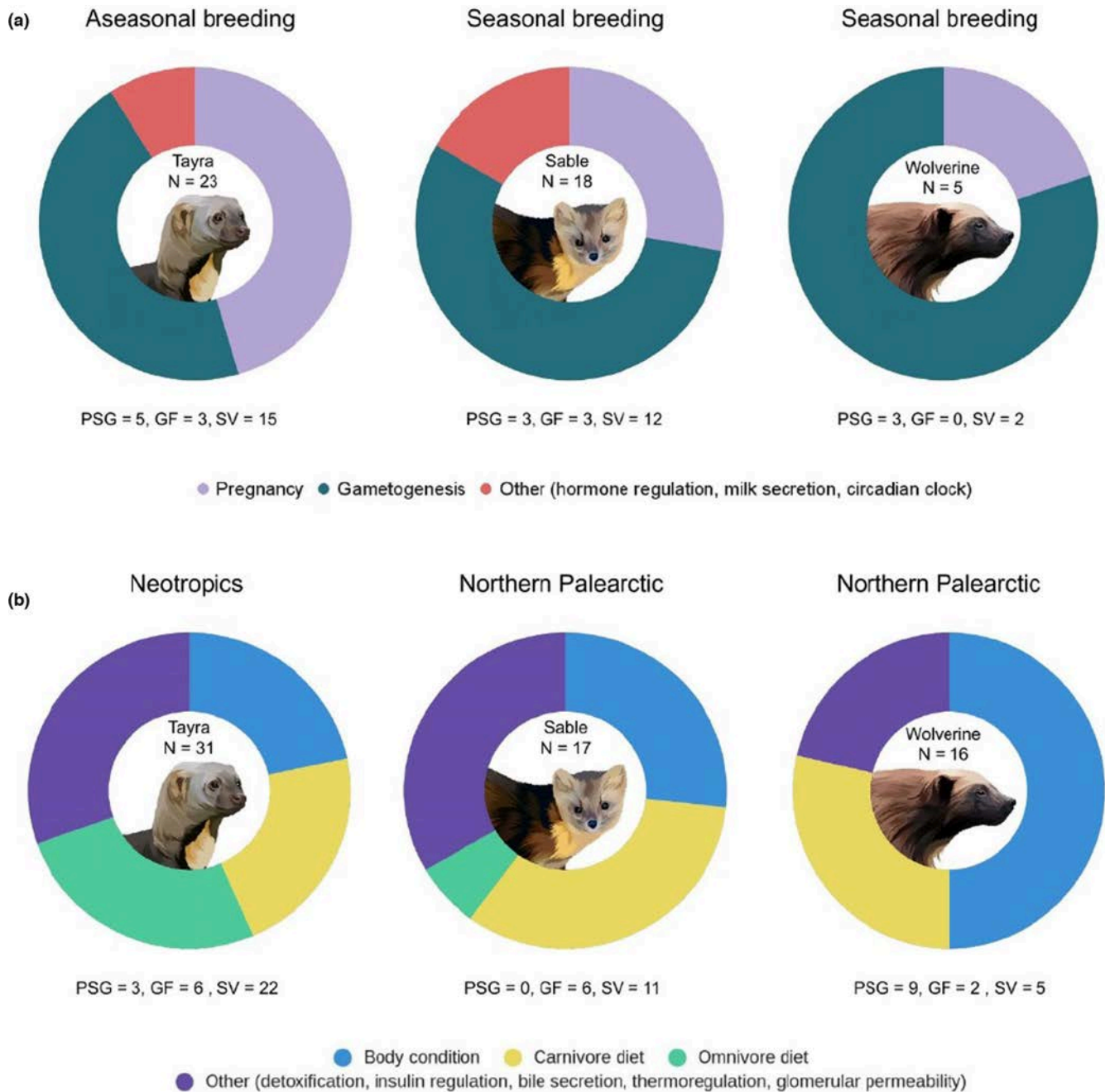


FIGURE 6 Summary of functional categories of (a) reproduction and (b) metabolism-related genes derived from analyses of positively selected genes (PSG), gene family expansion (GF) and structural variation (SV). N represents the total number of detected genes. Vector graphics of species are created based on royalty-free images (Source: Shutterstock)

represents the most basal taxon of the Guloninae and is an exception regarding its reproductive strategy.

We detected candidate genes in tayra that are related to pregnancy, and thus potentially to aseasonal breeding in this species. *ETV2* and *MUC15*, both under positive selection, are associated with placental and embryo development, and regulation of implantation (Poon et al., 2014; Singh et al., 2019). *SLC38A2*, which is duplicated in tayra, is upregulated in the pre-implantation period (Forde et al., 2014) and during late gestation, maintaining fetal growth when maternal growth is restricted by undernutrition

(Coan et al., 2010). *HSD17B10*, duplicated in tayra, is highly expressed in fetal and maternal livers, maintains pregnancy and provides protection against excitotoxicity (Hill et al., 2011). *RBP2*, also duplicated in tayra, regulates retinoids during oogenesis and embryogenesis, and positively impacts oocyte maturation in mice, cattle, pigs and sheep (Brown et al., 2003; Harney et al., 1993). Furthermore, six genes involved in placental development, implantation and embryogenesis (*HSF1*, *RSPO2*, *NLRP1*, *NLRP8*, *RNASEH2B* and *DNMT3A*) have been affected by partial deletions or duplications in tayra, raising the possibility of further

modification (e.g., functional or regulatory) of these pregnancy-related genes. Partial deletions or duplications overlapping one or more exons in protein-coding genes impact RNA splicing patterns and subsequently protein functions. They lead to production of protein isoforms with different structural and functional properties, or modulate mRNA translational efficiency, or lastly, lead to pseudogenization of a gene (Wang et al., 2015; Xing & Lee, 2006).

4.4 | Resource availability in the northern Palearctic: wolverine and sable

Surviving the winter is challenging for nonhibernating northern palearctic species and requires specific mechanisms to cope with adverse temperatures and food scarcity. Thus, efficient storage and mobilization of fat is very important in low-productivity environments (Inman et al., 2012). Three PSGs detected in the wolverine are involved in formation of adipose tissue, *MTPAP*, *OIP5* and *ZADH2* (Han et al., 2012; Inoue et al., 2014; Yu et al., 2013) and selective fatty acid mobilization stimulated by fasting periods (Inman et al., 2012; Krebs et al., 2004), as observed in mink (Nieminen et al., 2006) and raccoon dog (Mustonen et al., 2007).

One of the responses to prolonged periods of nutrient deprivation and extreme environmental conditions is suppressed bone resorption and formation (Lennox & Goodship, 2008; McGee-Lawrence et al., 2015). While the control of autophagy is important for the survival of blastocysts during delayed implantation (Lee et al., 2011; Lim & Song, 2014), it is also very important in maintaining bone homeostasis (DeSelm et al., 2011; Montaseri et al., 2020). It is thus of note that genes involved in bone mass regulation, resorption (*PPP1R18*, *TMEM38B*) and autophagy (*NBR1*) are under positive selection in wolverines. We also detected a duplication of the muscle growth-regulating gene *MTM1*. While a lack of *MTM1* will lead to muscle hypotrophy through unbalanced autophagy in humans and mice (Al-Qusairi et al., 2013), a gene duplication may facilitate muscle growth or counteract muscle reduction.

In sable, fatty acids are mobilized from fat deposits (Nieminen & Mustonen, 2007), and we observed duplications of *ASB6*, *SLC25A10*, *RNF186* and *BORCS6*, which regulate fat storage and response to nutrient availability (Mizuarai et al., 2005; Okamoto et al., 2020; Schweitzer et al., 2015; Wilcox et al., 2004). The partial deletions we detected in *APOD*, *PDHB*, *LDLR* and *CERS5*, all associated with lipoprotein metabolism (Carmo et al., 2009; Gosejacob et al., 2016; Serão et al., 2011; Tavori et al., 2015), indicate modification of genes in pathways associated with energy conservation in this species.

We observed partial deletions in *DNAJC7* in both sable and wolverine, indicating independent modification of this thermoregulation gene (Sonna et al., 2002) in gulonines inhabiting colder environments. Another gene in which we detected independent partial deletions in sable and wolverine is *GLUD1*, which regulates insulin homeostasis (Fahien & Macdonald, 2011). Modification of this gene

may impact “adaptive fasting” in these species, an adaptation to prolonged periods of nutrient deprivation observed in several carnivores (Martinez & Ortiz, 2017; Viscarra et al., 2013).

Sables are famous for their dense fur, and we observed two duplications that may be linked to this trait: *TCHHL1*, involved in hair morphogenesis (Wu et al., 2011), and *CDC42*, required for differentiation of hair follicle progenitor cells (Wu et al., 2006).

4.5 | Resource availability in the Neotropics: tayra

Tayras exploit diverse food sources and experience relatively stable resource availability all year round (Zhou et al., 2011). Shifts in dietary preferences have been linked to positive selection in single genes (Kosiol et al., 2008) and to copy number variation in metabolism-related gene families in mammals (Hecker et al., 2019; Rinker et al., 2019). In tayra, we found candidate genes associated with fructose metabolism, which may be associated with the addition of fruits and honey to this species' diet. For example, *UOX*, involved in regulation of purine metabolism and conversion of fructose to fat (Johnson & Andrews, 2010), and *DERA*, associated with catabolic processes, were both under positive selection in tayra, and part of significantly overrepresented GO categories in this species.

High rates of lineage-specific variation in gene family size, especially those families involved in immune response or detoxification of xenobiotic molecules (Thomas, 2007), are probably associated with environmental changes during speciation (Lynch & Conery, 2000; Zhang, 2003). We found a duplication of *N6AMT1*, which is associated with conversion of an arsenic metabolite, monomethylarsonous acid, to the less toxic dimethylarsonic acid (Ren et al., 2011). Arsenic with geothermal origins (e.g., volcanic activity) is common in Latin America, where it represents a severe threat to public health and the livelihoods of millions of people, with chronic exposure leading to various diseases (Morales-Simfors et al., 2020; Zhang et al., 2015). This duplication may represent an adaptation of tayra to this xenobiotic compound.

Finally, we also found candidate genes associated with lens fibre formation and retinal vascularization in tayra, including gene expansions of *ANKRD13A* (Avellino et al., 2013) and *RBP2* (D'Ambrosio et al., 2011). It has been suggested that tayras detect prey primarily by smell, as their eyesight has been described as being relatively poor (Defler, 1980; Wilson & Mittermeier, 2009). However, this has not been experimentally tested, and it is somewhat contradictory to the observed behaviour of caching of unripe but mature stages of both native and non-native fruits (Soley & Alvarado-Díaz, 2011). As tayras inhabit (sub)tropical forests, where mammals rely on vision, alongside olfaction, to forage and avoid potentially poisonous prey (Alatalo & Mappes, 1996; Nelson et al., 2011; Webb et al., 2008), we suggest that “poor” eyesight would not be advantageous, as this would impede recognition of noxious prey displaying conspicuous coloration (Blount et al., 2009). It may thus be appropriate to revisit tayras' visual acuity.

4.6 | Conclusion and future outlook

Mustelids are a remarkable example of adaptive radiation, and we show how positively selected loci, changes in gene family size and SVs have shaped genomes in this diverse taxonomic group. We demonstrated that, in particular, the latter two sources of variation contribute many loci potentially involved in adaptive genomic evolution. In the past, these types of genomic variation were often not considered in comparative genomic studies of nonmodel species, even though they encompass more nucleotides than SNPs. To fully explore the impact of different types of genomic variants on phenotypic variation, gene expression data would be necessary. Comparative analysis of gene expression patterns and elucidating protein interactions and pathways is a domain of functional genomics, and was unfortunately outside the scope of our study.

The mustelid subfamily Guloninae includes three monotypic genera (*Eira*, *Gulo* and *Pekania*) as well as the martens (eight *Martes* species). A feasible short-term goal regarding future genomics studies of this subfamily is the generation of reference genomes for all remaining Guloninae, which is a goal of the *Martes* Genome Consortium, launched in 2018. Additionally, existing reference genomes may be improved in contiguity using, for example, Hi-C approaches (e.g., Dudchenko et al., 2017; DNAzoo.org). This will be a strong foundation for both inter- and intraspecific genomics studies of Guloninae, which includes species of conservation concern (wolverine and Nilgiri martens: “vulnerable” on the IUCN red list), species that hybridize in nature (e.g., European pine martens and sables; Davison et al., 2001; Kassal & Sidorov, 2013), and species characterized by convergent evolution of ecological adaptations (e.g., delayed implantation, seasonal moulting, sociality, scent glands).

ACKNOWLEDGMENTS

We thank Dr R. Rafael from the Felidae Wildkatzen- und Artenschutzzentrum Barnim for kindly providing the tayra sample, and Michael Hofreiter from the Adaptive Genomics group (University of Potsdam) for assistance in generating the 10x Genomics linked-read library. David Duchêne was funded by a Carlsbergfondet postdoctoral fellowship (grant no. CF18-0223). Sergei Kliver, Andrey Tomarovsky and Azamat Totikov were funded by the Russian Foundation for Basic Research (grant no. 20-04-00808). Azamat Totikov and Andrey Tomarovsky were additionally funded by JetBrains Research.

AUTHOR CONTRIBUTIONS

L.D. conceived the study, conducted genome assembly, gene family evolution and structural variation analysis, interpreted the data, and wrote the manuscript. A.B. and S.J. conducted positive selection analysis, and interpreted the data. P.D. carried out demographic history reconstruction and data interpretation. D.A.D. designed and performed phylogenomic analysis and molecular dating, and wrote the manuscript. J.H.G. conducted repetitive elements analysis and interpreted the data. S.K., A.T. and A.T.

performed reference-based scaffolding, alignment to pseudochromosome assemblies, sex verification and nucleotide diversity assessment, interpreted the data and wrote the manuscript. K.P.K. advised on comparative genomic analyses, and edited the manuscript. D.M. supervised and performed gene family evolution analysis, and advised on other comparative genomic analyses. M.P. conducted whole genome sequencing and sequencing data interpretation. J.F. and D.W.F. conceived the study, interpreted the data, and wrote and edited the manuscript. All authors have reviewed and approved the manuscript.

OPEN RESEARCH BADGES



This article has been awarded Open Data Badges. All materials and data are publicly accessible via the Open Science Framework at <https://doi.org/10.5061/dryad.xgxd254hz> and <https://doi.org/10.5281/zenodo.6320689>.

DATA AVAILABILITY STATEMENT

The tayra principal genome assembly and raw sequencing reads have been deposited under the NCBI BioProject ID: [PRJNA732553](https://doi.org/10.5061/dryad.xgxd254hz), and the alternative assembly can be accessed under BioProject ID: [PRJNA732552](https://doi.org/10.5061/dryad.xgxd254hz). The vcf files with variant calling data are available at Dryad (<https://doi.org/10.5061/dryad.xgxd254hz>), and scripts at Zenodo (<https://doi.org/10.5281/zenodo.6320689>). All data generated or analysed during this study are included in this published article and its Supporting Information files.

ORCID

Lorena Derežanin <https://orcid.org/0000-0002-0707-573X>

Pavel Dobrynin <https://orcid.org/0000-0001-6995-5620>

REFERENCES

- Abduriyim, S., Nishita, Y., Abramov, A. V., Solovyev, V. A., Saveljev, A. P., Kosintsev, P. A., Kryukov, A. P., Raichev, E., Väinölä, R., Kaneko, Y., & Masuda, R. (2019). Variation in pancreatic amylase gene copy number among Eurasian badgers (Carnivora, Mustelidae, Meles) and its relationship to diet. *Journal of Zoology*, 308(1), 28–36.
- Alatalo, R. V., & Mappes, J. (1996). Tracking the evolution of warning signals. *Nature*, 382(6593), 708–710.
- Alonge, M., Soyk, S., Ramakrishnan, S., Wang, X., Goodwin, S., Sedlazeck, F. J., Lippman, Z. B., & Schatz, M. C. (2019). RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biology*, 20(1), 224. <https://doi.org/10.1186/s13059-019-1829-6>
- Al-Qusairi, L., Prokic, I., Amoasii, L., Kretz, C., Messaddeq, N., Mandel, J.-L., & Laporte, J. (2013). Lack of myotubularin (MTM1) leads to muscle hypotrophy through unbalanced regulation of the autophagy and ubiquitin-proteasome pathways. *FASEB Journal*, 27(8), 3384–3394. <https://doi.org/10.1096/fj.12-220947>
- Angelis, K., & Dos Reis, M. (2015). The impact of ancestral population size and incomplete lineage sorting on Bayesian estimation of species divergence times. *Current Zoology*, 61(5), 874–885.
- Anisimova, M., & Gascuel, O. (2006). Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic Biology*, 55(4), 539–552.

- Armstrong, E. E., Taylor, R. W., Prost, S., Blinston, P., van der Meer, E., Madzikanda, H., Mufute, O., Mandisodza-Chikerema, R., Stuelpnagel, J., Sillero-Zubiri, C., & Petrov, D. (2019). Cost-effective assembly of the African wild dog (*Lycaon pictus*) genome using linked reads. *GigaScience*, 8(2), <https://doi.org/10.1093/gigascience/giy124>
- Avellino, R., Carrella, S., Pirozzi, M., Risolino, M., Salierno, F. G., Franco, P., Stoppelli, P., Verde, P., Banfi, S., & Conte, I. (2013). miR-204 targeting of Ankrd13A controls both mesenchymal neural crest and lens cell migration. *PLoS One*, 8(4), e61099. <https://doi.org/10.1371/journal.pone.0061099>
- Barnett, R., Westbury, M. V., Sandoval-Velasco, M., Vieira, F. G., Jeon, S., Zazula, G., Martin, M. D., Ho, S. Y. W., Mather, N., Gopalakrishnan, S., Ramos-Madrugal, J., de Manuel, M., Zepeda-Mendoza, M. L., Antunes, A., Baez, A. C., De Cahsan, B., Larson, G., O'Brien, S. J., Eizirik, E., ... Gilbert, M. T. P. (2020). Genomic adaptations and evolutionary history of the extinct scimitar-toothed cat, *Homotherium latidens*. *Current Biology*, 30(24), 5018–5025.e5. <https://doi.org/10.1016/j.cub.2020.09.051>
- Beichman, A. C., Koepfli, K.-P., Li, G., Murphy, W., Dobrynin, P., Kilver, S., Tinker, M. T., Murray, M. J., Johnson, J., Lindblad-Toh, K., Karlsson, E. K., Lohmueller, K. E., & Wayne, R. K. (2019). Aquatic adaptation and depleted diversity: a deep dive into the genomes of the sea otter and giant otter. *Molecular Biology and Evolution*, 36(12), 2631–2655. <https://doi.org/10.1093/molbev/msz101>
- Blottner, S., Schön, J., & Jewgenow, K. (2006). Seasonally activated spermatogenesis is correlated with increased testicular production of testosterone and epidermal growth factor in mink (*Mustela vison*). *Theriogenology*, 66(6–7), 1593–1598. <https://doi.org/10.1016/j.theriogenology.2006.01.041>
- Blount, J. D., Speed, M. P., Ruxton, G. D., & Stephens, P. A. (2009). Warning displays may function as honest signals of toxicity. *Proceedings Biological Sciences*, 276(1658), 871–877.
- Broad Institute (2019). Picard Toolkit. GitHub Repository. <https://broadinstitute.github.io/picard/>
- Brown, J. A., Eberhardt, D. M., Schrick, F. N., Roberts, M. P., & Godkin, J. D. (2003). Expression of retinol-binding protein and cellular retinol-binding protein in the bovine ovary. *Molecular Reproduction and Development*, 64(3), 261–269. <https://doi.org/10.1002/mrd.10225>
- Buckley, R. M., Davis, B. W., Brashear, W. A., Farias, F. H. G., Kuroki, K., Graves, T., Hillier, L. W., Kremitzki, M., Li, G., Middleton, R. P., Minx, P., Tomlinson, C., Lyons, L. A., Murphy, W. J., & Warren, W. C. (2020). A new domestic cat genome assembly based on long sequence reads empowers feline genomic medicine and identifies a novel gene for dwarfism. *PLoS Genetics*, 16(10), e1008926. <https://doi.org/10.1371/journal.pgen.1008926>
- Cahill, J. A., Soares, A. E. R., Green, R. E., & Shapiro, B. (2016). Inferring species divergence times using pairwise sequential Markovian coalescent modelling and low-coverage genomic data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1699), <https://doi.org/10.1098/rstb.2015.0138>
- Carmo, S. D., Fournier, D., Mounier, C., & Rassart, E. (2009). Human apolipoprotein D overexpression in transgenic mice induces insulin resistance and alters lipid metabolism. *American Journal of Physiology-Endocrinology and Metabolism*, 296(4), E802–E811. <https://doi.org/10.1152/ajpendo.90725.2008>
- Catanach, A., Crowhurst, R., Deng, C., David, C., Bernatchez, L., & Wellenreuther, M. (2019). The genomic pool of standing structural variation outnumbers single nucleotide polymorphism by threefold in the marine teleost *Chrysophrys auratus*. *Molecular Ecology*, 28(6), 1210–1223.
- Cavagna, P., Menotti, A., & Stanyon, R. (2000). Genomic homology of the domestic ferret with cats and humans. *Mammalian Genome: Official Journal of the International Mammalian Genome Society*, 11(10), 866–870. <https://doi.org/10.1007/s003350010172>
- Chang, J.-M., Di Tommaso, P., & Notredame, C. (2014). TCS: a new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. *Molecular Biology and Evolution*, 31(6), 1625–1637. <https://doi.org/10.1093/molbev/msu117>
- Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A. J., Kruglyak, S., & Saunders, C. T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, 32(8), 1220–1222. <https://doi.org/10.1093/bioinformatics/btv710>
- Chiang, C., Layer, R. M., Faust, G. G., Lindberg, M. R., Rose, D. B., Garrison, E. P., Marth, G. T., Quinlan, A. R., & Hall, I. M. (2015). SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nature Methods*, 12(10), 966–968. <https://doi.org/10.1038/nmeth.3505>
- Chiang, C., Scott, A. J., Davis, J. R., Tsang, E. K., Li, X., Kim, Y., Hadzic, T., Damani, F. N., Ganel, L., Montgomery, S. B., Battle, A., Conrad, D. F., & Hall, I. M. (2017). The impact of structural variation on human gene expression. *Nature Genetics*, 49(5), 692–699. <https://doi.org/10.1038/ng.3834>
- Coan, P. M., Vaughan, O. R., Sekita, Y., Finn, S. L., Burton, G. J., Constancia, M., & Fowden, A. L. (2010). Adaptations in placental phenotype support fetal growth during undernutrition of pregnant mice. *The Journal of Physiology*, 588(Pt 3), 527–538. <https://doi.org/10.1113/jphysiol.2009.181214>
- Copeland, J. P., & Kucera, T. E. (1997). Wolverine (*Gulo gulo*). In: J. E. Harris, & C. V. Ogan, eds. *Mesocarnivores of Northern California: Biology, management, and survey techniques, workshop manual*. Humboldt State Univ. and the Wildlife Society, California North Coast Chapter, 127 p
- D'Ambrosio, D. N., Clugston, R. D., & Blaner, W. S. (2011). Vitamin A metabolism: An update. *Nutrients*, 3(1), 63–103. <https://doi.org/10.3390/nu3010063>
- Davison, A., Birks, J. D., Brookes, R. C., Messenger, J. E., & Griffiths, H. I. (2001). Mitochondrial phylogeography and population history of pine martens *Martes martes* compared with polecats *Mustela putorius*. *Molecular Ecology*, 10(10), 2479–2488. <https://doi.org/10.1046/j.1365-294X.2001.01381.x>
- Defler, T. R. (1980). Notes on interactions between the Tayra (*Eira barbara*) and the white-fronted capuchin (*Cebus albifrons*). *Journal of Mammalogy*, 61(1), 156. <https://doi.org/10.2307/1379979>
- DeSelm, C. J., Miller, B. C., Zou, W., Beatty, W. L., van Meel, E., Takahata, Y., Klumperman, J., Tooz, S. A., Teitelbaum, S. L., & Virgin, H. W. (2011). Autophagy proteins regulate the secretory component of osteoclastic bone resorption. *Developmental Cell*, 21(5), 966–974. <https://doi.org/10.1016/j.devcel.2011.08.016>
- Dobrynin, P., Liu, S., Tamazian, G., Xiong, Z., Yurchenko, A. A., Krashennikova, K., Kliver, S., Schmidt-Küntzel, A., Koepfli, K.-P., Johnson, W., Kuderna, L. F. K., García-Pérez, R., Manuel, M. D., Godinez, R., Komissarov, A., Makunin, A., Brukhin, V., Qiu, W., Zhou, L., ... O'Brien, S. J. (2015). Genomic legacy of the African cheetah, *Acinonyx jubatus*. *Genome Biology*, 16, 277. <https://doi.org/10.1186/s13059-015-0837-4>
- 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.
- Dudchenko, O., Shamim, M. S., Batra, S. S., Durand, N. C., Musial, N. T., Mostofa, R., Pham, M., St Hilaire, B. G., Yao, W., Stamenova, E., Hoeger, M., Nyquist, S. K., Korchina, V., Pletch, K., Flanagan, J. P., Tomaszewicz, A., McAloose, D., Estrada, C. P., Novak, B. J., Aiden, E. L. (2018). The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under \$1000 (p. 254797). <https://doi.org/10.1101/254797>
- Eklblom, R., Brechlin, B., Persson, J., Smeds, L., Johansson, M., Magnusson, J., Flagstad, Ø., & Ellegren, H. (2018). Genome sequencing and

- conservation genomics in the Scandinavian wolverine population. *Conservation Biology: the Journal of the Society for Conservation Biology*, 32(6), 1301–1312. <https://doi.org/10.1111/cobi.13157>
- Ellerman, D. A., Pei, J., Gupta, S., Snell, W. J., Myles, D., & Primakoff, P. (2009). Izumo is part of a multiprotein family whose members form large complexes on mammalian sperm. *Molecular Reproduction and Development*, 76(12), 1188–1199. <https://doi.org/10.1002/mrd.21092>
- Etherington, G. J., Heavens, D., Baker, D., Lister, A., McNelly, R., Garcia, G., Clavijo, B., Macaulay, I., Haerty, W., & Di Palma, F. (2020). Sequencing smart: De novo sequencing and assembly approaches for a non-model mammal. *GigaScience*, 9(5), <https://doi.org/10.1093/gigascience/giaa045>
- Fahien, L. A., & Macdonald, M. J. (2011). The complex mechanism of glutamate dehydrogenase in insulin secretion. *Diabetes*, 60(10), 2450–2454. <https://doi.org/10.2337/db10-1150>
- Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America*, 117(17), 9451–9457. <https://doi.org/10.1073/pnas.1921046117>
- Forde, N., Simintiras, C. A., Sturmey, R., Mamo, S., Kelly, A. K., Spencer, T. E., Bazer, F. W., & Lonergan, P. (2014). Amino acids in the uterine luminal fluid reflects the temporal changes in transporter expression in the endometrium and conceptus during early pregnancy in cattle. *PLoS One*, 9(6), e100010. <https://doi.org/10.1371/journal.pone.0100010>
- Frith, M. C., & Kawaguchi, R. (2015). Split-alignment of genomes finds orthologies more accurately. *Genome Biology*, 16, 106. <https://doi.org/10.1186/s13059-015-0670-9>
- Ge, S. X., Jung, D., & Yao, R. (2020). ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics*, 36(8), 2628–2629. <https://doi.org/10.1093/bioinformatics/btz931>
- Gosejacob, D., Jäger, P. S., Vom Dorp, K., Frejno, M., Carstensen, A. C., Köhnke, M., Degen, J., Dörmann, P., & Hoch, M. (2016). Ceramide synthase 5 is essential to maintain C16:0-ceramide pools and contributes to the development of diet-induced obesity. *Journal of Biological Chemistry*, 291(13), 6989–7003. <https://doi.org/10.1074/jbc.M115.691212>
- Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUASt: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8), 1072–1075. <https://doi.org/10.1093/bioinformatics/btt086>
- Han, M. V., Thomas, G. W. C., Lugo-Martinez, J., & Hahn, M. W. (2013). Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Molecular Biology and Evolution*, 30(8), 1987–1997. <https://doi.org/10.1093/molbev/mst100>
- Han, X., Jiang, T., Yu, L., Zeng, C., Fan, B., & Liu, B. (2012). Molecular characterization of the porcine MTPAP gene associated with meat quality traits: chromosome localization, expression distribution, and transcriptional regulation. *Molecular and Cellular Biochemistry*, 364(1–2), 173–180. <https://doi.org/10.1007/s11010-011-1216-4>
- Harney, J. P., Ott, T. L., Geisert, R. D., & Bazer, F. W. (1993). Retinol-binding protein gene expression in cyclic and pregnant endometrium of pigs, sheep, and cattle. *Biology of Reproduction*, 49(5), 1066–1073.
- Hecker, N., Sharma, V., & Hiller, M. (2019). Convergent gene losses illuminate metabolic and physiological changes in herbivores and carnivores. *Proceedings of the National Academy of Sciences of the United States of America*, 116(8), 3036–3041. <https://doi.org/10.1073/pnas.1818504116>
- Heldstab, S. A., Müller, D. W. H., Graber, S. M., Bingaman Lackey, L., Rensch, E., Hatt, J.-M., Zerbe, P., & Clauss, M. (2018). Geographical origin, delayed implantation, and induced ovulation explain reproductive seasonality in the Carnivora. *Journal of Biological Rhythms*, 33(4), 402–419. <https://doi.org/10.1177/0748730418773620>
- Hill, M., Pařízek, A., Kancheva, R., & Jirásek, J. E. (2011). Reduced progesterone metabolites in human late pregnancy. *Physiological Research*, 60(2), 225–241. <https://doi.org/10.33549/physiolres.932077>
- Hron, T., Elleder, D., & Gifford, R. J. (2019). Deltaretroviruses have circulated since at least the Paleogene and infected a broad range of mammalian species. *Retrovirology*, 16(1), 33. <https://doi.org/10.1186/s12977-019-0495-9>
- Inman, R. M., Magoun, A. J., Persson, J., & Mattisson, J. (2012). The wolverine's niche: Linking reproductive chronology, caching, competition, and climate. *Journal of Mammalogy*, 93(3), 634–644. <https://doi.org/10.1644/11-MAMM-A-319.1>
- Inoue, K., Maeda, N., Mori, T., Sekimoto, R., Tushima, Y., Matsuda, K., Yamaoka, M., Suganami, T., Nishizawa, H., Ogawa, Y., Funahashi, T., & Shimomura, I. (2014). Possible involvement of Opa-interacting protein 5 in adipose proliferation and obesity. *PLoS One*, 9(2), e87661. <https://doi.org/10.1371/journal.pone.0087661>
- Jeffares, D. C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., Balloux, F., Dessimoz, C., Bähler, J., & Sedlazeck, F. J. (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature Communications*, 8, 14061. <https://doi.org/10.1038/ncomms14061>
- Jewgenow, K., Goeritz, F., Neubauer, K., Fickel, J., & Naidenko, S. V. (2006). Characterization of reproductive activity in captive male Eurasian lynx (*Lynx lynx*). *European Journal of Wildlife Research*, 52(1), 34–38. <https://doi.org/10.1007/s10344-005-0002-6>
- Johnson, R. J., & Andrews, P. (2010). Fructose, uricase, and the back-to-Africa hypothesis. *Evolutionary Anthropology*, 19(6), 250–257. <https://doi.org/10.1002/evan.20266>
- Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermini, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, 14(6), 587–589. <https://doi.org/10.1038/nmeth.4285>
- Kashtanov, S. N., Svischeva, G. R., Pishchulina, S. L., Lazebny, O. E., Meshchersky, I. G., Simakin, L. V., & Rozhnov, V. V. (2015). Geographical structure of the sable (*Martes zibellina* L.) gene pool on the basis of microsatellite loci analysis. *Russian Journal of Genetics*, 51(1), 69–79.
- Kassal, B. Y., & Sidorov, G. N. (2013). Distribution of the sable (*Martes zibellina*) and the pine marten (*Martes martes*) in Omsk Oblast and biogeographic effects of their hybridization. *Russian Journal of Biological Invasions*, 4(2), 105–115. <https://doi.org/10.1134/S2075111713020070>
- Kim, B.-M., Lee, Y. J., Kim, J.-H., Kim, J.-H., Kang, S., Jo, E., Lee, S. J., Lee, J. H., Chi, Y. M., & Park, H. (2020). The genome assembly and annotation of the southern elephant seal *Mirounga leonina*. *Genes*, 11(2), <https://doi.org/10.3390/genes11020160>
- Koepfli, K.-P., Deere, K. A., Slater, G. J., Begg, C., Begg, K., Grassman, L., Lucherini, M., Veron, G., & Wayne, R. K. (2008). Multigene phylogeny of the Mustelidae: Resolving relationships, tempo and biogeographic history of a mammalian adaptive radiation. *BMC Biology*, 6, 10. <https://doi.org/10.1186/1741-7007-6-10>
- Kosiol, C., Vinar, T., da Fonseca, R. R., Hubisz, M. J., Bustamante, C. D., Nielsen, R., & Siepel, A. (2008). Patterns of positive selection in six mammalian genomes. *PLoS Genetics*, 4(8), e1000144. <https://doi.org/10.1371/journal.pgen.1000144>
- Krebs, J., Lofroth, E., Copeland, J., Banci, V., Cooley, D., Golden, H., Magoun, A., Mulders, R., & Shults, B. (2004). Synthesis of survival rates and causes of mortality in North American wolverines. *Journal of Wildlife Management*, 68(3), 493–502.
- Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F. A., & Zdobnov, E. M. (2019). OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes

- for evolutionary and functional annotations of orthologs. *Nucleic Acids Research*, 47(D1), D807–D811. <https://doi.org/10.1093/nar/gky1053>
- Kronenberg, Z. N., Osborne, E. J., Cone, K. R., Kennedy, B. J., Domyan, E. T., Shapiro, M. D., Elde, N. C., & Yandell, M. (2015). Wham: Identifying structural variants of biological consequence. *PLoS Computational Biology*, 11(12), e1004572. <https://doi.org/10.1371/journal.pcbi.1004572>
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie2. *Nature Methods*, 9(4), 357–359. <https://doi.org/10.1038/nmeth.1923>
- Larivière, S., & Ferguson, S. H. (2003). Evolution of induced ovulation in North American Carnivores. *Journal of Mammalogy*, 84(3), 937–947. <https://doi.org/10.1644/BME-003>
- Law, C. J., Slater, G. J., & Mehta, R. S. (2018). Lineage diversity and size disparity in Musteloidea: Testing patterns of adaptive radiation using molecular and fossil-based methods. *Systematic Biology*, 67(1), 127–144. <https://doi.org/10.1093/sysbio/syx047>
- Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: A probabilistic framework for structural variant discovery. *Genome Biology*, 15(6), R84. <https://doi.org/10.1186/gb-2014-15-6-r84>
- Lee, J.-E., Oh, H.-A., Song, H., Jun, J. H., Roh, C.-R., Xie, H., Dey, S. K., & Lim, H. J. (2011). Autophagy regulates embryonic survival during delayed implantation. *Endocrinology*, 152(5), 2067–2075. <https://doi.org/10.1210/en.2010-1456>
- Lennox, A. R., & Goodship, A. E. (2008). Polar bears (*Ursus maritimus*), the most evolutionarily advanced hibernators, avoid significant bone loss during hibernation. *Comparative Biochemistry and Physiology. Part A, Molecular & Integrative Physiology*, 149(2), 203–208. <https://doi.org/10.1016/j.cbpa.2007.11.012>
- 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), <https://doi.org/10.1093/gigascience/giz086>
- Li, B., Wolsan, M., Wu, D., Zhang, W., Xu, Y., & Zeng, Z. (2014). Mitochondrial genomes reveal the pattern and timing of marten (*Martes*), wolverine (*Gulo*), and fisher (*Pekania*) diversification. *Molecular Phylogenetics and Evolution*, 80, 156–164. <https://doi.org/10.1016/j.ympev.2014.08.002>
- Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357), 493–496.
- Lim, H. J., & Song, H. (2014). Evolving tales of autophagy in early reproductive events. *The International Journal of Developmental Biology*, 58(2–4), 183–187. <https://doi.org/10.1387/ijdb.130337hl>
- Liu, G., Zhao, C., Xu, D., Zhang, H., Monakhov, V., Shang, S., Gao, X., Sha, W., Ma, J., Zhang, W., Tang, X., Li, B., Hua, Y., Cao, X., Liu, Z., & Zhang, H. (2020). First draft genome of the sable, *Martes zibellina*. *Genome Biology and Evolution*, 12(3), 59–65. <https://doi.org/10.1093/gbe/evaa029>
- Lopes, F. L., Desmarais, J., Gevry, N. Y., Ledoux, S., & Murphy, B. D. (2003). Expression of vascular endothelial growth factor isoforms and receptors Flt-1 and KDR during the peri-implantation period in the mink, *Mustela vison*. *Biology of Reproduction*, 68(5), 1926–1933.
- Löytynoja, A. (2014). Phylogeny-aware alignment with PRANK. *Methods in Molecular Biology*, 1079, 155–170.
- Lynch, M., & Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes [Review of The evolutionary fate and consequences of duplicate genes]. *Science*, 290(5494), 1151–1155.
- Maglott, D., Ostell, J., Pruitt, K. D., & Tatusova, T. (2011). Entrez Gene: Gene-centered information at NCBI. *Nucleic Acids Research*, 39(Database issue), D52–D57. <https://doi.org/10.1093/nar/gkq1237>
- Martinez, B., & Ortiz, R. M. (2017). Thyroid hormone regulation and insulin resistance: Insights from animals naturally adapted to fasting. *Physiology*, 32(2), 141–151. <https://doi.org/10.1152/physiol.00018.2016>
- McGee-Lawrence, M., Buckendahl, P., Carpenter, C., Henriksen, K., Vaughan, M., & Donahue, S. (2015). Suppressed bone remodeling in black bears conserves energy and bone mass during hibernation. *The Journal of Experimental Biology*, 218(Pt 13), 2067–2074. <https://doi.org/10.1242/jeb.120725>
- McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl variant effect predictor. *Genome Biology*, 17(1), 122. <https://doi.org/10.1186/s13059-016-0974-4>
- McNutt, J. W., Groom, R., & Woodroffe, R. (2019). Ambient temperature provides an adaptive explanation for seasonal reproduction in a tropical mammal. *Journal of Zoology*, 309(3), 153–160. <https://doi.org/10.1111/jzo.12712>
- Mead, R. A. (1981). Delayed implantation in mustelids, with special emphasis on the spotted skunk. *Journal of Reproduction and Fertility* 29, 11–24.
- Mead, R. A. (1989). The physiology and evolution of delayed implantation in carnivores. In J. L. Gittleman (Ed.), *Carnivore behavior, ecology, and evolution* (pp. 437–464). Springer US.
- Mendes, F. K., & Hahn, M. W. (2016). Gene tree discordance causes apparent substitution rate variation. *Systematic Biology*, 65(4), 711–721. <https://doi.org/10.1093/sysbio/syw018>
- Mérot, C., Oomen, R. A., Tigano, A., & Wellenreuther, M. (2020). A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends in Ecology & Evolution*, 35(7), 561–572. <https://doi.org/10.1016/j.tree.2020.03.002>
- Minh, B. Q., Hahn, M. W., & Lanfear, R. (2020). New methods to calculate concordance factors for phylogenomic datasets. *Molecular Biology and Evolution*, 37(9), 2727–2733. <https://doi.org/10.1093/molbev/msaa106>
- Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, 37(5), 1530–1534. <https://doi.org/10.1093/molbev/msaa015>
- Miranda, I., Giska, I., Farello, L., Pimenta, J., Zimova, M., Bryk, J., Dalén, L., Mills, L. S., Zub, K., & Melo-Ferreira, J. (2021). Museomics dissects the genetic basis for adaptive seasonal colouration in the least weasel. *Molecular Biology and Evolution*, 38(10), 4388–4402. <https://doi.org/10.1093/molbev/msab177>
- Mizuarai, S., Miki, S., Araki, H., Takahashi, K., & Kotani, H. (2005). Identification of dicarboxylate carrier Slc25a10 as malate transporter in de novo fatty acid synthesis. *Journal of Biological Chemistry*, 280(37), 32434–32441. <https://doi.org/10.1074/jbc.M503152200>
- Monakhov, V. G. (2011). *Martes zibellina* (Carnivora: Mustelidae). *Mammalian Species*, 43(876), 75–86. <https://doi.org/10.1644/876.1>
- Montaseri, A., Giampietri, C., Rossi, M., Riccioli, A., Del Fattore, A., & Filippini, A. (2020). The role of autophagy in osteoclast differentiation and bone resorption function. *Biomolecules*, 10(10), <https://doi.org/10.3390/biom10101398>
- Morales-Simfors, N., Bundschuh, J., Herath, I., Inguaggiato, C., Caselli, A. T., Tapia, J., Choquehuayta, F. E. A., Armienta, M. A., Ormachea, M., Joseph, E., & López, D. L. (2020). Arsenic in Latin America: A critical overview on the geochemistry of arsenic originating from geothermal features and volcanic emissions for solving its environmental consequences. *Science of the Total Environment*, 716, 135564. <https://doi.org/10.1016/j.scitotenv.2019.135564>
- Mustonen, A.-M., Käkälä, R., Käkälä, A., Pyykönen, T., Aho, J., & Nieminen, P. (2007). Lipid metabolism in the adipose tissues of a carnivore, the raccoon dog, during prolonged fasting. *Experimental Biology and Medicine*, 232(1), 58–69.
- Mustonen, A.-M., Puukka, M., Saarela, S., Paakkonen, T., Aho, J., & Nieminen, P. (2006). Adaptations to fasting in a terrestrial mustelid, the sable (*Martes zibellina*). *Comparative Biochemistry and Physiology. Part A, Molecular & Integrative Physiology*, 144(4), 444–450. <https://doi.org/10.1016/j.cbpa.2006.03.008>

- Nelson, D. W. M., Crossland, M. R., & Shine, R. (2011). Foraging responses of predators to novel toxic prey: effects of predator learning and relative prey abundance. *Oikos*, 120(1), 152–158. <https://doi.org/10.1111/j.1600-0706.2010.18736.x>
- Nielsen, R., Bustamante, C., Clark, A. G., Glanowski, S., Sackton, T. B., Hubisz, M. J., Fledel-Alon, A., Tanenbaum, D. M., Civello, D., White, T. J., J. Sninsky, J., Adams, M. D., & Cargill, M. (2005). A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biology*, 3(6), e170. <https://doi.org/10.1371/journal.pbio.0030170>
- Nieminen, P., Käkälä, R., Pyykönen, T., & Mustonen, A.-M. (2006). Selective fatty acid mobilization in the American mink (*Mustela vison*) during food deprivation. *Comparative Biochemistry and Physiology. Part B, Biochemistry & Molecular Biology*, 145(1), 81–93. <https://doi.org/10.1016/j.cbpb.2006.06.007>
- Nieminen, P., & Mustonen, A.-M. (2007). Uniform fatty acid mobilization from anatomically distinct fat depots in the sable (*Martes zibellina*). *Lipids*, 42(7), 659–669. <https://doi.org/10.1007/s11745-007-3061-5>
- O'Brien, S. J., Graphodatsky, A. S., & Perelman, P. L. (2020). *Atlas of mammalian chromosomes*. John Wiley & Sons.
- Okamoto, T., Imaizumi, K., & Kaneko, M. (2020). The role of tissue-specific ubiquitin ligases, RNF183, RNF186, RNF182 and RNF152, in disease and biological function. *International Journal of Molecular Sciences*, 21(11), <https://doi.org/10.3390/ijms21113921>
- Olson, M. V. (1999). When less is more: gene loss as an engine of evolutionary change. *American Journal of Human Genetics*, 64(1), 18–23. <https://doi.org/10.1086/302219>
- Pacifici, M., Santini, L., Di Marco, M., Baisero, D., Francucci, L., Grottole Marasini, G., Visconti, P., & Rondinini, C. (2013). Generation length for mammals. *Nature Conservation*, 5, 89–94. <https://doi.org/10.3897/natureconservation.5.5734>
- Papin, C., Rouget, C., Lorca, T., Castro, A., & Mandart, E. (2004). XCdh1 is involved in progesterone-induced oocyte maturation. *Developmental Biology*, 272(1), 66–75. <https://doi.org/10.1016/j.ydbio.2004.04.018>
- Pasitschniak-Arts, M., & Larivière, S. (1995). *Gulo gulo*. *Mammalian Species*, 499, 1–10.
- Peng, C., Niu, L., Deng, J., Yu, J., Zhang, X., Zhou, C., Xing, J., & Li, J. (2018). Can-SINE dynamics in the giant panda and three other Caniformia genomes. *Mobile DNA*, 9, 32. <https://doi.org/10.1186/s13100-018-0137-0>
- Peng, X., Alföldi, J., Gori, K., Eisfeld, A. J., Tyler, S. R., Tisoncik-Go, J., Brawand, D., Law, G. L., Skunca, N., Hatta, M., Gasper, D. J., Kelly, S. M., Chang, J., Thomas, M. J., Johnson, J., Berlin, A. M., Lara, M., Russell, P., Swofford, R., ... Katze, M. G. (2014). The draft genome sequence of the ferret (*Mustela putorius furo*) facilitates study of human respiratory disease. *Nature Biotechnology*, 32(12), 1250–1255. <https://doi.org/10.1038/nbt.3079>
- Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: convergence diagnosis and output analysis for MCMC. *R News*, 6(1), 7–11.
- Poglayen-Neuwall, I., Durrant, B. S., Swansen, M. L., Williams, R. C., & Barnes, R. A. (1989). Estrous cycle of the tayra, *Eira barbara*. *Zoo Biology*, 8(2), 171–177. <https://doi.org/10.1002/zoo.1430080208>
- Poon, C. E., Lecce, L., Day, M. L., & Murphy, C. R. (2014). Mucin 15 is lost but mucin 13 remains in uterine luminal epithelial cells and the blastocyst at the time of implantation in the rat. *Reproduction, Fertility, and Development*, 26(3), 421–431. <https://doi.org/10.1071/RD12313>
- Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M. J., Neale, B., MacArthur, D. G., & Banks, E. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*, 201178. <https://doi.org/10.1101/201178>
- Porubsky, D., Sanders, A. D., Höps, W., Hsieh, P., Sulovari, A., Li, R., Mercuri, L., Sorensen, M., Murali, S. C., Gordon, D., Cantsilieris, S., Pollen, A. A., Ventura, M., Antonacci, F., Marschall, T., Korbel, J. O., & Eichler, E. E. (2020). Recurrent inversion toggling and great ape genome evolution. *Nature Genetics*, 52(8), 849–858. <https://doi.org/10.1038/s41588-020-0646-x>
- Proulx, G., & Aubry, K. B. (2017). The *Martes* complex: A monophyletic clade that shares many life-history traits and conservation challenges. In: G. Proulx (Ed.). *The Martes Complex in the 21st Century: ecology and conservation* (pp. 3–24). Mammal Research Institute, Polish Academy of Sciences.
- Ranwez, V., Harispe, S., Delsuc, F., & Douzery, E. J. P. (2011). MACSE: Multiple alignment of coding SEquences accounting for frameshifts and stop codons. *PLoS One*, 6(9), e22594. <https://doi.org/10.1371/journal.pone.0022594>
- Reis, A., Chang, H.-Y., Levasseur, M., & Jones, K. T. (2006). APCcdh1 activity in mouse oocytes prevents entry into the first meiotic division. *Nature Cell Biology*, 8(5), 539–540. <https://doi.org/10.1038/ncb1406>
- Ren, X., Aleshin, M., Jo, W. J., Dills, R., Kalman, D. A., Vulpe, C. D., Smith, M. T., & Zhang, L. (2011). Involvement of N-6 adenine-specific DNA methyltransferase 1 (N6AMT1) in arsenic biomethylation and its role in arsenic-induced toxicity. *Environmental Health Perspectives*, 119(6), 771–777.
- Reynolds, A., Qiao, H., Yang, Y., Chen, J. K., Jackson, N., Biswas, K., Holloway, J. K., Baudat, F., de Massy, B., Wang, J., Höög, C., Cohen, P. E., & Hunter, N. (2013). RNF212 is a dosage-sensitive regulator of crossing-over during mammalian meiosis. *Nature Genetics*, 45(3), 269–278. <https://doi.org/10.1038/ng.2541>
- Rinker, D. C., Specian, N. K., Zhao, S., & Gibbons, J. G. (2019). Polar bear evolution is marked by rapid changes in gene copy number in response to dietary shift. *Proceedings of the National Academy of Sciences of the United States of America*, 116(27), 13446–13451. <https://doi.org/10.1073/pnas.1901093116>
- Sato, J. J., Wolsan, M., Prevosti, F. J., D'Elia, G., Begg, C., Begg, K., Hosoda, T., Campbell, K. L., & Suzuki, H. (2012). Evolutionary and biogeographic history of weasel-like carnivorans (Musteloidea). *Molecular Phylogenetics and Evolution*, 63(3), 745–757. <https://doi.org/10.1016/j.ympev.2012.02.025>
- Sayyari, E., & Mirarab, S. (2016). Fast coalescent-based computation of local branch support from quartet frequencies. *Molecular Biology and Evolution*, 33(7), 1654–1668. <https://doi.org/10.1093/molbev/msw079>
- Schweitzer, L. D., Comb, W. C., Bar-Peled, L., & Sabatini, D. M. (2015). Disruption of the rag-ulator complex by c17orf59 inhibits mTORC1. *Cell Reports*, 12(9), 1445–1455. <https://doi.org/10.1016/j.celrep.2015.07.052>
- Serão, N. V. L., Veroneze, R., Ribeiro, A. M. F., Verardo, L. L., Braccini Neto, J., Gasparino, E., Campos, C. F., Lopes, P. S., & Guimarães, S. E. F. (2011). Candidate gene expression and intramuscular fat content in pigs. *Journal of Animal Breeding and Genetics*, 128(1), 28–34. <https://doi.org/10.1111/j.1439-0388.2010.00887.x>
- Shi, G., Xing, L., Liu, Z., Qu, Z., Wu, X., Dong, Z., Wang, X., Gao, X., Huang, M., Yan, J., Yang, L., Liu, Y., Ptáček, L. J., & Xu, Y. (2013). Dual roles of FBXL3 in the mammalian circadian feedback loops are important for period determination and robustness of the clock. *Proceedings of the National Academy of Sciences of the United States of America*, 110(12), 4750–4755. <https://doi.org/10.1073/pnas.1302560110>
- Shumate, A., & Salzberg, S. L. (2020). *Liftoff: an accurate gene annotation mapping tool*. <https://doi.org/10.1101/2020.06.24.169680>
- Siepkha, S. M., Yoo, S.-H., Park, J., Song, W., Kumar, V., Hu, Y., Lee, C., & Takahashi, J. S. (2007). Circadian mutant Overtime reveals F-box

- protein FBXL3 regulation of cryptochrome and period gene expression. *Cell*, 129(5), 1011–1023. <https://doi.org/10.1016/j.cell.2007.04.030>
- Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19), 3210–3212. <https://doi.org/10.1093/bioinformatics/btv351>
- Singh, B. N., Gong, W., Das, S., Theisen, J. W. M., Sierra-Pagan, J. E., Yannopoulos, D., Skie, E., Shah, P., Garry, M. G., & Garry, D. J. (2019). Etv2 transcriptionally regulates Yes1 and promotes cell proliferation during embryogenesis. *Scientific Reports*, 9(1), 9736. <https://doi.org/10.1038/s41598-019-45841-5>
- Smit, A. F. A. (2004). Repeat-masker open-3.0. <http://www.repeatmasker.org> and <https://ci.nii.ac.jp/naid/10029514778/>
- Soley, F. G., & Alvarado-Díaz, I. (2011). Prospective thinking in a mustelid? *Eira barbara* (Carnivora) cache unripe fruits to consume them once ripened. *Die Naturwissenschaften*, 98(8), 693–698. <https://doi.org/10.1007/s00114-011-0821-0>
- Sonna, L. A., Fujita, J., Gaffin, S. L., & Lilly, C. M. (2002). Invited review: Effects of heat and cold stress on mammalian gene expression. *Journal of Applied Physiology*, 92(4), 1725–1742. <https://doi.org/10.1152/jappphysiol.01143.2001>
- Subramanian, A. R., Kaufmann, M., & Morgenstern, B. (2008). DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. *Algorithms for Molecular Biology*, 3, 6. <https://doi.org/10.1186/1748-7188-3-6>
- Svishcheva, G. R., & Kashtanov, S. N. (2011). Reproductive strategy of the sable (*Martes zibellina* Linnaeus, 1758): An analysis of litter size inheritance in farm-raised populations. *Russian Journal of Genetics: Applied Research*, 1(3), 221–225. <https://doi.org/10.1134/S2079-059711030129>
- Tavori, H., Rashid, S., & Fazio, S. (2015). On the function and homeostasis of PCSK9: Reciprocal interaction with LDLR and additional lipid effects. *Atherosclerosis*, 238(2), 264–270. <https://doi.org/10.1016/j.atherosclerosis.2014.12.017>
- The UniProt Consortium (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1), D158–D169.
- Thomas, J. H. (2007). Rapid birth-death evolution specific to xenobiotic cytochrome P450 genes in vertebrates. *PLoS Genetics*, 3(5), e67. <https://doi.org/10.1371/journal.pgen.0030067>
- Tigano, A., Colella, J. P., & MacManes, M. D. (2020). Comparative and population genomics approaches reveal the basis of adaptation to deserts in a small rodent. *Molecular Ecology*, 29(7), 1300–1314. <https://doi.org/10.1111/mec.15401>
- Totikov, A., Tomarovsky, A., Prokopov, D., Yakupova, A., Bulyonkova, T., Derežanin, L., Rasskazov, D., Wolfsberger, W. W., Koepfli, K.-P., Oleksyk, T. K., & Kliver, S. (2021). Chromosome-level genome assemblies expand capabilities of genomics for conservation biology. *Genes*, 12(9), 1336. <https://doi.org/10.3390/genes12091336>
- Viscarra, J. A., Rodriguez, R., Vazquez-Medina, J. P., Lee, A., Tift, M. S., Tavoni, S. K., Crocker, D. E., & Ortiz, R. M. (2013). Insulin and GLP-1 infusions demonstrate the onset of adipose-specific insulin resistance in a large fasting mammal: Potential glucogenic role for GLP-1. *Physiological Reports*, 1(2), e00023. <https://doi.org/10.1002/phy2.23>
- Wang, J., Lu, Z.-X., Tokheim, C. J., Miller, S. E., & Xing, Y. (2015). Species-specific exon loss in human transcriptomes. *Molecular Biology and Evolution*, 32(2), 481–494. <https://doi.org/10.1093/molbev/msu317>
- Webb, J. K., Brown, G. P., Child, T., Greenlees, M. J., Phillips, B. L., & Shine, R. (2008). A native dasyurid predator (common planigale, *Planigale maculata*) rapidly learns to avoid a toxic invader. *Austral Ecology*, 33(7), 821–829.
- Weischenfeldt, J., Symmons, O., Spitz, F., & Korbel, J. O. (2013). Phenotypic impact of genomic structural variation: Insights from and for human disease. *Nature Reviews. Genetics*, 14(2), 125–138. <https://doi.org/10.1038/nrg3373>
- Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., & Jaffe, D. B. (2017). Direct determination of diploid genome sequences. *Genome Research*, 27(5), 757–767. <https://doi.org/10.1101/gr.214874.116>
- Weissensteiner, M. H., Bunikis, I., Catalán, A., Francoijs, K.-J., Knief, U., Heim, W., Peona, V., Pophaly, S. D., Sedlazeck, F. J., Suh, A., Warmuth, V. M., & Wolf, J. B. W. (2020). Discovery and population genomics of structural variation in a songbird genus. *Nature Communications*, 11(1), 3403. <https://doi.org/10.1038/s41467-020-17195-4>
- Wellenreuther, M., & Bernatchez, L. (2018). Eco-evolutionary genomics of chromosomal inversions. *Trends in Ecology & Evolution*, 33(6), 427–440. <https://doi.org/10.1016/j.tree.2018.04.002>
- Wilcox, A., Katsanakis, K. D., Bheda, F., & Pillay, T. S. (2004). Asb6, an adipocyte-specific ankyrin and SOCS box protein, interacts with APS to enable recruitment of elongins B and C to the insulin receptor signaling complex. *Journal of Biological Chemistry*, 279(37), 38881–38888. <https://doi.org/10.1074/jbc.M406101200>
- Wilson, D. E., Mittermeier, R. A. (2009). Family Mustelidae. In: *Handbook of the mammals of the world* (vol. 1, pp. 627–637). Lynx Editions.
- Wu, X., Quondamatteo, F., Lefever, T., Czuchra, A., Meyer, H., Chrostek, A., Paus, R., Langbein, L., & Brakebusch, C. (2006). Cdc42 controls progenitor cell differentiation and beta-catenin turnover in skin. *Genes & Development*, 20(5), 571–585.
- Wu, Z., Latendorf, T., Meyer-Hoffert, U., & Schröder, J.-M. (2011). Identification of trichohyalin-like 1, an s100 fused-type protein selectively expressed in hair follicles. *Journal of Investigative Dermatology*, 131(8), 1761–1763. <https://doi.org/10.1038/jid.2011.71>
- Xing, Y., & Lee, C. (2006). Alternative splicing and RNA selection pressure—evolutionary consequences for eukaryotic genomes. *Nature Reviews. Genetics*, 7(7), 499–509. <https://doi.org/10.1038/nrg1896>
- Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8), 1586–1591. <https://doi.org/10.1093/molbev/msm088>
- Yu, Y.-H., Chang, Y.-C., Su, T.-H., Nong, J.-Y., Li, C.-C., & Chuang, L.-M. (2013). Prostaglandin reductase-3 negatively modulates adipogenesis through regulation of PPAR γ activity. *Journal of Lipid Research*, 54(9), 2391–2399. <https://doi.org/10.1194/jlr.M037556>
- Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, 19(Suppl 6), 153. <https://doi.org/10.1186/s12859-018-2129-y>
- Zhang, H., Ge, Y., He, P., Chen, X., Carina, A., Qiu, Y., Aga, D. S., & Ren, X. (2015). Interactive effects of N6AMT1 and As3MT in arsenic biomethylation. *Toxicological Sciences: An Official Journal of the Society of Toxicology*, 146(2), 354–362. <https://doi.org/10.1093/toxsci/kfv101>
- Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, 18(6), 292–298. [https://doi.org/10.1016/S0169-5347\(03\)00033-8](https://doi.org/10.1016/S0169-5347(03)00033-8)
- Zhang, J., Nielsen, R., & Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular Biology and Evolution*, 22(12), 2472–2479. <https://doi.org/10.1093/molbev/msi237>
- Zhang, Z., Li, J., Zhao, X.-Q., Wang, J., Wong, G.-K.-S., & Yu, J. (2006). KaKs_Calculator: Calculating Ka and Ks through model selection and model averaging. *Genomics, Proteomics & Bioinformatics*, 4(4), 259–263. [https://doi.org/10.1016/S1672-0229\(07\)60007-2](https://doi.org/10.1016/S1672-0229(07)60007-2)

Zhou, Y.-B., Newman, C., Xu, W.-T., Buesching, C. D., Zalewski, A., Kaneko, Y., Macdonald, D. W., & Xie, Z.-Q. (2011). Biogeographical variation in the diet of Holarctic martens (genus *Martes*, Mammalia: Carnivora: Mustelidae): adaptive foraging in generalists. *Journal of Biogeography*, 38(1), 137–147. <https://doi.org/10.1111/j.1365-2699.2010.02396.x>

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Derežanin, L., Blažytė, A., Dobrynin, P., Duchêne, D. A., Grau, J. H., Jeon, S., Kliver, S., Koepfli, K.-P., Meneghini, D., Preick, M., Tomarovsky, A., Totikov, A., Fickel, J., & Förster, D. W. (2022). Multiple types of genomic variation contribute to adaptive traits in the mustelid subfamily Guloninae. *Molecular Ecology*, 31, 2898–2919. <https://doi.org/10.1111/mec.16443>

Chapter II

Comparative analyses inform the genomic consequences of the population bottleneck in the endangered black-footed ferret

Lorena Derežanin^{1*}, Yana Safanova^{2*}, Sergei Kliver³, Claudia Fontseré⁴, Aitor Serres-Armero⁴, Andrei Tomarovsky^{3,5,6}, Azamat Totikov^{3,5,6}, Graham Etherington⁷, Wilfried Haerty^{7,8}, Federica Di Palma^{7,8,9}, Polina L. Perelman³, Violetta Beklemisheva³, Natalia Serdyukova³, Alexander Graphodatsky³, José Melo-Ferreira^{10,11,12}, Paul Marinari¹³, Tomas Marques-Bonet^{4,14,15,16}, Jörns Fickel^{1,17}, Daniel W. Förster¹, Klaus-Peter Koepfli^{5,13,18}

¹Leibniz Institute for Zoo and Wildlife Research (IZW), Alfred Kowalke Straße 17, 10315 Berlin, Germany

²Department of Computer Science, Johns Hopkins University, Baltimore, Maryland 21218, USA

³Department of the Diversity and Evolution of Genomes, Institute of Molecular and Cellular Biology SB RAS, 630090 Novosibirsk, Russia

⁴Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Barcelona, Spain

⁵Computer Technologies Laboratory, ITMO University, 197101 Saint Petersburg, Russia

⁶Novosibirsk State University, 630090 Novosibirsk, Russia

⁷Earlham Institute, Norwich Research Park, Colney Ln, Norwich NR4 7UZ, United Kingdom

⁸School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, United Kingdom

⁹Research and Innovation, Genome British Columbia, 575 W 8th Ave #400, Vancouver BC V5Z 0C4, Canada

¹⁰CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Campus de Vairão, Universidade Do Porto, Vairão, Rua Padre Armando Quintas, nr.7, 4485-661, Vairão, Portugal

¹¹BIOPOLIS Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, 4485-661, Vairão, Portugal

¹²Departamento de Biologia, Faculdade de Ciências, Universidade Do Porto, 4099-002, Porto, Portugal

¹³Center for Species Survival, Smithsonian's National Zoo and Conservation Biology Institute, Front Royal, VA, 22630, USA

¹⁴CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology, Barcelona, Spain

¹⁵Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain

¹⁶Catalan Institution of Research and Advanced Studies (ICREA), Barcelona, Spain

¹⁷Institute for Biochemistry and Biology, Faculty of Mathematics and Natural Sciences, University of Potsdam, Karl-Liebknecht- Straße 24-25 Haus 25, 14476 Potsdam OT Golm, Germany

¹⁸Smithsonian-Mason School of Conservation, George Mason University, Front Royal, VA, 22630, USA

*These authors contributed equally.

Manuscript prepared for submission to *Current Biology*.

Summary

Comparing the genomes of threatened and non-threatened species can reveal important insights into the impact of demographic history on genetic diversity, inbreeding, and functional variation. The black-footed ferret (*Mustela nigripes*) was once declared extinct as a result of habitat loss, persecution of its primary prey, and disease susceptibility. In 1981, the last colony of black-footed ferrets was rediscovered and the surviving 18 individuals (an estimated 7 founders) were used to start a successful *ex situ* breeding program which enabled the reintroduction of the species into its former range in North America.

We analyzed the genomes of black-footed ferrets and four other species in the genus *Mustela* to characterize differences in genomic diversity, structural variation, and the organization of the immunoglobulin gene repertoire. Black-footed ferrets showed the lowest levels of heterozygosity and highest burden of runs of homozygosity, with up to 70% of the genome occupied by blocks greater than 10 Mb. Deletions and inversions were the most frequent structural variants observed, and we found a high amount of intra-individual variation in the number and distribution of structural variants in black-footed ferrets, suggesting a hitherto unknown reservoir of genetic diversity with potential impacts on gene function.

Lastly, relative to their congeners, black-footed ferrets lack variable gene clusters at two of three immunoglobulin loci examined and possess an altered complementarity-determining region (CDR) at one of these clusters. Our findings demonstrate the power of comparative analyses for expanding our understanding of the genomic composition of threatened species, thereby informing efforts to manage and protect the genetic diversity that remains.

Identification and functional annotation of genome-wide structural variants in Mustela genomes

Besides SNPs and short indels, structural variants (SVs) represent an important source of genomic variation across a range of species (Wold et al. 2021; Chain and Feulner 2014), but the estimation of their functional impact has remained challenging (Ho, Urban, and Mills 2020). SVs encompass deletions, duplications, insertions, inversions, and translocations of at least 50 bp in size (Alkan, Coe, and Eichler 2011). These variants can be either balanced, with no loss or gain of genetic material, such as inversions or a majority of translocations within or between chromosomes, or unbalanced, where genetic information is lost or gained, also referred to as copy number variation (CNV) (Escaramís, Docampo, and Rabionet 2015).

Compared to SNPs, CNVs encompass an order of magnitude more nucleotides and have higher mutation rates (Zarrei et al. 2015). Besides their key role in ecological adaptation and speciation (Mérot et al. 2020; Rinker et al. 2019; Wellenreuther et al. 2019; Axelsson et al. 2013), SVs are often associated with genomic disorders and diseases (Payer et al. 2017; Weischenfeldt et al. 2013), with CNVs mainly contributing to disease susceptibility, affecting traits related to immune gene functions (Aitman et al. 2006; Fellersmann et al. 2006).

Black-footed ferrets underwent a fairly recent but severe population bottleneck leading to a decrease in genetic diversity. Their genetic diversity was found to be lower compared to the unaffected steppe polecat population and similar to the bottlenecked European polecat population (Wisely et al. 2002). Furthermore, the small size of a black-footed ferret founder population in the first conservation breeding program led to a decrease in seminal quality of male individuals, subsequently affecting pregnancy and litter size (Santymire et al. 2019).

As structural variation represents a significant source of functional genomic variation it can complement SNP-based approaches in conservation genomic assessments and breeding programs to enhance species recovery and survival (Wold et al. 2021). Here we examined the evolutionary dynamics and diversity of SV in black-footed ferrets compared to closely related, more outbred *Mustela* species.

Methods

We aligned short-read data from three black-footed ferret individuals (SB6536, SB7462, SB8055), least weasel (*Mustela nivalis*), steppe polecat (*Mustela eversmanii*) and European polecat (*Mustela putorius*, (Etherington et al. 2020) to the chromosome-length genome assembly of the domestic ferret (*Mustela putorius furo*) (Dudchenko et al. 2018, 2017). Prior to alignment with Bowtie2 v.2.3.5.1 (Langmead and Salzberg 2012), the 10x linked-read barcodes were trimmed from *M.eversmanii* and *M.putorius* reads. Paired-end reads of *M.eversmanii* and *M.nivalis* were trimmed from 250 to 150 bp length to maintain uniformity among samples, and reduce the number of overlapping reads leading to potential spurious variant calls. Adapter and quality trimming (Q30, min. length 80 bp) were performed on all samples with TrimGalore v.0.6.4. Duplicated reads were removed with Picard v.2.23. Aligned reads of all samples were downsampled to ~30x coverage to reduce bias during structural variant calling.

Structural variant calling was conducted using an ensemble approach, consisting of three SV callers that rely on different detection methods (Manta v.1.6.0, Chen et al. 2016; Whamg v.1.7.0, Kronenberg et al. 2015; and Lumpy v.0.2.13, Layer et al. 2014). We retained Manta calls with minimum paired-read (PR) and split-read (SR) support of $PR \geq 5$ or $SR \geq 5$, respectively. Potential translocation events were removed from the variant sets, in order to decrease the proportion of false-positive calls, as they could not be reliably called in species with different karyotypes. Calls with total evidence supporting a variant (PR and/or SR) below 10 were removed from the Whamg and Lumpy call sets.

Moreover, SV call sets were genotyped with Svtlyper v0.7.1 (Chiang et al. 2015) and were then filtered for genotype quality of $GQ \geq 30$. Only scaffolds assigned to chromosomes were further analyzed. To further reduce the number of potentially unreliable variant calls, variants overlapping gaps and high coverage regions ($> 100x$) in the reference genome were identified and removed. Survivor v.1.0.7 (Jeffares et al. 2017) was used to merge and compare SV call sets within and among samples. Union of SV calls among all samples containing sample-specific and shared variants (Figure 1A.) was annotated using Liftoff v.1.5.1 (Shumate and Salzberg 2020) and Ensembl Variant Effect Predictor v. 101.0 (McLaren et al. 2016) to identify variants affecting protein-coding genes.

Functional classification of genes was conducted through detailed literature and database search (OrthoDB v10, Kriventseva et al. 2019; Uniprot, The UniProt Consortium 2017; NCBI Entrez gene, Maglott et al. 2011), and Gene Ontology enrichment analysis of biological processes (Shiny GO v0.61, FDR < 0.05 , Ge, Jung, and Yao 2020) for SVs encompassing larger gene blocks (>5 genes). As the precise effect of inversions overlapping large sets of genes is still challenging to determine, we inspected inversions

affecting up to 20 genes for significantly enriched biological processes, and no limitation was imposed on other SV types.

Results and Discussion

Overall SV landscape in Mustela species

We characterized four types of SVs (DEL, DUP, INS, INV) in the least weasel (*Mustela nivalis*), steppe polecat (*Mustela eversmanii*), European polecat (*Mustela putorius*) and three black-footed ferrets (*Mustela nigripes*), based on a majority call from an ensemble SV calling approach.

We detected the highest number of inter- and intraspecific SVs (> 50bp) in *M. nivalis* (7138/5743), and the lowest number in *M. putorius* (677/264) (Table 1, Figure 1). Thus, we observed a relationship between the number of SVs detected and the relatedness of a given species to the domestic ferret, the reference species, with an increasing number of SVs detected in more distantly related species (see e.g. phylogenies in Law, Slater, and Mehta 2018; Koepfli et al. 2008). This finding shows a pattern of SV accumulation over time, with the most SVs also found in the most outbred species, the least weasel (*M. nivalis*), which harbours the highest SNP diversity. Similarly, Weissensteiner et al. (2020), reported a positive correlation between SNP and SV diversity in corvid species. Both SNP and SV numbers were lowest in the highly inbred Hawaiian crow compared to more outbred, closely related species.

In all mustelids, deletions were the most frequent SV type, ranging from 524 deletions in *M. putorius* up to 6149 in *M. nivalis* (Table 1). The least abundant SV type were insertions, ranging from 2 in the *M. nigripes* sample SB6536 to 149 in the *M. nigripes* sample SB8055, all detected in non-coding genomic regions. Only one insertion affected the genic region, detected in *M. nivalis* (Figure 2A).

Table 1. Total SV counts in all samples split by variant type.

Species	DEL	DUP	INS	INV	Total
<i>M. putorius</i>	524	44	5	104	677
<i>M. eversmanii</i>	1797	74	24	263	2158
<i>M. nigripes</i> _SB6536	2109	74	2	319	2504
<i>M. nigripes</i> _SB7462	2675	85	28	418	3206
<i>M. nigripes</i> _SB8055	2703	116	149	422	3390
<i>M. nivalis</i>	6149	160	47	782	7138

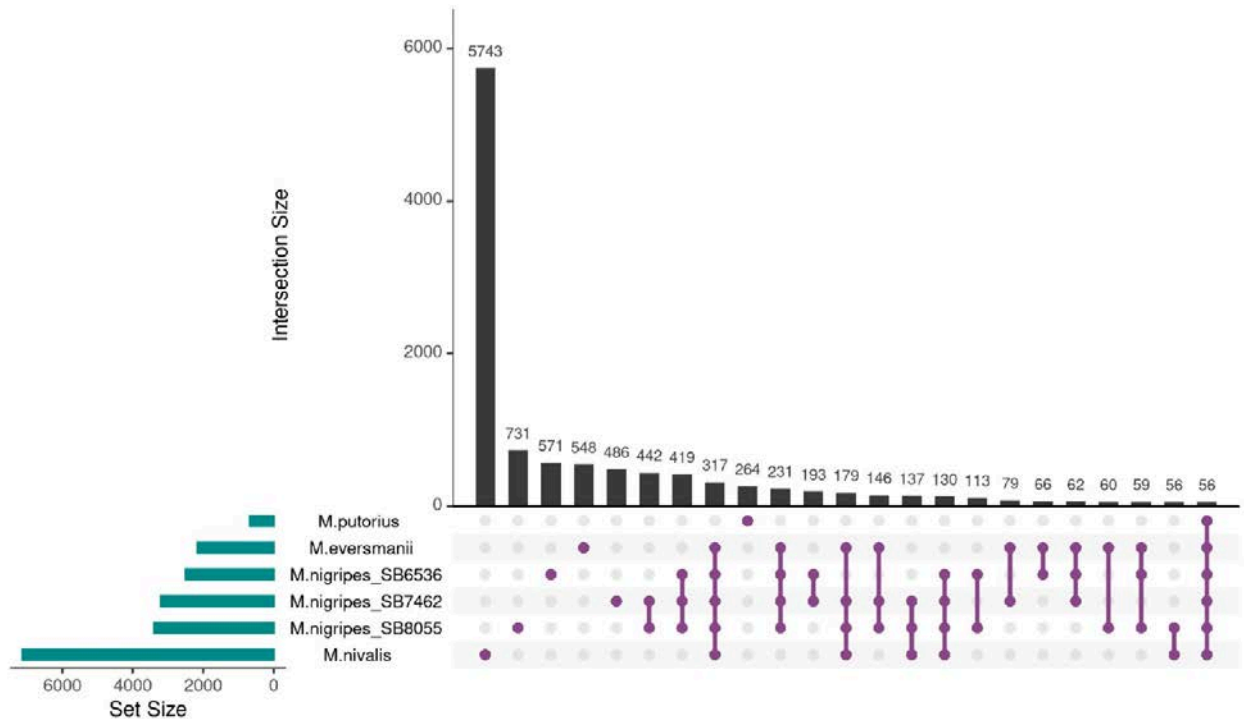


Figure 1. Intra and interspecific structural variation in *Mustela* sp.

Distribution of shared and sample-specific structural variants in three black-footed ferret individuals (SB6536, SB8055, SB7462), European polecat (*M_putorius*), steppe polecat (*M_eversmanii*) and least weasel (*M_nivalis*). The green barplot on the left indicates total SV counts per sample.

Genic SVs in Mustela species

The majority of SVs in investigated mustelids were located in non-coding regions (> 96%), with a smaller proportion of SVs either completely or partially overlapping protein-coding regions and intergenic regions putatively implicated in the regulation of nearby genes (Figure 2A). Structural variants impacting coding regions were all classified either as high-impact variants or modifiers (Variant Effect Predictor (VEP), McLaren et al. 2016).

The highest number of sample-specific genic and nearby intergenic SVs was 229 (3.99% of the total number of sample-specific SVs) observed in *M. nivalis*, while the lowest was 18 (3.15%), detected in the black-footed ferret SB6536 (Figure 2A). In all mustelids, the majority of genic and nearby intergenic variants were in the size class with lengths of 50 bp - 20kb (Figure 2B).

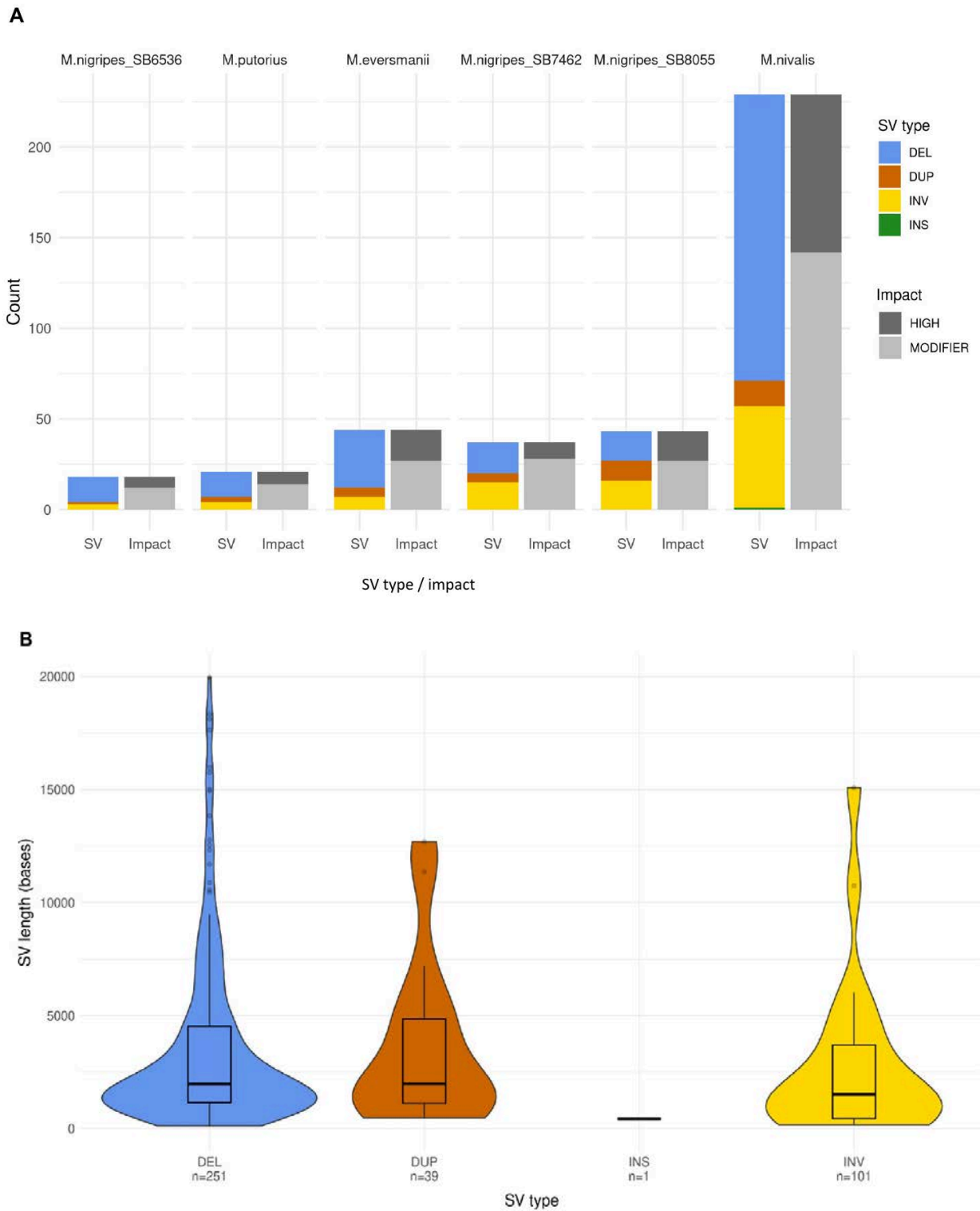


Figure 2. Structural variation type, variant impact and size distribution.

(A) The number of different structural variant types (deletion - DEL, duplication - DUP, inversion - INV, insertion - INS) overlapping sample-specific genic and nearby intergenic regions. Gray bars indicate SV impact classified by VEP based on the severity of consequences (high or modifier). (B) Length distribution of structural variants split by type. Aggregated lengths of sample-specific genic SVs and nearby intergenic regions (max SV length shown = 20 kb).

Most genic SVs were heterozygous in all samples, with deletions being the primary SV type in all samples except for SB8055, for which heterozygous inversions are more prevalent (Supplementary Figure S1). Inversions are known to strongly impact genome evolution by suppressing recombination and facilitating the protection of favourable allelic combinations in heterozygous individuals (Dobigny, Britton-Davidian, and Robinson 2017; Hammer, Schimenti, and Silver 1989). To determine the functions of genes affected by SVs, we classified genes into eight functional categories (Figure 3).

In *M. nivalis* the functional group with the highest number of genes (145) is cell cycle processes (e.g. DNA and RNA transcription, protein modification, mitosis), followed by genes related to the immune system (28), metabolism (25), and nervous system (24). In one of the black-footed ferret individuals (*M. nigripes* SB8055), we found a high number of genes affected by large SVs, associated with the cell cycle (102), metabolism (32), followed by sensory perception (25). In *M. nigripes* SB6536, the species with the lowest number of genic SVs, functional categories are mostly involved in cell cycle processes (22) and the nervous system (6).

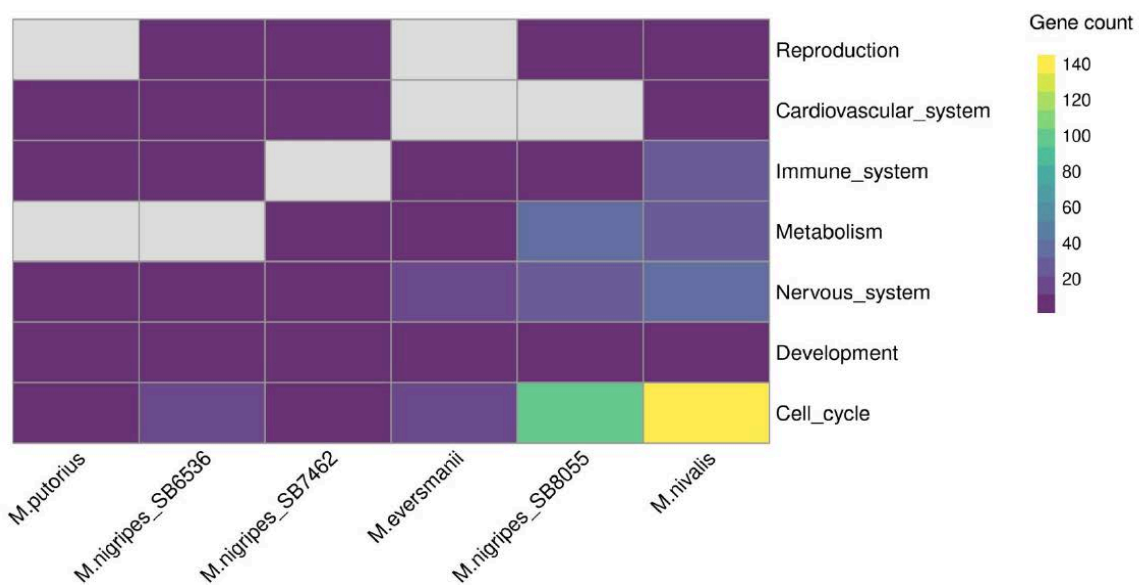


Figure 3. Functional categories of genes overlapped by SV.

Functional categories of genes affected by sample-specific SVs. Categories with no genes detected for a given sample are presented with a grey color.

Genic SVs in black-footed ferrets

Among the 419 SVs exclusively shared by all three black-footed ferrets (Figure 1), we detected 305 DELs, 10 DUPs and 104 INVs. Private genic SVs detected in each of the black-footed ferret individuals are represented in relation to domestic ferret chromosomes (Figure 4 A-C). Only ten of these shared SVs overlap coding sequence (CDS) regions (Figure 4D). Unlike a large number of sample-specific genic SVs with predicted high impact (Figure 2A), only two of the ten shared SVs among black-footed ferret individuals are flagged as high-impact variants. The majority (417) of shared variants are identified as modifiers, and the two deletions annotated as high-impact, are putatively inducing severe consequences, affecting two to five exons of the genes (*ENSMYPUG00000011741*, *ENSMYPUG00000006045* and *ENSMYPUG00000005865*) leading to truncation of the CDS. We excluded two shared inversions overlapping large sets of genes (> 20 genes) from further GO analysis as the precise effect on multiple genes is challenging to determine without support from transcriptome data.

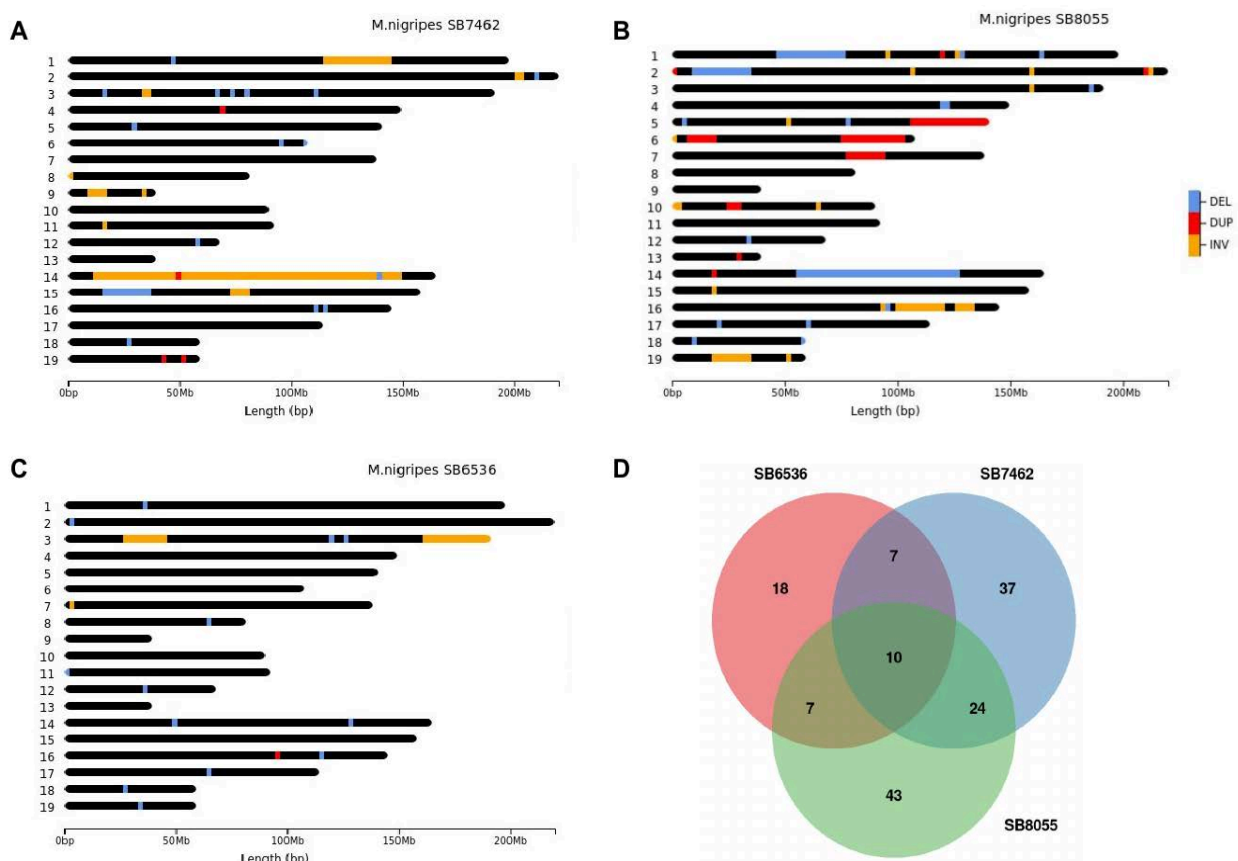


Figure 4. SV composition and distribution in three black-footed ferrets.

(A-C) Private structural variants overlapping genic regions presented on domestic ferret chromosomes for each of the three black-footed ferrets (SB6536, SB8055, SB7462). (D) SV counts of shared and private SVs detected in the coding sequence (CDS) in all three individuals.

Functional groups of genes overlapped by SVs shared in three black-footed ferrets include cell cycle (63 genes), nervous system (29, out of which 4 are sensory perception-related), development (12), metabolism/ energy conversion (8), reproduction (6), cardiovascular system (3), and immune system (1).

Shared genic SVs in all tree individuals are localized on autosomes. Notably, we detected short homozygous deletions (~1-2 kb long) within four reproduction-related genes. The *OSBP2* is involved in spermatogenesis, lipid trafficking and apoptosis (Annis et al. 2002). In mice lacking this gene, a condition that includes low sperm number, low sperm motility and abnormal sperm morphology (a common cause of male infertility) has been observed. In male mice that are homozygous mutant for *OSBP2*, sperm cell proliferation and subsequent meiosis occur normally, but the morphology of cells is severely distorted, with spermatozoa having little to no motility and no fertilizing ability *in vitro*. On the other hand, females display normal fecundity (Udagawa et al. 2014). Moreover, we detected *SPACA1*, a gene also associated with spermatogenesis, and acrosome assembly. The loss of function of this gene in male mice and human individuals has been found to cause globozoospermia - a condition involving malformation or loss of the acrosome, and subsequently leads to infertility (Fujihara et al. 2012; Chen et al. 2021).

Furthermore, *RECK* and *TANCI*, genes involved with embryo implantation, blood vessel maturation, and embryonic development (Welm, Mott, and Werb 2002; Han et al. 2010) harbor a short deletion in ferrets, respectively. The lack of *RECK* in mice embryos has been shown to halt *in utero* angiogenesis and was lethal (Chandana et al. 2010), while *TANCI*-knockout mice displayed impaired neuronal development (Han et al. 2010).

However, a homozygous duplication has been detected in all three black-footed ferrets, within *SETD3*, a gene associated with muscle cell differentiation and regulation of uterine smooth muscle contraction (Eom et al. 2011). *SETD3*-deficient female mice have severely decreased litter sizes owing to primary maternal dystocia, leading to no births or incomplete delivery with fetuses remaining *in utero* (Wilkinson et al. 2019). Duplication within this gene may be implicated in the regulation of smooth muscle contractility of the uterus during labor.

Additionally, we detected a heterozygous inversion in ferrets, within *IHO1*, a gene involved in gametogenesis and homologous chromosome pairing at meiosis (Stanzione et al. 2016). In mice, both female and male *IHO-1*-knockouts were found to be infertile. Oocytes were depleted in six-week-old *Iho1*^{-/-} females, and spermatocytes in males underwent early apoptosis, suggestive of meiotic recombination defects (Burgoyne, Mahadevaiah, and Turner 2009). This polymorphic inversion may imply a putative adaptive potential in ferrets.

In all three black-footed ferret individuals, there is persistent diversity in the form of SVs in parts of their genomes. Some of the genes affected by SVs may be associated with adaptive processes, despite the presence of potentially deleterious SVs. Similarly, retention of variation in parts of the genome associated with adaptation, despite the presence of detrimental anatomical and physiological effects linked to inbreeding, was observed in a small and isolated Apennine brown bear population (Benazzo et al. 2017). Random fixation of SVs certainly carries an increased extinction risk in small populations. Still, it can be tolerated if balancing selection prevents random loss of variation in regions of adaptive potential.

Supplementary Figure

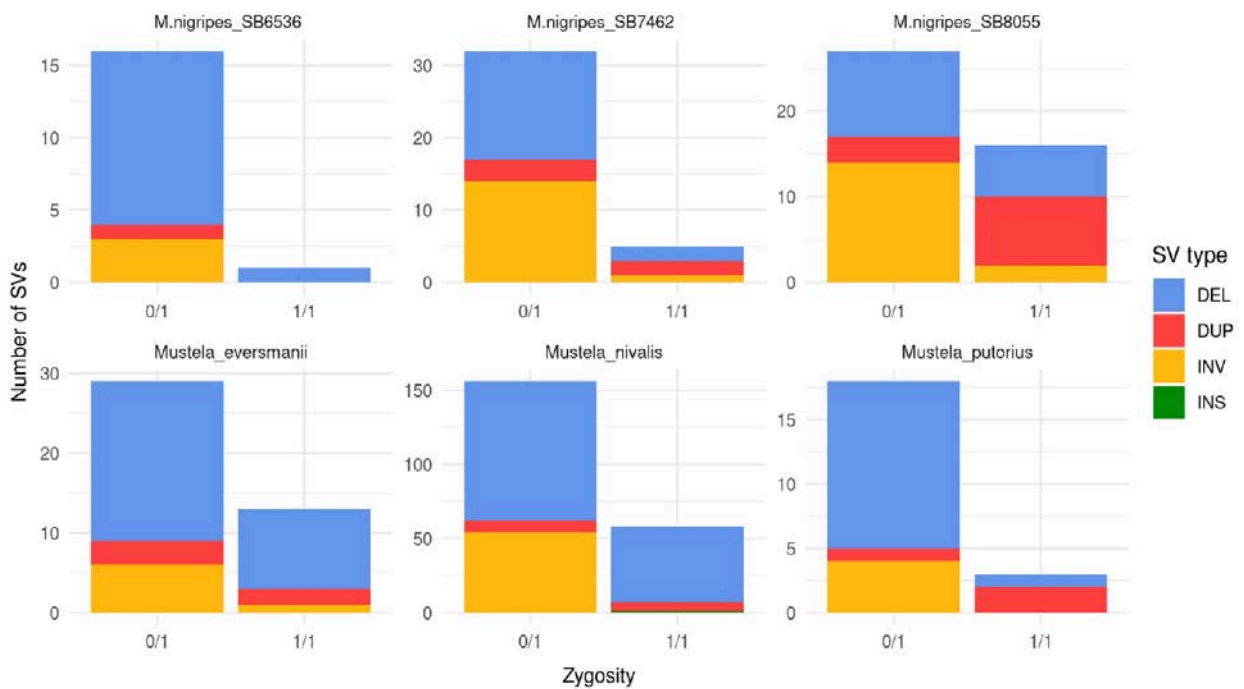


Figure S1. Heterozygous and homozygous sample-specific genic and nearby intergenic variants. Heterozygous variants are denoted with 0/1 and homozygous with 1/1. Note the difference in the y-axis scale.

Chapter III

Genomic characterization of the world's longest selection experiment in mouse reveals the complexity of polygenic traits

Sergio E. Palma-Vera^{1*}, Henry Reyer², Martina Langhammer³, Norbert Reinsch³, **Lorena Derezanin**^{1,4}, Joerns Fickel^{4,5}, Saber Qanbari³, Joachim M. Weitzel¹, Soeren Franzenburg⁵, Georg Hemmrich-Stanisak⁶, Jennifer Schoen^{1,7}

¹*Institute of Reproductive Biology, Research Institute for Farm Animal Biology (FBN), Dummerstorf Germany*

²*Institute of Genome Biology, Research Institute for Farm Animal Biology (FBN), Dummerstorf Germany*

³*Institute of Genetics and Biometry, Research Institute for Farm Animal Biology (FBN), Dummerstorf Germany*

⁴*Department of Evolutionary Genetics, Research Institute for Zoo and Wildlife Research (IZW), Berlin, Germany*

⁵*University of Potsdam, Institute for Biochemistry and Biology, Potsdam, Germany*

⁶*Institute of Clinical Molecular Biology (IKMB), Kiel, Germany*

⁷*Department of Reproduction Biology, Research Institute for Zoo and Wildlife Research (IZW), Berlin, Germany*

*Corresponding author


Published in *BMC Biology*, 2021, DOI: 10.1186/s12915-022-01248-9

RESEARCH ARTICLE

Open Access



Genomic characterization of the world's longest selection experiment in mouse reveals the complexity of polygenic traits

Sergio E. Palma-Vera^{1*} , Henry Reyer², Martina Langhammer³, Norbert Reinsch³, Lorena Derezanin^{1,4}, Joerns Fickel^{4,5}, Saber Qanbari³, Joachim M. Weitzel¹, Soeren Franzenburg⁶, Georg Hemmrich-Stanisak⁶ and Jennifer Schoen^{1,7}

Abstract

Background: Long-term selection experiments are a powerful tool to understand the genetic background of complex traits. The longest of such experiments has been conducted in the Research Institute for Farm Animal Biology (FBN), generating extreme mouse lines with increased fertility, body mass, protein mass and endurance. For >140 generations, these lines have been maintained alongside an unselected control line, representing a valuable resource for understanding the genetic basis of polygenic traits. However, their history and genomes have not been reported in a comprehensive manner yet. Therefore, the aim of this study is to provide a summary of the breeding history and phenotypic traits of these lines along with their genomic characteristics. We further attempt to decipher the effects of the observed line-specific patterns of genetic variation on each of the selected traits.

Results: Over the course of >140 generations, selection on the control line has given rise to two extremely fertile lines (>20 pups per litter each), two giant growth lines (one lean, one obese) and one long-distance running line. Whole genome sequencing analysis on 25 animals per line revealed line-specific patterns of genetic variation among lines, as well as high levels of homozygosity within lines. This high degree of distinctiveness results from the combined effects of long-term continuous selection, genetic drift, population bottleneck and isolation. Detection of line-specific patterns of genetic differentiation and structural variation revealed multiple candidate genes behind the improvement of the selected traits.

Conclusions: The genomes of the Dummerstorf trait-selected mouse lines display distinct patterns of genomic variation harbouring multiple trait-relevant genes. Low levels of within-line genetic diversity indicate that many of the beneficial alleles have arrived to fixation alongside with neutral alleles. This study represents the first step in deciphering the influence of selection and neutral evolutionary forces on the genomes of these extreme mouse lines and depicts the genetic complexity underlying polygenic traits.

Keywords: Mouse, Fertility, Body mass, Endurance, Selective breeding, Genetic drift, Bottleneck, Whole genome sequencing, Single-nucleotide polymorphism, Structural variation

Background

Artificial selection is the selective breeding of organisms by which desired phenotypic traits evolve in a population [1]. Farm animals are the result of this selective breeding process to achieve efficient food production.

*Correspondence: serpalma.v@gmail.com

¹ Institute of Reproductive Biology, Research Institute for Farm Animal Biology (FBN), Dummerstorf, Germany

Full list of author information is available at the end of the article



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

However, artificial selection can also be applied experimentally in other species in order to connect genes and other genomic elements to selection response for complex traits such as behaviour [2] and limb elongation [3]. More generally, experimental evolution, which includes artificial selection experiments, is a powerful approach to understand response to selection across multiple traits and organisms [4].

The worldwide longest selection experiment on mice began in 1969 at the former Forschungszentrum für Tierproduktion (FZT), nowadays called Research Institute for Farm Animal Biology (FBN) located in Dummerstorf, Germany [5, 6]. Starting from a single founder line developed from four outbred and four inbred mouse strains [5, 6], selection lines for different complex traits were bred with population sizes of 60–100 breeding pairs per line. An unselected control line from the same founder line was maintained over the entire selection period with a larger population size (125–200 breeding pairs) [5, 6]. Over the course of >140 generations, selection has shaped the genomes of the Dummerstorf trait-selected mouse lines, and led to extreme phenotypes that include increased litter size (approx. double the litter size of the unselected mouse line) [7], body mass (approx. 90g body weight at 6 weeks of age) [8] and endurance (more than 3× higher untrained running capacity) [9, 10]. Therefore, in order to elucidate the unpredictable polygenic background of these complex traits, where multiple genes, regulatory elements and pathways act in conjunction, the Dummerstorf trait-selected mouse lines represent a valuable resource.

Other selection experiments have generated mice with increased litter size [11–14], as well as mice with enhanced body weight (see [15, 16] for a list of body weight mouse lines) and exercise performance [17], yet few studies have examined the polygenic background of these traits through genomic analysis. For example, a genome-wide association study of the high-fertility inbred strain QSi5 corroborated multiple previously reported loci associated with reproductive performance [18]. Likewise, a multi-line approach detected shared loci controlling body weight across seven high body weight selection lines, including an inbred subline of the Dummerstorf's body mass line [16]. Finally, a comprehensive genomic analysis of mice from the "High Runner" selection experiment found widespread regions with significant genetic differentiation between selected and unselected replicate lines (4 per group) [19].

The Dummerstorf mouse lines expand the repertoire of polygenic mouse models to understand the genetic basis of fertility, body weight and endurance. Each of these lines arose from almost the same genetic diversity and has been maintained to this day for about half

a century. Here we describe the selection history of this unique selection experiment, characterize line-specific patterns of genetic variation and identify genes that are likely associated to each selection trait.

Results and discussion

Phenotypic impact of selection

Over the course of more than 140 generations (Table 1), the selected traits (Table 2) have shown remarkable increments in each line (Table 1, Fig. 1, Additional file 2: Figure S1). The span and number of generations makes the present study the longest selection experiment ever reported in mice. Relative to the unselected control line FZTDU (exposed to genetic drift only), reproductive performance has doubled in DUK (Fertility mouse line 1) and DUC (Fertility mouse line 2) (Fig. 1A,B, E,G, Additional file 2: Figure S1). Even though these two trait-selected lines have achieved comparable litter sizes at first delivery (>20 offspring) [20], their reproductive lifespan differs, with 5.8 and 2.7 litters in average per lifetime for DUK and DUC, respectively [20]. A remarkable level of divergence has been achieved by the increased body size lines (Fig. 1C,D, Additional file 2: Figure S1). Individuals of the body mass line (DU6) have almost tripled their weight compared to FZTDU (Fig. 1H, Additional file 2: Figure S1), whereas mice of the protein mass line (DU6P) not only have become larger and heavier than FZTDU mice, but their level of muscularity is also considerably higher (Fig. 1D,I, Additional file 2: Figure S1). In terms of running distance capacity, the treadmill performance line (DUhLB) can on average cover three times more distance than FZTDU (Fig. 1J, Additional file 2: Figure S1).

With the exception of the obese line DU6 [21], each one of the trait-selected mouse lines has developed an extreme phenotype without obvious detrimental effects on their general health, well-being, and longevity. All these lines are still maintained, but selection only continues for DUK, DUC and DU6. Due to the long span of this selection experiment, lines have been given alternative names (Table 1, Additional file 3 [6, 8, 10, 20–41]: Table S1) and selected at variable intensities (Additional file 2: Figure S2).

Whole genome sequencing (WGS) analysis and short variant detection

After quality filtering and trimming, >90% of the raw reads were mapped to the genome as pairs, with a mean insert size of ~380 bp. For samples sequenced at a target coverage of 30×, mean genome-wide coverage averaged ~24×, with ~95% of genome territory covered at least 5×; samples sequenced at a target coverage of 5× averaged ~8× and ~72%, respectively (for a summary across all samples see Table 3 and for details, Additional file 4: Data S1).

Table 1 Summary selection history of the Dummerstorf mouse lines

	FZTDU	DUK	DUC	DU6	DU6P	DUHLB
Established (year)	1969	1971	1971	1975	1975	1982
No. Founders (BPs)	NR	60	60	80	80	100
Trait increment	–	2×	.2×	3×	2×	3×
Percentage selected ^a	–	25–80	25–80	45–90	45–70	40–100
Relocation ^b at generation	160–164	165	163–164	154–155	154–155	120–121
BPs per generation before relocation	200	60–100	60–100	60–80	60–80	60–100
BPs after relocation (founders)	55	19	24	7	19	22
BPs per generation (current)	125	60	60	60–120	60	60
End of selection (at generation)	–	Ongoing	Ongoing	Ongoing	152	141
No. generations under selection ^c	–	182/189	180/187	169/177	152/152	117/117
WGS at generation(s)	188/195	188/195	186/193	177/185	177/184	143/150
Alternative names ^d	Fzt: DU, DUK, Ctrl	DU-K, FL1	DU-C, FL2	BW, Titan	PA	DU-hTP

BPs breeding pairs, WGS whole genome sequencing, NR no records

Trait increment: mean trait expression in the sampled generation compared with trait expression in starting generation

^a Percentage selected: percentage of litters from which parents were chosen

^b Transfer of animals to a new housing building in 2011

^c Total generations under selection until first and second sampling

^d See Additional file 3: Table S1 for references on alternative names

Table 2 Selection criteria for Dummerstorf trait-selected mouse lines

Line-ID	Selected Sex	Trait
FZTDU	–	Unselected
DUK	Females	Number of offspring in first litter and litter weight at birth
DUC	Females	Number of offspring in first litter and litter weight at birth
DU6	Males	Body mass at day 42 of age
DU6P	Males	Protein amount in carcass at day 42 of age
DUHLB	Males	Submaximal untrained running distance on treadmill

The final variant call set contained 5,099,945 single-nucleotide polymorphisms (SNPs) and 766,655 insertions-deletions (INDELs) (374,604 insertions; 392,051 deletions, Additional file 2: Figure S3B). The trait-selected lines had much fewer variants than FZTDU and these variants were mostly fixed, whereas FZTDU variants were mostly polymorphic (Fig. 2, Additional file 2: Figure S4, Additional file 3: Table S2). This reduction in genetic diversity could be explained by the fact that the trait-selected lines have been maintained at smaller population sizes and

were relocated with fewer founders (Table 1). In fact, it has been shown that artificial selection for complex traits does not affect the number of segregating sites [3], nor the number of SNP sites and heterozygosity [19]. Interestingly, more than 89% of the variants observed in the trait-selected lines were also detected in the control line FZTDU (Fig. 2A, Additional file 3: Table S2), indicating that despite genetic drift, the control line preserves most of the alleles underlying each selected trait and that it still is a proxy of the original founder population.

(See figure on next page.)

Fig. 1 Phenotypic characteristics of the five trait-selected Dummerstorf mouse lines and the unselected control line FZTDU. Representative subjects showing the impressive litter size of DUK and DUC (**A, B, F, G**) and the considerable body size difference at 6 weeks of age between DU6 (**C, H**) or DU6P (**D, H, I**) and FZTDU. **E** Untrained mice undergoing a treadmill running endurance trial and the increased running performance of DUHLB due to selection (**J**). Stars signify differences ($p < 0.05$) after conducting a t -test between trait-selected lines and FZTDU. Sample sizes are indicated below tick labels (x -axis)

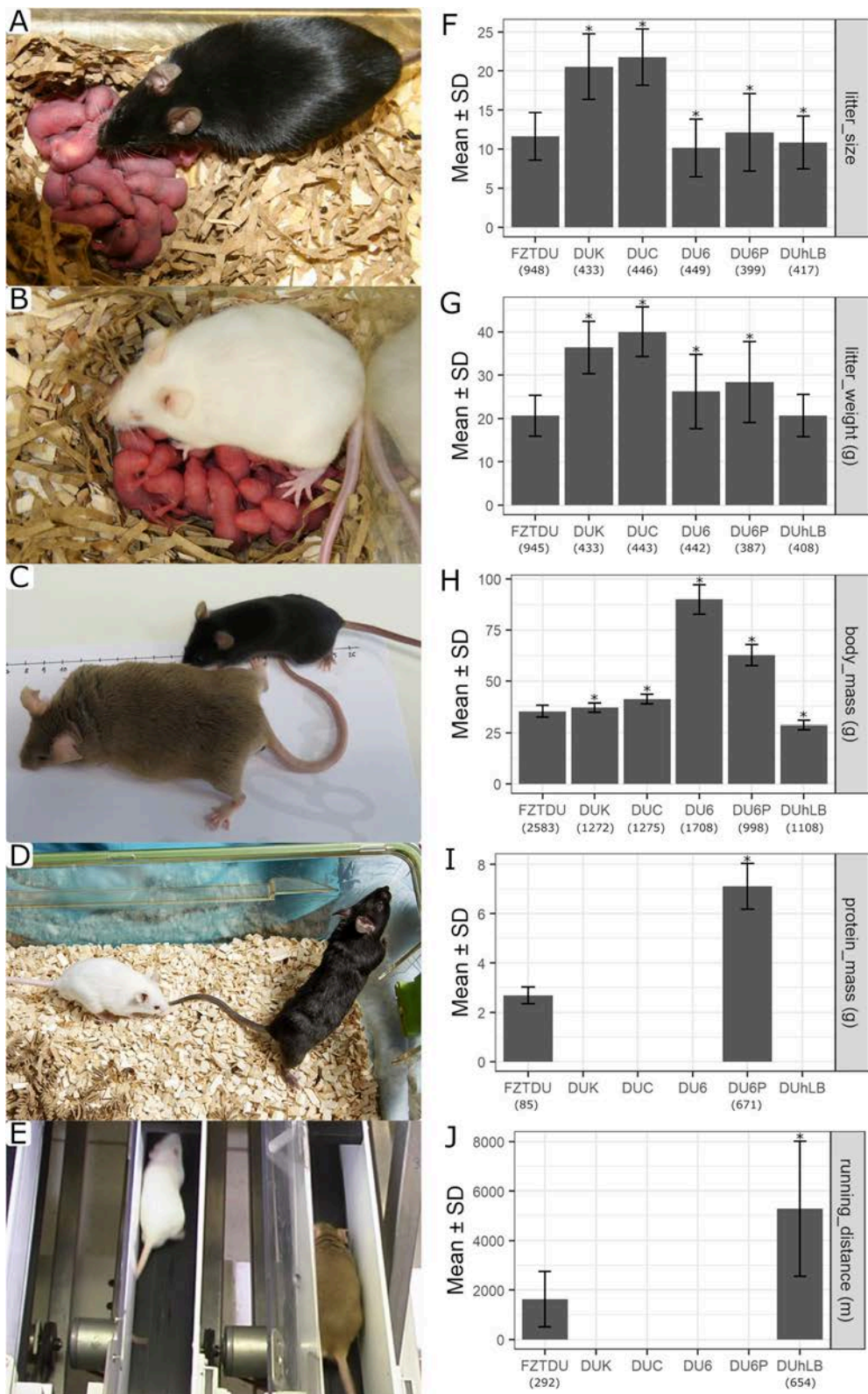
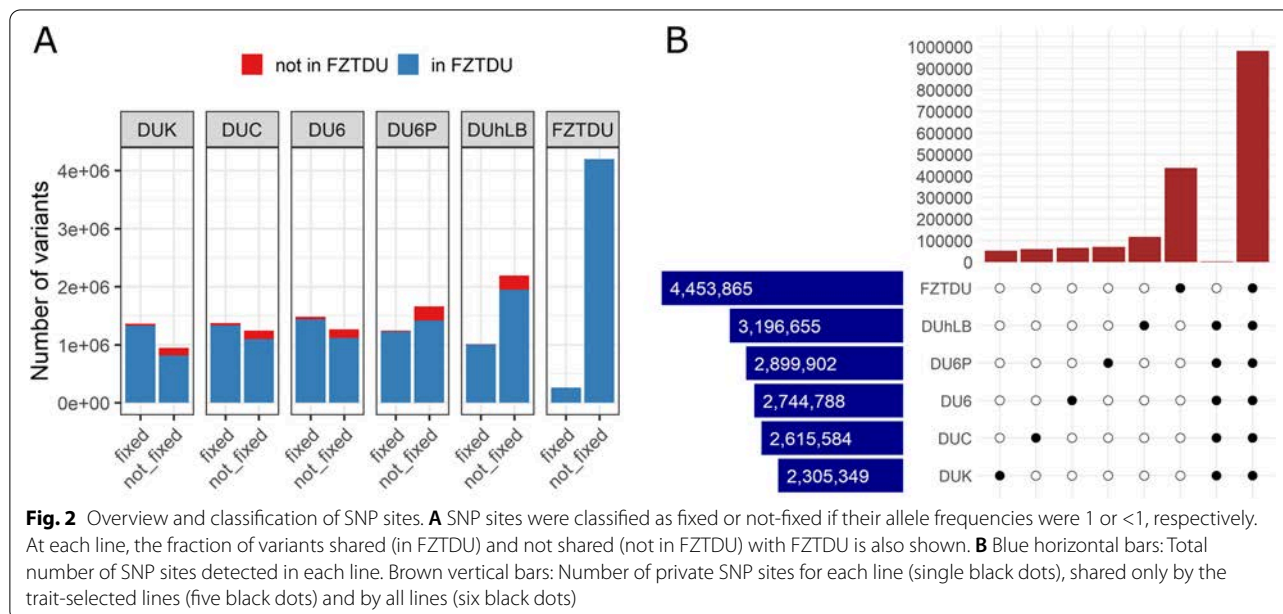


Fig. 1 (See legend on previous page.)

Table 3 Summary metrics WGS data

	Target coverage 30×	Target coverage 5×
Sample size	60	90
Mean number of reads mapped as pairs	90.72%	93.07%
Mean insert size	347.73 bp	401.65 bp
Mean genome-wide coverage	24.08×	7.89×
Mean genome territory covered ≥ 5×	95.57%	71.82%



Most (~90%) INDELs were no longer than 10 bp (Additional file 2: Figure S3A, Additional File 3: Table S4), with slightly more deletions than inversions (Additional file 2: Figure S3B). The proportion of SNPs and INDELs overlapping dbSNP was 95% and 55%, respectively. This discrepancy is not necessarily due to a high number of artefacts in the INDEL set, but rather by the fact that INDELs are a much less characterized type of genetic variant in comparison [42].

The number of alleles present in all six lines was ~1M, but very few alleles were shared by the trait-selected lines exclusively (~3.3K) (Fig. 2B). The lines DU6P and DUhLB were the most polymorphic of the trait-selected lines, followed by DU6. The two fertility lines (DUK, DUC) were the least polymorphic ones (Fig. 2B, Additional file 3: Table S2).

Almost all SNPs and INDELs (~97%) occurred in non-coding regions (introns ~56%; intergenic ~41%). This is not an unexpected outcome considering that only ~2% of the genome codes for proteins and genetic variation is widespread. Inter-genic variants could affect regulatory elements of gene expression, as well as transcripts not

yet described [43], whereas intronic variants could affect gene splicing [44].

Based on assessment of variant annotations, a very small number of variants (20,236 SNPs and 1,801 INDELs) were classified as high-impact and moderate-impact mutations, and could interfere with gene transcription or translation. These “impact variants” were screened for (i) being private for any trait-selected line (Additional file 3: Table S3) and (ii) the functional categories their affected genes belonged to. The number of genes affected by these private “impact variants” was twice as large in DUhLB (1027 genes) than in the other trait-selected lines (465–546 genes). However, there was no obvious coherence between significantly enriched functions and the selected traits (Additional file 4: Data S2).

Runs of homozygosity (RoH) and linkage disequilibrium (LD)

While for the five trait-selected lines, most of the SNP loci (57.5–81.5%) were already fixed for either the reference or the alternative allele, in the control line FZTDU

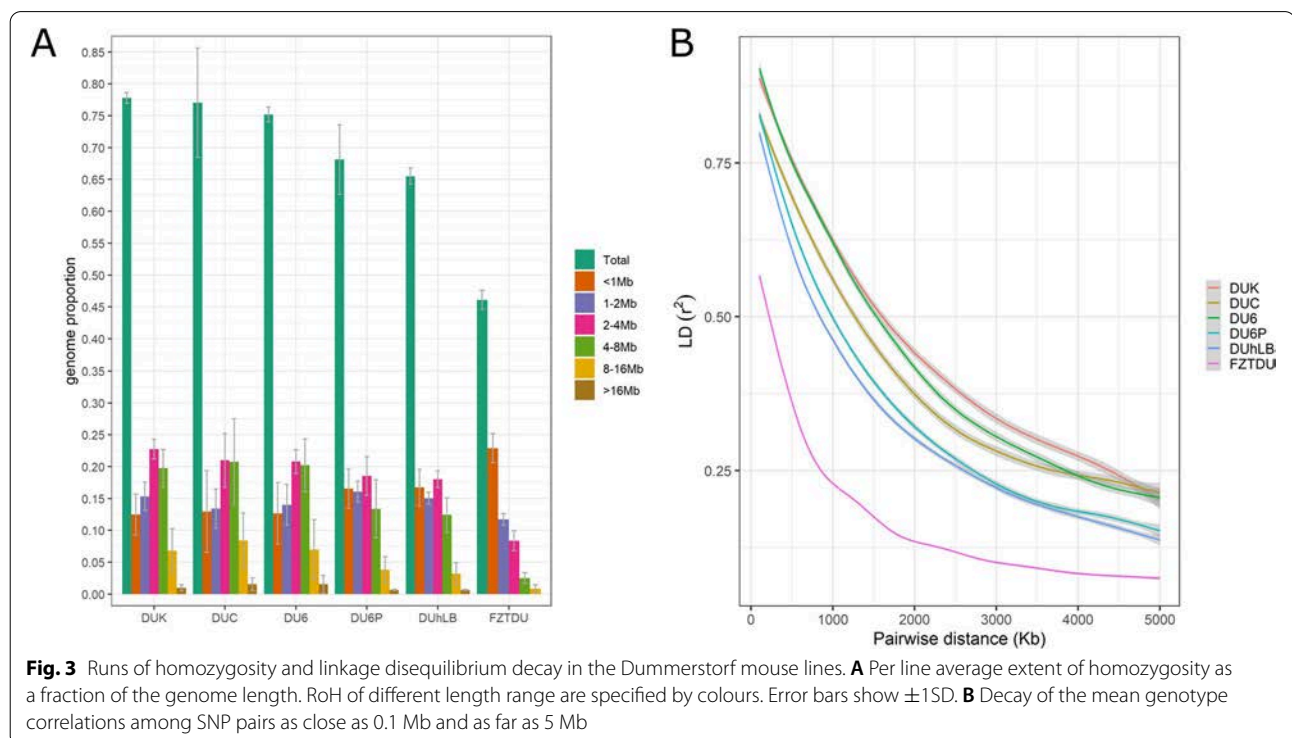
alleles were mostly (>75%) polymorphic (Additional file 2: Figure S5). This disparity was also reflected by the distribution of frequencies for the alternative allele, displaying a “U” shape that is much more pronounced in the trait-selected lines than in the control line (Additional file 2: Figure S6). Genomes of mice from the control line FZTDU also had higher nucleotide diversity (Additional file 2: Figure S7 and S8). Accordingly, RoH covered between ~65 and ~78% (~50% as 1–8 Mb tracts) of the genome length of the trait-selected lines, but only ~45% (~23% as 1–8 Mb tracts) of the genome length of FZTDU (Fig. 3A). Analysing RoH shared among individuals of a population can aid to detect past selection events [45]; however, this is applicable as long as RoH events are rare in the genome (RoH islands), which is not the case here, where RoH are widespread, indicating that the observed degree of homozygosity is the result of a combination of multiple evolutionary forces.

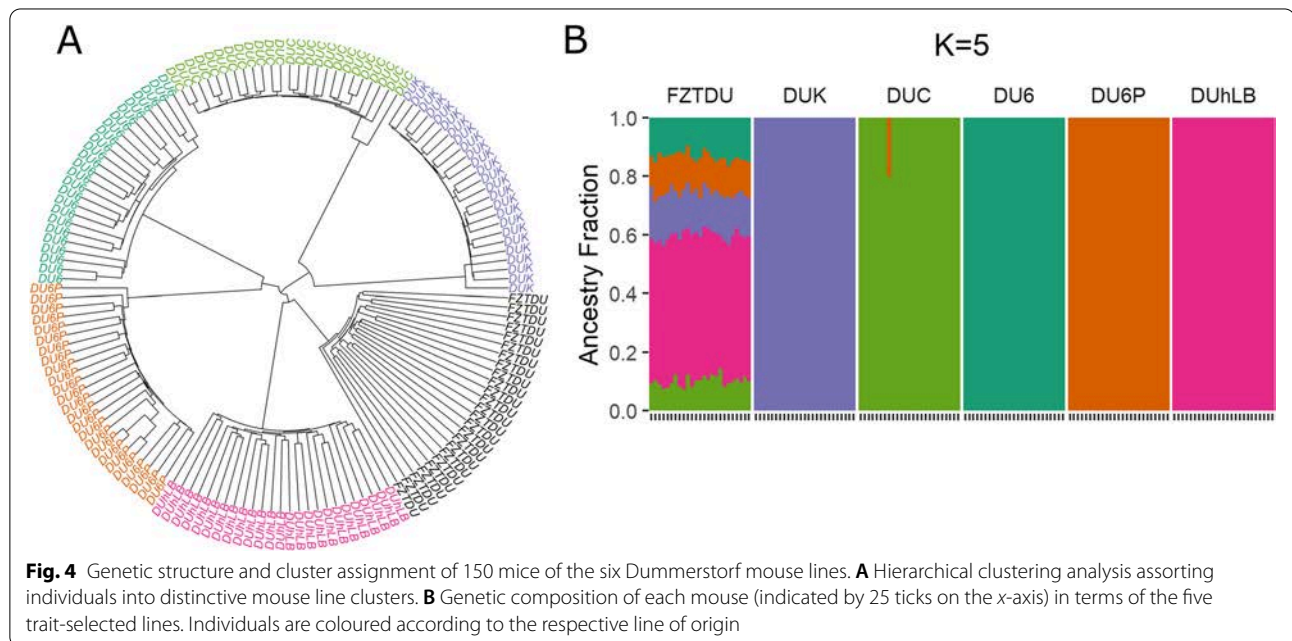
Linkage disequilibrium decay, represented by the genotype correlation (r^2) between pairs of SNP sites within min. 0.1 Mb and max. 5 Mb, can be classified into three patterns with decreasing decay strength; one for the three most homozygous trait-selected lines (DUK, DUC and DU6; upper three lines Fig. 3B), a second for the two least homozygous trait-selected lines (DU6P and DUhLB; middle two lines Fig. 3B) and a third for the unselected line FZTDU (bottom

line Fig. 3B). Overall, r^2 clearly differs between trait-selected lines and FZTDU. Comparable levels of r^2 have been reported in mountain gorillas, in which population decline has led to high levels of inbreeding [46]. Likewise, strong levels of LD have been observed in laboratory mice [47]. However, other populations with high levels of inbreeding, such as dog [48] and horse breeds [49], do not display such strong genotypic correlations, highlighting the impact of the bottleneck in the genetic diversity of the Dummerstorf mouse lines.

Population structure of the Dummerstorf mouse lines

The genetic relationship among the 150 Dummerstorf mice was assessed by hierarchical clustering (HC) and admixture analysis using the 5,099,945 SNPs obtained after variant calling. Samples formed a hierarchical group structure that represented each of the Dummerstorf lines (Fig. 4A). There was no admixture present in the trait-selected lines, except for one DUC animal sharing ancestry with mice from DU6P (Fig. 4B). FZTDU is represented as an admixture of all the trait-selected lines with similarly large contributions of the four older lines and a significantly larger contribution of DUhLB (Fig. 4B). This is expected because this mouse line is the youngest and has had the least number of generations that underwent selection.





Genetic differentiation of the trait-selected lines

Mean genome-wide pairwise genetic differentiation among trait-selected lines estimated with the genetic differentiation index (F_{ST}) ranged from 0.44 to 0.61 (Fig. 5B). The highest level of differentiation was found between either one of the fertility lines and the body mass line DU6 ($F_{ST(DUK-DU6)} = 0.61$ and $F_{ST(DUC-DU6)} = 0.59$; Fig. 5B), followed by the differentiation between the two fertility lines themselves ($F_{ST(DUK-DUC)} = 0.57$; Fig. 5B). Although pairwise genetic differentiation between trait-selected lines and the control line was similar in all comparisons ($F_{ST} \sim 0.3$), it was lowest in the pairwise comparison between the two most polymorphic lines ($F_{ST(DUHLB-FZTDU)} = 0.26$; Fig. 5B). Such strong levels of differentiation occur mainly as a result of reproductive isolation and genetic drift [50]; however, it is expected that a subset of alleles that have arrived to fixation due to selection contribute to genetic differentiation as well. The challenge is thus to sort out which genomic regions contain such beneficial alleles.

Trait-specific regions of genetic differentiation

Genome-wide scans were conducted in order to detect genomic regions of consistent genetic differentiation

between each trait-selected line and FZTDU. The pseudo-line of DUK and DUC combined (FERT) was also included, for a total of six F_{ST} contrasts. Overall, outstanding regions of particularly extreme genetic differentiation were not observed, but rather a uniform genome-wide level of high F_{ST} (Fig. 5A). Choosing genomic regions of interest by focusing on the most differentiated regions (95th or 99th percentile of the F_{ST} distribution) resulted in the detection of multiple loci in every chromosome (Fig. 5A). Because these regions were frequent and did not sufficiently depart from the global level of genetic differentiation to be considered genomic outliers (i.e. max. zF_{ST} : 2.89–3.47, Fig. 5C), a more stringent approach was applied to identify line-specific regions of high genetic differentiation (Fig. 6D and Fig. 7D), while reducing the influence of genetic drift. These regions of distinct genetic differentiation (hereafter referred to as RDDs) appeared simultaneously in the top 5% F_{ST} windows of the target contrast and in the bottom 10% of all the remaining contrasts, occurring close to each other in only a subset of chromosomes and containing multiple genes (Fig. 6A–C, Fig. 7A–C, Additional file 4: Data S3-S14), some of which were related to the selected traits (see below).

(See figure on next page.)

Fig. 5 Genetic differentiation of the Dummerstorf trait-selected lines. **A** Genome-wide scans of genetic differentiation in sliding window mode (size = 50 kb, step = 25 kb) contrasting each trait-selected line to FZTDU. Each window is the average F_{ST} of at least 10 SNPs. **B** Pairwise genomic mean F_{ST} among all six Dummerstorf lines. **C** F_{ST} distribution as z-scores, illustrating the departure of each window from the mean genomic level of genetic differentiation. Dotted lines indicate the 95th (red) and 99th (blue) percentiles and black dots correspond to data points larger than 1.5 the interquartile range (outliers)

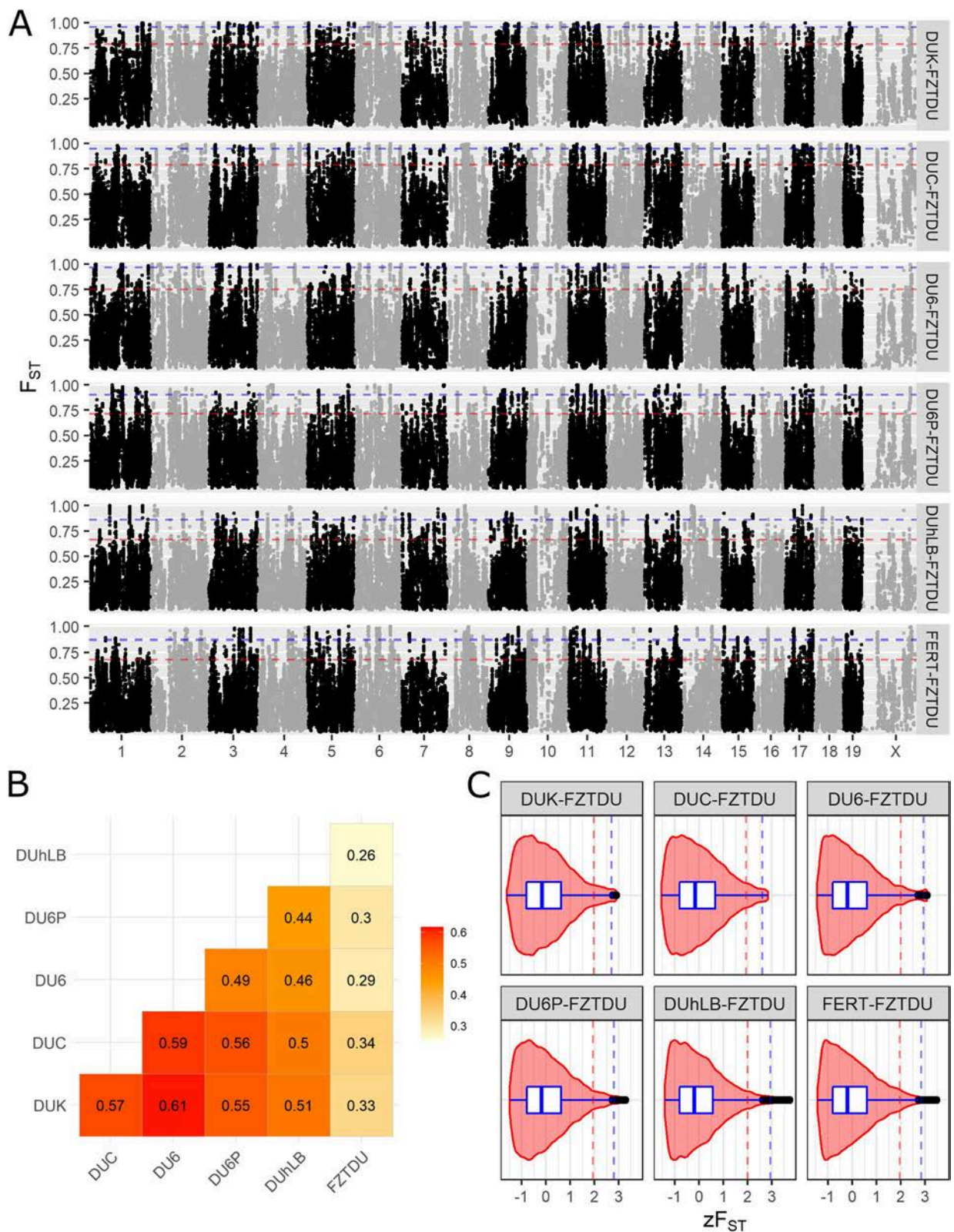
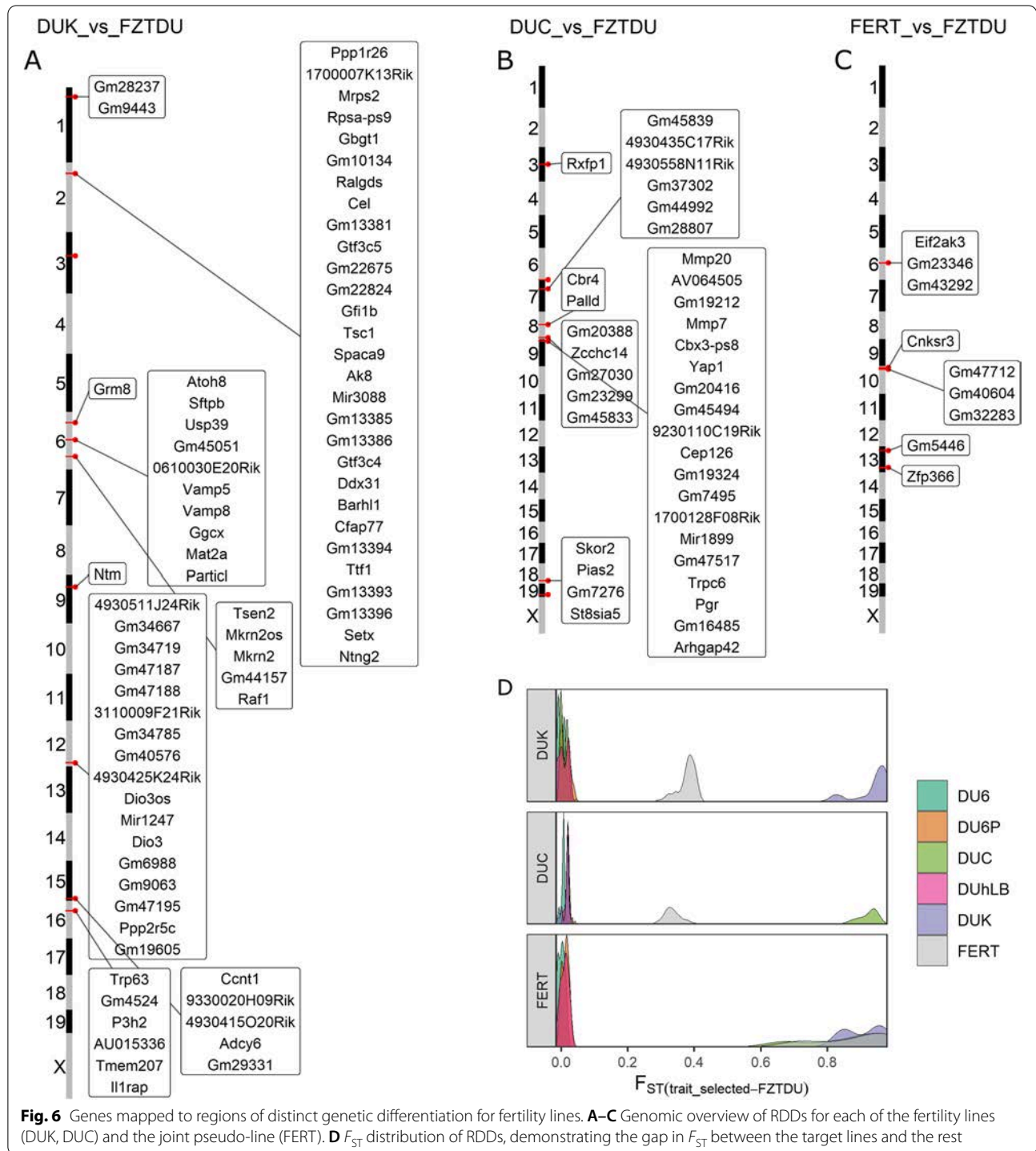


Fig. 5 (See legend on previous page.)



These thresholds were empirically determined based on a similar study comparing two extremely differentiated inbred maize lines [51]. Neutrality simulations were not conducted due to the lack of genetic material from founders and incomplete

pedigrees. This information is critical to identify discrete candidate targets of selection for complex traits, in which selection response occurs gradually and myriads of loci with small effects are expected to be involved [3].

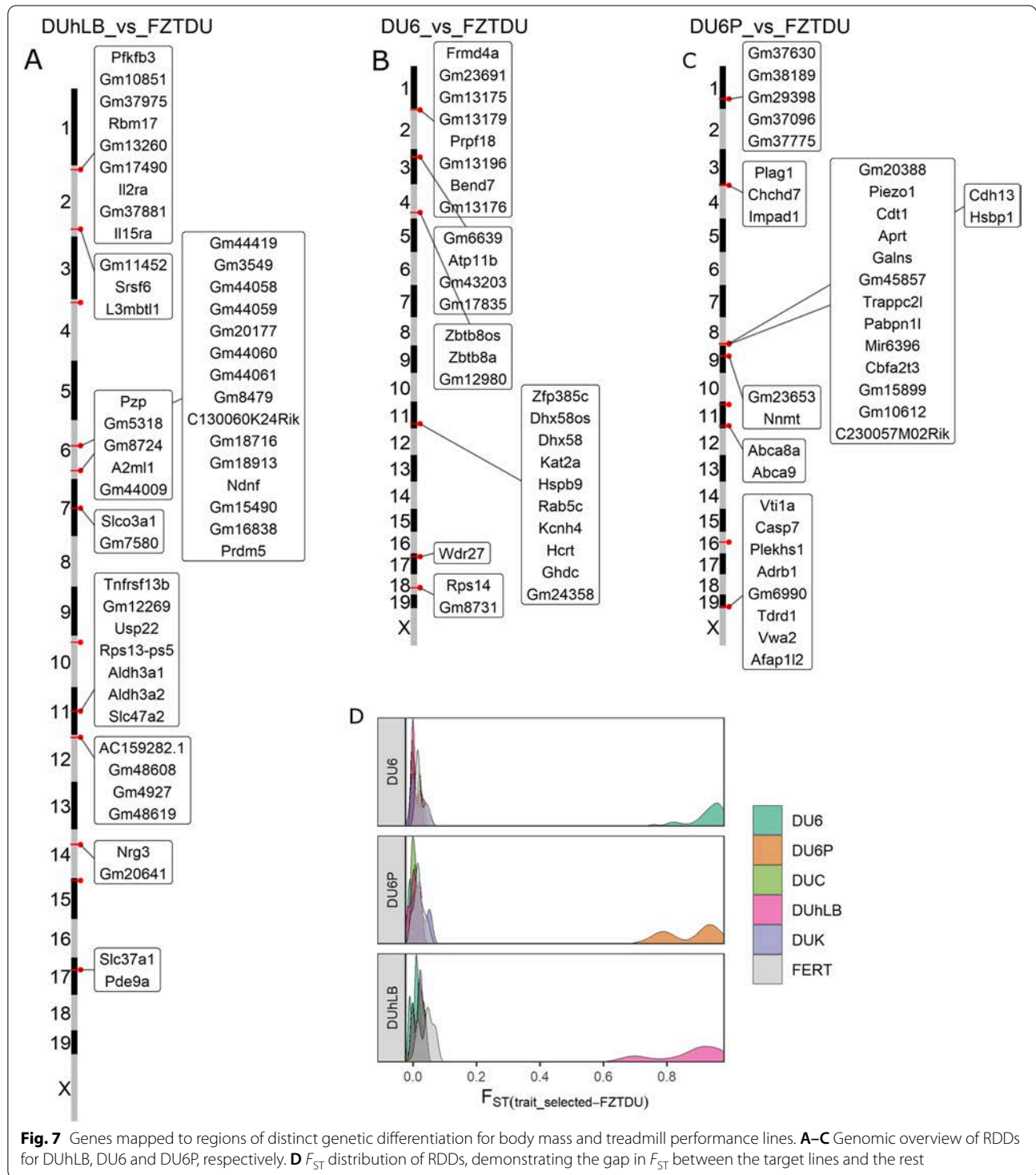


Fig. 7 Genes mapped to regions of distinct genetic differentiation for body mass and treadmill performance lines. **A–C** Genomic overview of RDDs for DUhLB, DU6 and DU6P, respectively. **D** F_{ST} distribution of RDDs, demonstrating the gap in F_{ST} between the target lines and the rest

Line-specific patterns of structural variation (SV)

Despite primarily thought to be deleterious and implicated in disease phenotypes [52, 53], large chromosomal rearrangements such as deletions, duplications and inversions have an important role in local adaptation

and divergence of populations [54]. These structural variants can lead to gene expression differences by disrupting genes and altering gene dosage [55]. Because copy number variation often results in notable phenotypic differences, it is likely a subject to selection during

domestication [56]. For example, genes related to metabolic activity and production traits have been shown to be affected by copy number variation during artificial selection of cattle [57], goats [58] and pigs [59].

After calling and filtering, only duplications, deletions and inversions remained in the final SV data set. Insertions did not occur in enough samples to be included in the analysis. Also, because of the lower detectability in the low sequencing coverage samples, most SVs were found in high coverage samples (Additional file 3: Table S10). Nevertheless, the final SV call set contained the union of good-quality SVs detected in both coverage sets.

SVs were predominantly located in non-coding regions (98%) where they could affect gene expression. Also, SVs (Table 4) were more abundant in the trait-selected lines (deletions (DEL) 5560–4339; duplications (DUP) 48–20; inversions (INV) 1508–530) than in the control line (DEL 3902; DUP 14, INV 605) implying that large genomic rearrangements could contribute to the development of the selected traits. In order to associate SVs to each selected trait, line-specific SVs overlapping protein-coding genes were identified and characterized in greater detail (Additional file 4: Data S15). The total number of these line-specific SVs ranged from 9 (FZTDU) to 36 (DUC), comprising mostly deletions and inversions (Table 4). Most SVs were polymorphic and large length differences were observed between polymorphic and fixed SVs (Additional file 3: Table S6). Fixed line-specific deletions were detected in all lines, whereas duplications were found only in DU6P, and inversions only in DUC, DU6P and DUhLB (Additional file 3: Table S7).

The number of genes affected by fixed line-specific SVs varied from 1 (DUC, DU6P, FZTDU) to 5 (DUK), but went up to more than a thousand for genes affected by large polymorphic inversions (Additional file 3: Table S8). These genes were classified in functional groups based on the biological processes they are associated with (Additional file 3: Table S9). The most gene-rich functional groups are the ones associated with sensory perception,

predominantly olfaction (found in the fertility lines DUK and DUC), followed by “cell cycle and nucleic acid transcription and translation” (in DUC), and “metabolism and energy conversion” (DUC, DU6P).

Genes associated with fertility

Genes detected in RDDs for DUK were enriched for “phospholipase D signalling pathway” (Additional file 3: Table S5). In granulosa cells, phospholipase D activity is stimulated by GnRH, thereby inducing or inhibiting cell differentiation depending on the maturation state of the ovarian follicle [60]. Other genes encode for proteins involved in the ovarian development and maintenance of the primordial follicle reserve (*Tsc1* [61], *Trp63* [62]), in the vascularization of the placenta (*Atoh8* [63]) and facilitate maternal supplied lipids and dietary fat digestion in neonatal mice (*Cel* [64, 65]). Furthermore, DUK shares a fecundity associated region (*Sftpb*, *Usp39*, *Tmem150*, *Rnf181*, *Vamp5*, *Vamp8*, *Cgcx*, *Mat2a*) with Qsi5 mice [18], an inbred mouse line known for its increased litter size, and candidate genes associated with birth rate and male fertility in humans (*Ntm* [66]) and litter size in cattle, goats and pigs (*Dio3* [67–69]). Interestingly, analysis of private SVs detected a 317-bp deletion affecting *Olfr279* (Additional file 4: Data S15). This gene has been associated to mouse male sub-fertility [70] and more generally, olfactory receptors could regulate fertilization [71, 72].

Significantly enriched terms for DUC included “intracellular steroid hormone receptor signalling pathway” (Additional file 3: Table S5), involving progesterone receptor (*Pgr*) carrying a missense mutation, which is fixed in and specific for DUC (Additional file 2: Figure S9B). Progesterone is one of the main steroid hormones regulating reproductive processes and critical for (i.a.) pregnancy maintenance and mammary gland development [73, 74]. It remains to be proven if a connection exists between this missense and potentially deleterious (Sorting Intolerant From Tolerant (SIFT) score = 0.04) mutation and the fact that DUC females display increased

Table 4 Summary of structural variants detected in all mouse lines

	Total				Line-specific-genic			
	DEL	DUP	INV	Total	DEL	DUP	INV	Total
DUK	4633	32	530	5195	11	2	7	20
DUC	5560	48	1248	6856	10	2	24	36
DU6	5025	27	551	5603	11	0	7	18
DU6P	4339	23	2091	6453	9	1	14	24
DUhLB	4614	20	1508	6142	10	1	9	20
FZTDU	3902	14	605	4521	4	0	5	9

levels of progesterone [22]. Interestingly, a Neanderthal missense mutation in *Pgr* associated with increased fertility was recently reported to segregate in human populations [75]. Further candidates in DUC control ovarian follicle development, uterine growth and endometrial angiogenesis during pregnancy (*Yap1* [76], *Rxfp1* [77, 78]). In the context of preparation of the endometrium for implantation and pregnancy and progesterone signaling, the gene *Rrm2* [78] was identified by the structural variation analysis of the DUC genome.

The fertility lines DUK and DUC have been bred according to the same criteria, share the same evolutionary history, and both have been able to more than double the number of pups per litter since the beginning of selection. Despite these commonalities, improved fertility is achieved via different physiological pathways in each line [22]. For example, females from both fertility lines have an increased ovulation rate, but only DUK exhibits follicles containing multiple oocytes; DUC on the other hand shows an increased progesterone level compared to DUK and FZTDU [22]. The scarce number of RDDs in the combined FERT population also illustrates this discrepancy. Candidate RDD and line-specific SV overlapping genes in both fertility lines likely affect the reproductive process on multiple levels such as ovarian physiology, placentation, sex steroid signalling and milk composition.

Genes associated with body size and endurance

Two of the Dummerstorf trait-selected mouse lines have increased their body weight in response to selection. The “giant” DU6 line (selected for body mass at 6 weeks of age) exhibits an obese phenotype [8] while the protein-mass line DU6P (selected for protein mass in the carcass) is lean and muscular [25].

In line with the obese phenotype, DU6 candidate genes overlapping RDDs regulate energy metabolism and food intake (*Hprt* [79]) and are linked to feed efficiency (*Wdr27* [80]) and body composition in other species (*Atp11b* [81]). DU6 mice also exhibit larger bones [21], and the analysis of SVs detected *Smad5*, a modulator of bone formation [82], to be partially overlapped by a heterozygous deletion and a heterozygous inversion. Though DU6 gave origin to DUHi, one of the lines used to detect parallel selected regions (PSRs) for high body weight, none of the RDDs intersected with PSRs [16]. This is partly explained by the fact that DUHi was established after sampling DU6 mice on generation 85 (well before bottleneck, see Table 1) and further maintenance of these animals under inbreeding [15].

Candidate genes in the RDDs for DU6P conform with growth-related major quantitative trait loci found in sheep and are known to influence stature and body size in cattle, pigs and human (*Plag1* [83, 84], *Chchd7*

[83–85], *Impad1* [86]). In line with this, an SV (deletion) was found overlapping *Fam92a*, a gene that is involved in limb development [87]. Further candidates for lean body mass are the RDD overlapping genes *Piezo1* (myotube formation [88, 89]) and *Cdh13* (control of lipid content in developing adipocytes [90–92]).

Finally, genes specific for the endurance line DUhLB participate in lipid metabolism (these animals display faster mobilization of lipids during exercise). Only two DUhLB genes (*Aldh3a1* and *Aldh3a2*, the later containing 3 missense SNPs (Additional file 2: Figure S10C)) caused the significant enrichment of the “Histidine metabolism” and “beta-Alanine metabolism” pathways (Additional file 3: Table S5). The “marathon mice” DUhLB have developed a striking metabolic phenotype characterized by accelerated browning of subcutaneous fat and altered mitochondrial biogenesis in response to selection for high treadmill performance [29]. Likewise, detected RDD candidate genes are involved in the development of brown adipocytes (*Srsf6* [93]), removal of toxic waste products from lipid metabolism (*Aldh3a2* [94]), mobilization of fatty acids, mitochondria content and cristae complexity (*Il15r* [95]) and in the regulation of glycolysis associated to obesity and weight gain (*Pfkfb3* [96, 97]). Moreover, SV analysis detected a ~2.8 kb inversion in *Atp5j* whose overexpression has been shown to counteract exercise-induced cardiac hypertrophy in mice [98]. Interestingly, the genes identified here did not overlap with significantly differentiated genes of the “High Runner” selection experiment [19], highlighting the fact that these two studies produced phenotypically different mice (i.e. DUhLB shows lower running wheel activity compared to controls [31]).

Limitations

There are five main weaknesses in this study. First, due to gaps in pedigree documentation over more than 140 generations, modelling neutrality was not feasible. In turn, the thresholds to evaluate line-specific genetic differentiation were chosen empirically by setting conservative limits that minimize the presence of false positives.

Second, at its origin in 1969, the study was not designed to conduct genomic analyses. Thus, genetic material from the founders is not available. Unfortunately, this and the incomplete pedigree information hamper the detection of signatures of selection. However, the genomic data generated here still allows deriving biological interpretations based on the line-specific patterns of genetic differentiation, which is the subject of this study.

Third, relocation of the mouse lines by embryo transfer resulted in a genetic bottleneck and random fixation events. This further obscures insight into the selection response mechanisms of these mouse lines. Still, the

current strong phenotypic divergence of the lines is the result of long-term selection.

Fourth, except for the fertility lines DUK and DUC, trait-selected lines were not replicated in order to identify overlapping genomic signatures. Interestingly, these two lines are markedly different both physiologically and genetically, despite having the same selection criteria.

Finally, SVs were detected using short pair-end reads (150bp) and this is not an optimal approach for SV discovery. For this, long reads provide much greater accuracy and sensitivity [99, 100].

Conclusions

The genomes of the Dummerstorf trait-selected mouse lines have evolved in response to selective breeding and neutral forces, exhibit low genetic diversity and display distinct patterns of genetic variation. Distinguishing between selection and neutral evolution is a challenging task and will require further research. However, by focusing on regions of distinct genetic differentiation, we were able to identify genes with important functions associated to the selected traits.

Over the span of this selection experiment, traits have improved continuously and have not decayed despite the dramatic loss of genetic diversity within lines. This implies that many of the alleles that contribute to trait improvement have arrived to fixation and that these lines are highly enriched for such alleles. Therefore, a deeper understanding of the genomes of the trait-selected Dummerstorf mouse lines will provide valuable insights into the genetic basis of important polygenic traits and constitutes an unprecedented scientific resource for geneticists, physiologists and the wider biomedical research community.

Methods

Selection history of the Dummerstorf trait-selected mouse lines

The selection experiment started in 1969 (Tables 1 and 2, for more detail see Additional file 1: Supplementary Methods [5, 6, 22, 101–114]) with the establishment of a founder line FZTDU (Forschungszentrum für Tierproduktion Dummerstorf) [5, 6] by systematic crossing of four outbred strains (NMRI orig., Han:NMRI, CFW, CF1) and four inbred strains (CBA/Bln, AB/Bln, C57BL/Bln, XVII/Bln). From FZTDU, five lines were established through selective breeding: two lines were selected for increased litter size (DUK and DUC), one for increased body mass (DU6), and one each for protein mass (DU6P) and treadmill running endurance (DUHLB) (Table 2, Fig. 1, Additional file 2: Fig. S1).

Sample collection and whole genome sequencing

All animal procedures were performed in accordance with national and international guidelines and approved by the Animal Protection Board of the Institute for Farm Animal Biology. Genomic DNA was purified from tail biopsy samples using QIAamp DNA Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's recommendations. A total of 25 females per line (150 animals in total) were sampled at two different time-points (Table 1). For the first time-point, 10 females per line with the lowest kinship coefficient were chosen. Kinship was determined using the programme INBREED implemented in the software SAS/STAT® (v9.4, SAS Institute Inc., USA). For the second time-point, 15 females per line were chosen at random since the kinship coefficient is similar among subjects of the same line. The study was originally designed with 10 females per line, sequenced at high coverage (target 30×, time-point 1) to capture as much line-specific genetic variability as possible. Due to the low genetic variability in each line resulting from the preliminary data analysis, 15 additional females per line were sequenced. As this was intended to verify the low degree of genetic variability at the initially detected loci and to increase the number of total observations for each line, the samples of the second batch were sequenced with a lower sample coverage (target: 5×, time-point 2).

Library preparation and sequencing were carried out at the Competence Centre for Genomic Analysis (Kiel). Paired-end sequencing libraries were prepared using the TruSeq Nano DNA Library Prep kit following the manufacturer's specifications (Illumina Inc., San Diego, CA, USA). Out of the 150 libraries, 60 were sequenced on a HiSeq 4000 platform (Illumina Inc.), and 90 samples were sequenced on a NovaSeq 6000 (Illumina Inc.) platform. The target coverage was 30× (high coverage set) and 5× (low coverage set), respectively. Read length was 151 nucleotides. Samples sequenced at 30× ($n = 60$) were distributed in 9 lanes for a total of 540 pairs of read files. Ten of those samples had to be supplemented with extra sequencing data due to not reaching the expected 30× coverage. Samples sequenced at 5× were not lane-distributed, amounting to 90 pairs of read files. In total, 640 pairs of read files were produced. Sample-wise WGS data is summarized in Additional file 4: Data S1.

Analysis of WGS data

Adapter removal and quality trimming were done using Trimmomatic v0.38 [115] for HiSeq reads and FASTP v0.19.6 [116] for NovaSeq reads. Read quality was evaluated before and after processing with FastQC v0.11.5 [117]. Reads were aligned to the mouse genome build GRCm38.p6 [118, 119] from Ensembl version 93 [120]

using the Burrow-Wheeler Aligner software in MEM mode (BWA-MEM) [121] coupled with SAMtools v1.5 [122] in order to store alignments as Binary Alignment Map (BAM) files. Per sample BAM files were processed sequentially with Picard tools [123] by adding read group information (*AddOrReplaceReadGroups*), merging alignments from different read groups (*MergeSamFiles*), and by sorting (*SortSam*) and marking duplicated (*MarkDuplicates*) reads.

Short variant calling and annotation

Short variants were detected according to GATK's best practices for germline short variant discovery (GATK v 4.0.6.0) [124–127]. Systematic errors in base quality were corrected using *BaseRecalibrator* and dbSNP [128] version 150 for *Mus musculus* (Ensembl version 93 [129]). For each sample, variants were called with *HaplotypeCaller* and then combined with *GenomicsDBImport*. Joint genotyping was done with *GenotypeGVCFs* and then only bi-allelic variants (SNPs and INDELS) were retained. Filtering was applied separately for SNPs and INDELS. Site-level filtering was done following the Variant Quality Score Recalibration (VQSR) procedure. This comprised an internal variant set used as truth-training resource, created after stringent site-level filtering of the bi-allelic variants obtained from joint genotype calling, plus an external pre-filtered training variant set provided by the Mouse Genomes Project (MGP version 5 [130]). Variants were genotyped as missing if the depth of coverage (DP) was either too low (<4), too high (3 standard deviations higher than the sample mean depth) or if the genotype quality (GQ) was too low (<20). INDELS overlapping microsatellites [131] were excluded. The final set consisted of variants present in at least 15 samples per line (except for DU6 that had a lower coverage, so this threshold was lowered to 12 samples). Annotations were done using SnpEff v4.3t [132] and missense mutations were further evaluated with Ensembl Variant Effect Predictor (VEP) v.101.0 [103] to obtain their corresponding SIFT scores [133] and to predict amino acid changes affecting protein function.

Structural variant calling and annotation

Processed BAM files used for short variant calling were also used to detect large structural variants (SVs). SVs correspond to deletions, duplications, insertions, inversions and translocations of at least 50 bp in size [134]. Because of the considerable difference in coverage of the two sequence data sets, this was done independently for the high and the low coverage set.

Three SV callers (Manta v.1.6.0 [110], Whamg v.1.7.0 [111] and Lumpy v.0.2.13 [112]) were applied per line and per coverage set yielding six call sets per line (see

Additional file 1 for more detailed information). Specific filters were applied depending upon the call set. SVs detected by Manta were site-filtered by excluding SVs with poor mapping quality (Mapping Quality (MAPQ) < 30) or with excessive coverage (>3 × the median chromosome depth) that could be due to reads originated from low complexity regions. For each sample, only SVs with GQ ≥ 20 and read depth ≥ 5× were accepted. Whamg SV calls with sizes <50 bp and >2 Mb were filtered out to improve call accuracy. Here too, only calls with read depth ≥ 5× were accepted. Calls with GQ < 20 were filtered out. To reduce the number of false positive calls, high cross-chromosomal mappings were excluded, as Whamg is aware of but does not specifically call translocations. Likewise, SVs in poorly mapped regions were also removed. Lumpy SV calls for which supporting evidence (FORMAT/SU field) was below 5 (SU<5) were excluded, as well as SV calls with GQ<20. Since both Whamg and Lumpy do not have a built-in genotyping module, SV call sets were genotyped with Svtlyper v0.7.1 [101] prior filtering for genotype quality. For each line and coverage set, SVs called by at least two SV callers were merged using Survivor v.1.0.7 [102] and kept if they were found in at least 10 samples. The final set consisted of the union of SVs detected in the high and low coverage read sets. We then intersected SV calls among all six mouse lines to obtain SVs private for each line (line-specific) and SVs shared among lines. SVs were annotated with Ensembl's VEP [103] focusing on variants affecting protein-coding genes with the maximum SV size set to 200 Mb.

Functional classification was conducted after thorough literature and database search (OrthoDB v10 [104], UniProt [107], NCBI Entrez gene [105]), plus Gene Ontology enrichment analysis (Shiny GO [106], false discovery rate [FDR] < 0.05). To further minimize false positives, SV calls overlapping gaps and high coverage regions (>80×) in the reference genome assembly were filtered out.

Population genetics analysis

Genetic structure among all 150 samples was assessed using HC analysis and genetic admixture. HC was computed using SNPRelate v1.22.0 [135]. The ape v5.0 package was used for visualization of HC results [136]. Genetic admixture was estimated with ADMIXTURE v1.3.0 [137] after transforming the Variant Calling File (VCF) file into a BED file using PLINK v2.00a2LM [138, 139]. Linkage disequilibrium (LD) was evaluated after thinning the main VCF file with vcftools v0.1.13 [140] retaining sites at least 100 kb apart and then calculating r^2 within windows of 5 Mb using PLINK v2.00a2LM [138, 139]. Runs of homozygosity were estimated for each sample using the RoH extension [141] in SAMtools/BCFtools

v1.5 [122]. For this, allele frequencies at each SNP site and a constant recombination rate (average recombination rate mouse genome: 0.51 cM/Mb [142, 143]) were provided. These parameters, plus the genotype likelihoods stored in the VCF containing the sample, allow to identify RoHs using a hidden Markov model.

Genetic differentiation and diversity analysis

The genomes of the trait-selected lines were compared to the neutrally evolving control line (FZTDU). For this, genetic differentiation was estimated using the F_{ST} index [144] in sliding window mode (size = 50 kb, step = 25 kb, min 10 SNPs) using vcfTools v0.1.13 [140]. Since F_{ST} calculations are based on allele counts and not read counts, differences in depth between low and high coverage samples are not expected to have a direct effect in the estimation of genetic differentiation. The average number of SNP sites per window was ~ 125 (Additional file 3: Table S11). At each window, the arithmetic mean of the SNP-specific F_{ST} was calculated and then transformed into z -scores to represent its departure from the genomic mean. Additionally, all samples of the two fertility lines (DUK and DUC) were combined (pseudo-line: FERT) and compared to FZTDU as well. Since autosomes and the X-chromosomes have different effective population sizes, the X-chromosome was standardized individually. In order to identify RDDs, F_{ST} windows appearing simultaneously in the 95th percentile of a given contrast and in the bottom 10th percentile of all other contrast were identified. These thresholds are not derived from modeling neutrality, rather they were chosen empirically based on a previous study [51] and after testing multiple combinations of ≥ 95 th percentiles and ≤ 10 th percentiles, choosing the combination in which RDDs could be found in all contrasts. The upper threshold is suitable to evaluate genetic differentiation [49, 145, 146], while the bottom threshold ensures that there is practically no genetic differentiation between any of the other trait-selected lines and the control line (Fig. 6D and Fig. 7D). Genome-wide diversity patterns were assessed by measuring the nucleotide diversity (π) [147] in sliding windows of 50 kb size (step size = 25 kb) using vcfTools v0.1.13 [140].

Gene annotation and enrichment analysis

Genes overlapping RDDs were identified using GenomicRanges [148] and Ensembl 93's [120] *Mus musculus* gene set. In order to sort out the most relevant genes for each of the selected traits, thorough inspection of functional annotations, literature and SNP effects was conducted. This also included testing for enrichment of Gene Ontology Biological Processes (GOBP) [149, 150] and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [151–153] using WebGestalt [154–157] using the whole

genome as reference set. A FDR threshold of 10% was used as cutoff for significant enrichment of a term or pathway. Finally, genes in quantitative trait loci (QTLs) were identified by finding overlaps with QTL data compiled in the Mouse Genome Database [158, 159].

Data handling and visualization

Data processing and visualizations were done using R [160] and the tidyverse package [161].

Abbreviations

FZT: Forschungszentrum für Tierproduktion Dummerstorf; FBN: Research Institute for Farm Animal Biology; DUK: Fertility mouse line 1; DUC: Fertility mouse line 2; DU6: Body mass mouse line; DU6P: Protein mass mouse line; DUhLB: Treadmill performance mouse line; FZTDU: Control mouse line; BP: Breeding pairs; WGS: Whole Genome Sequencing; SNP: Single-nucleotide polymorphism; INDEL: Insertion deletion; RoH: Runs of homozygosity; LD: Linkage disequilibrium; HC: Hierarchical clustering; F_{ST} : Genetic differentiation index; RDD: Region of distinct genetic differentiation; SV: Structural variant; DEL: Deletion; DUP: Duplication; INV: Inversion; SIFT: Sorting Intolerant From Tolerant; FERT: Fertility pseudo-population composed of DUK and DUC; BLUP: Best linear unbiased prediction; VQS: Variant Quality Score Recalibration; MGP: Mouse Genomes Project; GQ: Genotype quality; DP: Genotype depth; BAM: Binary Alignment Map; VCF: Variant Calling File; VEP: Ensembl Variant Effect Predictor; MAPQ: Mapping quality; FDR: False discovery rate; GOBP: Gene Ontology Biological Processes; KEGG: Kyoto Encyclopedia of Genes and Genomes; QTL: Quantitative trait loci.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12915-022-01248-9>.

Additional file 1. Establishment of the Dummerstorf mouse lines, Structural Variant Calling.

Additional file 2: Figure S1. Response to selection throughout the selection experiment. **Figure S2.** Proportion of litters supplying parents for the next generation. **Figure S3.** Distribution of INDEL lengths. **Figure S4.** Number of private and shared INDELS among lines. **Figure S5.** SNP allele frequency state classification. **Figure S6.** Alternative allele frequency distribution. **Figure S7.** Nucleotide diversity (π) distribution in the Dummerstorf mouse lines. **Figure S8.** Example of one chromosome representative of the level of genetic diversity observed in the Dummerstorf mouse lines. **Figure S9.** Allele frequency heatmap of non-synonymous mutations in RDD genes. **Figure S10.** Allele frequency heatmap of non-synonymous mutations in RDD genes.

Additional file 3: Table S1. Alternative names of the Dummerstorf mouse lines. **Table S2.** Number of SNP and INDEL sites discovered in each line. **Table S3.** Number of private variants with predicted high/moderate effects according to SnpEff. **Table S4.** Counts per length up to the 90% most frequent INDELS sorted in decreasing order of frequency. **Table S5.** Significantly enriched terms based on RDD gene lists. **Table S6.** Proportion of line-specific fixed and polymorphic structural variants in genic regions. **Table S7.** Types and lengths of line-specific fixed structural variants in genic regions. **Table S8.** Number of genes affected by line-specific fixed and polymorphic structural variants. **Table S9.** Number of genes in functional groups affected by line-specific structural variants. **Table S10.** Summary of structural variants detected in low and high coverage variant calling sets for each mouse line. **Table S11.** Number of SNP sites per window analysed with FST.

Additional file 4: Data S1. WGS data overview. **Data S2.** Significantly enriched terms for genes affected by line-specific SNPs and/or INDELS with high/moderate impact. **Data S3.** Genomic information of regions of

distinct genetic differentiation for DUK. **Data S4.** Genomic information of regions of distinct genetic differentiation for DUC. **Data S5.** Genomic information of regions of distinct genetic differentiation for DU6. **Data S6.** Genomic information of regions of distinct genetic differentiation for DU6P. **Data S7.** Genomic information of regions of distinct genetic differentiation for DUhLB. **Data S8.** Genomic information of regions of distinct genetic differentiation for FERT. **Data S9.** Genes in regions of line-specific genetic differentiation associated to increased fertility (DUK). **Data S10.** Genes in regions of line-specific genetic differentiation associated to increased fertility (DUC). **Data S11.** Genes in regions of line-specific genetic differentiation associated to increased body mass (DU6). **Data S12.** Genes in regions of line-specific genetic differentiation associated to increased lean body mass (DU6P). **Data S13.** Genes in regions of line-specific genetic differentiation associated to increased treadmill endurance (DUhLB). **Data S14.** Genes in regions of specific genetic differentiation for the pseudo-line FERT (DUK and DUC combined). **Data S15.** Gene-spanning line-specific structural variants.

Acknowledgements

We particularly thank the staff of the FBN Service Group Lab Animal Facility for excellent animal care and cooperation: Benita Lucht, Karin Ullerich, Sabine Maibohm, Ines Müntzel, Hildburg Meyer. And furthermore we thank Erika Wyrwat for the distinguished data base management with the historical mouse data.

Authors' contributions

SEP conducted bioinformatic and population genomic analysis, interpreted the data and wrote the manuscript. HR provided genomic DNA material, interpreted the data and advised population genomic analysis. ML supervised mouse breeding, sample collection and phenotypic records. NR conceived the study, interpreted the data and supervised population genomic analysis and edited the manuscript. LD conducted structural variation analysis and wrote the respective manuscript part accordingly. JF conceived the study, interpreted the data, supervised population genomic analysis and edited the manuscript. SQ advised population genomic analysis and provided suggestions to improve the manuscript. JW interpreted the data and edited the manuscript. SF conducted whole genome sequencing. GHS conducted whole genome sequencing and advised bioinformatics analysis. JS conceived the study, interpreted the data and wrote the manuscript. All authors have reviewed and approved the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. This study was funded by the Leibniz Collaborative Excellence programme (K52/2017, SOS-FERT). This work was also supported by the DFG Research Infrastructure NGS_CC (project 407495230) as part of the Next Generation Sequencing Competence Network (project 423957469).

Availability of data and materials

The datasets generated and analysed during the current study are available in the European Nucleotide Archive (raw sequencing data; accession: PRJEB44248 [162]) and in the European Variation Archive (variant calling files; accession: PRJEB45961 [163]). Scripts used to generate the results of this publication are available in [164].

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details

¹Institute of Reproductive Biology, Research Institute for Farm Animal Biology (FBN), Dummerstorf, Germany. ²Institute of Genome Biology, Research Institute for Farm Animal Biology (FBN), Dummerstorf, Germany. ³Institute of Genetics and Biometry, Research Institute for Farm Animal Biology (FBN), Dummerstorf, Germany. ⁴Department of Evolutionary Genetics, Research Institute for Zoo and Wildlife Research (IZW), Berlin, Germany. ⁵University of Potsdam, Institute for Biochemistry and Biology, Potsdam, Germany. ⁶Institute of Clinical Molecular Biology (IKMB), Kiel, Germany. ⁷Department of Reproduction Biology, Research Institute for Zoo and Wildlife Research (IZW), Berlin, Germany.

Received: 14 July 2021 Accepted: 7 February 2022

Published online: 21 February 2022

References

1. Conner JK. Artificial Selection. In: Kliman R, editor. *Encyclopedia of Evolutionary Biology*. Oxford: Academic Press; 2016. p. 107–13.
2. Kukekova AV, Johnson JL, Xiang X, Feng S, Liu S, Rando HM, et al. Red fox genome assembly identifies genomic regions associated with tame and aggressive behaviours. *Nat Ecol Evol*. 2018;2:1479–91.
3. Castro JP, Yancoskie MN, Marchini M, Belohlavy S, Hiramatsu L, Kučka M, et al. An integrative genomic analysis of the Longshanks selection experiment for longer limbs in mice. *Elife*. 2019;8:e42014.
4. Boulding EG. *Experimental evolution: concepts, methods, and applications of selection experiments*. 1st ed. Garland T, Rose MR, editors. Berkeley, CA: University of California Press; 2009.
5. Schueler L. Mouse strain Fzt:DU and its use as model in animal breeding research. *Arch für Tierzucht (Archives Anim Breeding)*. 1985;28:357–63.
6. Dietl G, Langhammer M, Renne U. Model simulations for genetic random drift in the outbred strain Fzt: DU. *Arch für Tierzucht (Archives Anim Breeding)*. 2004;47:595–604.
7. Langhammer M, Michaelis M, Hartmann MF, Wudy SA, Sobczak A, Nürnberg G, et al. Reproductive performance primarily depends on the female genotype in a two-factorial breeding experiment using high-fertility mouse lines. *Reproduction*. 2017;153:361–8.
8. Renne U, Langhammer M, Brenmoehl J, Walz C, Zeissler A, Tuhscherer A, et al. Lifelong obesity in a polygenic mouse model prevents age- and diet-induced glucose intolerance—obesity is no road to late-onset diabetes in mice. *PLoS One*. 2013;8:e79788.
9. Brenmoehl J, Walz C, Renne U, Ponsuksili S, Wolf C, Langhammer M, et al. Metabolic adaptations in the liver of born long-distance running mice. *Med Sci Sport Exerc*. 2013;45:841–50.
10. Ohde D, Moeller M, Brenmoehl J, Walz C, Ponsuksili S, Schwerin M, et al. Advanced running performance by genetic predisposition in male Dummerstorf marathon mice (DUhTP) reveals higher sterol regulatory element-binding protein (SREBP) related mRNA expression in the liver and higher serum levels of progesterone. *PLoS One*. 2016;11:e0146748.
11. Holt M, Nicholas FW, James JW, Moran C, Martin ICA. Development of a highly fecund inbred strain of mice. *Mamm Genome*. 2004;15:951–9.
12. Bayon Y, Fuente L, Primitivo FS. Selection for increased and decreased total number of young born in the first three parities in mice. *Genet Sel Evol*. 1988;20:259–66.
13. Joakimsen Ø, Baker RL. Selection for Litter Size in Mice. *Acta Agric Scand*. 1977;27:301–18.
14. Ribeiro EL, van Engelen MA, Nielsen MK. Embryonal survival to 6 days in mice selected on different criteria for litter size. *J Anim Sci*. 1996;74:610–5.
15. Bünger L, Laidlaw A, Bulfield G, Eisen EJ, Medrano JF, Bradford GE, et al. Inbred lines of mice derived from long-term growth selected lines: unique resources for mapping growth genes. *Mamm Genome*. 2001;12:678–86.
16. Chan YF, Jones FC, McConnell E, Bryk J, Bünger L, Tautz D. Parallel selection mapping using artificially selected mice reveals body weight control loci. *Curr Biol*. 2012;22:794–800.
17. Schwartz NL, Patel BA, Garland T, Horner AM. Effects of selective breeding for high voluntary wheel-running behavior on femoral nutrient canal size and abundance in house mice. *J Anat*. 2018;233:193–203.

18. Wei J, Ramanathan P, Thomson PC, Martin IC, Moran C, Williamson P. An integrative genomic analysis of the superior fecundity phenotype in QSI5 mice. *Mol Biotechnol*. 2013;53:217–26.
19. Hillis DA, Yadgary L, Weinstock GM, Pardo-Manuel de Villena F, Pomp D, Fowler AS, et al. Genetic basis of aerobically supported voluntary exercise: results from a selection experiment with house mice. *Genetics*. 2020;216:781–804.
20. Langhammer M, Wyrwat E, Michaelis M, Schön J, Tuchscherer A, Reinsch N, et al. Two mouse lines selected for large litter size display different lifetime fecundities. *Reproduction*. 2021;161:721–30.
21. Müller-Eigner A, Sanz-Moreno A, De-Diego I, Venkatasubramani AV, Langhammer M, Gerlini R, et al. Dietary intervention improves health metrics and life expectancy of the genetically obese DU6 (Titan) mouse. *bioRxiv*. 2021. <https://doi.org/10.1101/2020.05.11.088625>.
22. Langhammer M, Michaelis M, Hoeflich A, Sobczak A, Schoen J, Weitzel JM. High-fertility phenotypes: two outbred mouse models exhibit substantially different molecular and physiological strategies warranting improved fertility. *Reproduction*. 2014;147:427–33.
23. Michaelis M, Sobczak A, Koczan D, Langhammer M, Reinsch N, Schön J, et al. Testicular transcriptional signatures associated with high fertility. *Reproduction*. 2018;155:219–31.
24. Meng J, Mayer M, Wyrwat E, Langhammer M, Reinsch N. Turning observed founder alleles into expected relationships in an intercross population. *G3 Genes, Genomes, Genet*. 2019;9:889–99.
25. Bünger L, Renne U, Dietl G, Kuhla S. Long-term selection for protein amount over 70 generations in mice. *Genet Res*. 1998;72:93–109.
26. Bünger L, Renne U, Buis RC. Body weight limits in mice - long term selection and single genes. In: Reeve ECR, editor. *Chicago: Fitzroy Dearborn*; 2001. p. 337–60.
27. Falkenberg H, Langhammer M, Renne U. Comparison of biochemical blood traits after long-term selection on high or low locomotory activity in mice. *Arch Anim Breed*. 2000;43:513–22.
28. Ohde D, Brenmoehl J, Walz C, Tuchscherer A, Wirthgen E, Hoeflich A. Comparative analysis of hepatic miRNA levels in male marathon mice reveals a link between obesity and endurance exercise capacities. *J Comp Physiol B Biochem Syst Environ Physiol*. 2016;186:1067–78.
29. Brenmoehl J, Ohde D, Albrecht E, Walz C, Tuchscherer A, Hoeflich A. Browning of subcutaneous fat and higher surface temperature in response to phenotype selection for advanced endurance exercise performance in male DUHTP mice. *J Comp Physiol B Biochem Syst Environ Physiol*. 2017;187:361–73.
30. Brenmoehl J, Walz C, Spitschak M, Wirthgen E, Walz M, Langhammer M, et al. Partial phenotype conversion and differential trait response to conditions of handbreeding in mice. *J Comp Physiol B Biochem Syst Environ Physiol*. 2018;188:527–39.
31. Brenmoehl J, Ohde D, Walz C, Langhammer M, Schultz J, Hoeflich A. Analysis of activity-dependent energy metabolism in mice reveals regulation of mitochondrial fission and fusion mRNA by voluntary physical exercise in subcutaneous fat from male marathon mice (DUHTP). *Cells*. 2020;9:2697.
32. Walz C, Brenmoehl J, Trakooljul N, Noce A, Caffier C, Ohde D, et al. Control of protein and energy metabolism in the pituitary gland in response to three-week running training in adult male mice. *Cells*. 2021;10:736.
33. Walz M, Chau L, Walz C, Sawitzky M, Ohde D, Brenmoehl J, et al. Overlap of Peak Growth Activity and Peak IGF-1 to IGF1R Ratio: Delayed increase of IGF1R versus IGF-1 in serum as a mechanism to speed up and down postnatal weight gain in mice. *Cells*. 2020;9:1516.
34. Vanselow J, Kucia M, Langhammer M, Koczan D, Rehfeldt C, Metges CC. Hepatic expression of the GH/JAK/STAT/IGF pathway, acute-phase response signalling and complement system are affected in mouse offspring by prenatal and early postnatal exposure to maternal high-protein diet. *Eur J Nutr*. 2011;50:611–23.
35. Kucia M, Langhammer M, Goers S, Albrecht E, Hammon HM, Nrnberg G, et al. High-protein diet during gestation and lactation affects mammary gland mRNA abundance, milk composition and pre-weaning litter growth in mice. *Animal*. 2011;5:268–77.
36. Vanselow J, Kucia M, Langhammer M, Koczan D, Metges CC. Maternal high-protein diet during pregnancy, but not during suckling, induced altered expression of an increasing number of hepatic genes in adult mouse offspring. *Eur J Nutr*. 2016;55:917–30.
37. Schüller L, Renne U, Bünger L. Selection for litter weight on the 21st day after long-term selection for first litter performance in laboratory mice. *J Anim Breed Genet*. 1990;107:161–8.
38. Spitschak M, Langhammer M, Schneider F, Renne U, Vanselow J. Two high-fertility mouse lines show differences in component fertility traits after long-term selection. *Reprod Fertil Dev*. 2007;19:815.
39. Vanselow J, Nurnberg G, Koczan D, Langhammer M, Thiesen H-JJ, Reinsch N, et al. Expression profiling of a high-fertility mouse line by microarray analysis and qPCR. *BMC Genomics*. 2008;9:307.
40. Alm H, Kuhlmann S, Langhammer M, Tuchscherer A, Torner H, Reinsch N. Occurrence of polyovular follicles in mouse lines selected for high fecundity. *J Reprod Dev*. 2010;56:449–53.
41. Michaelis M, Langhammer M, Höflich A, Reinsch N, Schön J, Weitzel JM, et al. Initial characterization of an outbreed mouse model for male factor (in)fertility. *Andrology*. 2013;1:772–8.
42. Hu J, Ng PC. Predicting the effects of frameshifting indels. *Genome Biol*. 2012;13:R9.
43. Bartonicek N, Clark MB, Quek XC, Torpy JR, Pritchard AL, Maag JLV, et al. Intergenic disease-associated regions are abundant in novel transcripts. *Genome Biol*. 2017;18:241.
44. Scotti MM, Swanson MS. RNA mis-splicing in disease. *Nat Rev Genet*. 2016;17:19–32.
45. Kim E-S, Cole JB, Huson H, Wiggans GR, Van Tassel CP, Crooker BA, et al. Effect of artificial selection on runs of homozygosity in US Holstein cattle. *PLoS One*. 2013;8:e80813.
46. Xue Y, Prado-Martinez J, Sudmant PH, Narasimhan V, Ayub Q, Szpak M, et al. Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science* (80-). 2015;348:242–5.
47. Laurie CC, Nickerson DA, Anderson AD, Weir BS, Livingston RJ, Dean MD, et al. Linkage disequilibrium in wild mice. *PLoS Genet*. 2007;3:e144.
48. Davis BW, Williams FJ, Ostrander EA, Parker HG, Plassais J, Kim J, et al. Genetic selection of athletic success in sport-hunting dogs. *Proc Natl Acad Sci*. 2018;115:E7212–21.
49. Kim H, Lee T, Park W, Lee JW, Kim J, Lee B-Y, et al. Peeling back the evolutionary layers of molecular mechanisms responsive to exercise-stress in the skeletal muscle of the racing horse. *DNA Res*. 2013;20:287–98.
50. Foote AD, Vijay N, Ávila-Arcos MC, Baird RW, Durban JW, Fumagalli M, et al. Genome-culture coevolution promotes rapid divergence of killer whale ecotypes. *Nat Commun*. 2016;7:11693.
51. Gage JL, Jarquin D, Romay C, Lorenz A, Buckler ES, Kaeppeler S, et al. The effect of artificial selection on phenotypic plasticity in maize. *Nat Commun*. 2017;8:1348.
52. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. *Annu Rev Med*. 2010;61:437–55.
53. Weischenfeldt J, Symmons O, Spitz F, Korbel JO. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet*. 2013;14:125–38.
54. Johnson ME, Viggiano L, Bailey JA, Abdul-Rauf M, Goodwin G, Rocchi M, et al. Positive selection of a gene family during the emergence of humans and African apes. *Nature*. 2001;413:514–9.
55. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. *Nature*. 2006;444:444–54.
56. Paudel Y, Madsen O, Megens HJ, Frantz LAF, Bosse M, Bastiaansen JWM, et al. Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication. *BMC Genomics*. 2013;14:449.
57. Gao Y, Jiang J, Yang S, Hou Y, Liu GE, Zhang S, et al. CNV discovery for milk composition traits in dairy cattle using whole genome resequencing. *BMC Genomics*. 2017;18:265.
58. Zhang RQ, Wang JJ, Zhang T, Zhai HL, Shen W. Copy-number variation in goat genome sequence: a comparative analysis of the different litter size trait groups. *Gene*. 2019;696:40–6.
59. Chen C, Qiao R, Wei R, Guo Y, Ai H, Ma J, et al. A comprehensive survey of copy number variation in 18 diverse pig populations and identification of candidate copy number variable genes associated with complex traits. *BMC Genomics*. 2012;13:733.
60. Amsterdam A, Dantes A, Liscovitch M. Role of phospholipase-D and phosphatidic acid in mediating gonadotropin-releasing hormone-induced inhibition of preantral granulosa cell differentiation. *Endocrinology*. 1994;135:1205–11.

61. Adhikari D, Zheng W, Shen Y, Gorre N, Hämäläinen T, Cooney AJ, et al. Tsc/mTORC1 signaling in oocytes governs the quiescence and activation of primordial follicles. *Hum Mol Genet.* 2009;19:397–410.
62. Tuppi M, Kehrloesser S, Coutandin DW, Rossi V, Luh LM, Strubel A, et al. Oocyte DNA damage quality control requires consecutive interplay of CHK2 and CK1 to activate p63. *Nat Struct Mol Biol.* 2018;25:261–9.
63. Böing M, Brand-Saberi B, Napirei M. Murine transcription factor Math6 is a regulator of placenta development. *Sci Rep.* 2018;8:14997.
64. Qiu Y, Sun S, Yu X, Zhou J, Cai W, Qian L. Carboxyl ester lipase is highly conserved in utilizing maternal supplied lipids during early development of zebrafish and human. *Biochim Biophys Acta - Mol Cell Biol Lipids.* 1865;2020:158663.
65. Miller R, Lowe ME. Carboxyl ester lipase from either mother's milk or the pancreas is required for efficient dietary triglyceride digestion in suckling mice. *J Nutr.* 2008;138:927–30.
66. Kosova G, Scott NM, Niederberger C, Prins GS, Ober C. Genome-wide association study identifies candidate genes for male fertility traits in humans. *Am J Hum Genet.* 2012;90:950–61.
67. Coster A, Madsen O, Heuven HCM, Dibbits B, Groenen MAM, van Arendonk JAM, et al. The imprinted gene DIO3 is a candidate gene for litter size in pigs. *PLoS One.* 2012;7:e31825.
68. Magee DA, Berry DP, Berkowicz EW, Sikora KM, Howard DJ, Mullen MP, et al. Single nucleotide polymorphisms within the bovine DLK1-DIO3 imprinted domain are associated with economically important production traits in cattle. *J Hered.* 2011;102:94–101.
69. Tao L, He XY, Jiang YT, Lan R, Li M, Li ZM, et al. Combined approaches to reveal genes associated with litter size in Yunshang black goats. *Anim Genet.* 2020;51:924–34.
70. Morgan K, Harr B, White MA, Payseur BA, Turner LM. Disrupted gene networks in subfertile hybrid house mice. *Mol Biol Evol.* 2020;37:1547–62.
71. Flegel C, Vogel F, Hofreuter A, Schreiner BSP, Osthold S, Veitinger S, et al. Characterization of the Olfactory receptors expressed in human spermatozoa. *Front Mol Biosci.* 2016;2:73.
72. Daei-Farshbaf N, Aflatoonian R, Amjadi FS, Taleahmad S, Ashrafi M, Bakhtiyari M. Expression pattern of olfactory receptor genes in human cumulus cells as an indicator for competent oocyte selection. *Turkish J Biol.* 2020;44:371–80.
73. Arck P, Hansen PJ, Jericevic BM, Piccinni MP, Szekeres-Bartho J. Progesterone during pregnancy: endocrine-immune cross talk in mammalian species and the role of stress. *Am J Reprod Immunol.* 2007;58:268–79.
74. Taraborrelli S. Physiology, production and action of progesterone. *Acta Obstet Gynecol Scand.* 2015;94:8–16.
75. Zeberg H, Kelso J, Pääbo S. The Neandertal Progesterone Receptor. *Mol Biol Evol.* 2020;37:2655–60.
76. Lv X, He C, Huang C, Wang H, Hua G, Wang Z, et al. Timely expression and activation of YAP1 in granulosa cells is essential for ovarian follicle development. *FASEB J.* 2019;33:10049–64.
77. Anand-Ivell R, Ivell R. Regulation of the reproductive cycle and early pregnancy by relaxin family peptides. *Mol Cell Endocrinol.* 2014;382:472–9.
78. Lei W, Feng XH, Deng WB, Ni H, Zhang ZR, Jia B, et al. Progesterone and DNA damage encourage uterine cell proliferation and decidualization through up-regulating ribonucleotide reductase 2 expression during early pregnancy in mice. *J Biol Chem.* 2012;287:15174–92.
79. Tsuneki H, Wada T, Sasaoka T. Role of orexin in the regulation of glucose homeostasis. *Acta Physiol.* 2010;198:335–48.
80. Taussat S, Boussaha M, Ramayo-Caldas Y, Martin P, Venot E, Cantalapie-dra-Hijar G, et al. Gene networks for three feed efficiency criteria reveal shared and specific biological processes. *Genet Sel Evol.* 2020;52:1–14.
81. Zhang Y, Kent JW, Olivier M, Ali O, Broeckel U, Abdou RM, et al. QTL-based association analyses reveal novel genes influencing pleiotropy of metabolic syndrome (MetS). *Obesity.* 2013;21:2099–111.
82. Liu B, Mao N. Smad5: Signaling roles in hematopoiesis and osteogenesis. *Int J Biochem Cell Biol.* 2004;36:766–70.
83. Taye M, Yoon J, Dessie T, Cho S, Oh SJ, Lee HK, et al. Deciphering signature of selection affecting beef quality traits in Angus cattle. *Genes and Genomics.* 2018;40:63–75.
84. Jiao S, Maltecca C, Gray KA, Cassady JP. Feed intake, average daily gain, feed efficiency, and real-time ultrasound traits in Duroc pigs: II. Genome-wide association. *J Anim Sci.* 2014;92:2846–60.
85. Xu H, Li H, Wang Z, Abudureyimu A, Yang J, Cao X, et al. A deletion downstream of the CHCHD7 gene is associated with growth traits in sheep. *Animals.* 2020;10:1–10.
86. An B, Xia J, Chang T, Wang X, Xu L, Zhang L, et al. Genome-wide association study reveals candidate genes associated with body measurement traits in Chinese Wagyu beef cattle. *Anim Genet.* 2019;50:386–90.
87. Schrauwen I, Giese APJ, Aziz A, Lafont DT, Chakchouk I, Santos-Cortez RLP, et al. FAM92A underlies nonsyndromic postaxial polydactyly in humans and an abnormal limb and digit skeletal phenotype in mice. *J Bone Miner Res.* 2019;34:375–86.
88. Tsuchiya M, Hara Y, Okuda M, Itoh K, Nishioka R, Shiomi A, et al. Cell surface flip-flop of phosphatidylserine is critical for PIEZO1-mediated myotube formation. *Nat Commun.* 2018;9:1–15.
89. Rode B, Shi J, Endesh N, Drinkhill MJ, Webster PJ, Lotteau SJ, et al. Piezo1 channels sense whole body physical activity to reset cardiovascular homeostasis and enhance performance. *Nat Commun.* 2017;8:1–11.
90. Göddeke S, Knebel B, Fahlbusch P, Hörbelt T, Poschmann G, Van De Velde F, et al. CDH13 abundance interferes with adipocyte differentiation and is a novel biomarker for adipose tissue health. *Int J Obes.* 2018;42:1039–50.
91. Teng MS, Wu S, Hsu LA, Chou HH, Ko YL. Differential associations between CDH13 genotypes, adiponectin levels, and circulating levels of cellular adhesive molecules. *Mediators Inflamm.* 2015;2015:635751.
92. Philippova M, Joshi MB, Kyriakakis E, Pfaff D, Erne P, Resink TJ. A guide and guard: the many faces of T-cadherin. *Cell Signal.* 2009;21:1035–44.
93. Lin JC, Chi YL, Peng HY, Lu YH. RBM4–Nova1–SRSF6 splicing cascade modulates the development of brown adipocytes. *Biochim Biophys Acta.* 1859;2016:1368–79.
94. Keller MA, Zander U, Fuchs JE, Kreutz C, Watschinger K, Mueller T, et al. A gatekeeper helix determines the substrate specificity of Sjögren-Larsson Syndrome enzyme fatty aldehyde dehydrogenase. *Nat Commun.* 2014;5:1–12.
95. Loro E, Jang C, Quinn WJ, Baur JA, Arany ZP, Khurana TS. Effect of interleukin-15 receptor alpha ablation on the metabolic responses to moderate exercise simulated by in vivo isometric muscle contractions. *Front Physiol.* 2019;10:1439.
96. Jiao H, Kaaman M, Dungner E, Kere J, Arner P, Dahlman I. Association analysis of positional obesity candidate genes based on integrated data from transcriptomics and linkage analysis. *Int J Obes.* 2008;32:816–25.
97. Duran J, Navarro-Sabate A, Pujol A, Perales JC, Manzano A, Obach M, et al. Overexpression of ubiquitous 6-phosphofructo-2-kinase in the liver of transgenic mice results in weight gain. *Biochem Biophys Res Commun.* 2008;365:291–7.
98. Sagara S, Osanai T, Itoh T, Izumiyama K, Shibutani S, Hanada K, et al. Overexpression of coupling factor 6 attenuates exercise-induced physiological cardiac hypertrophy by inhibiting PI3K/Akt signaling in mice. *J Hypertens.* 2012;30:778–86.
99. Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science (80-).* 2021;372:eabf7117.
100. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. *Nature.* 2015;526:75–81.
101. Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, et al. SpeedSeq: Ultra-fast personal genome analysis and interpretation. *Nat Methods.* 2015;12:966–8.
102. Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat Commun.* 2017;8:1–11.
103. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. *Genome Biol.* 2016;17:122.
104. Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, et al. OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* 2019;47:D807–11.
105. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez gene: gene-centered information at NCBI. *Nucleic Acids Res.* 2011;39:D52–7.
106. Ge SX, Jung D, Jung D, Yao R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics.* 2020;36:2628–9.
107. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 2021;49:D480–9.

108. Walsh B, Lynch M. Evolution and selection of quantitative traits. 1st ed. Oxford: Oxford University Press; 2018.
109. Henderson CR. Best linear unbiased estimation and prediction under a selection model. *Biometrics*. 1975;31:423–47.
110. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. 2016;32:1220–2.
111. Kronenberg ZN, Osborne EJ, Cone KR, Kennedy BJ, Domyan ET, Shapiro MD, et al. Wham: identifying structural variants of biological consequence. *PLoS Comput Biol*. 2015;11:e1004572.
112. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol*. 2014;15:R84.
113. Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol*. 2019;20:1–18.
114. Pirooznia M, Goes F, Zandi PP. Whole-genome CNV analysis: advances in computational approaches. *Front Genet*. 2015;6:138.
115. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30:2114–20.
116. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 2018;34:i884–90.
117. Andrews S. FastQC: a quality control tool for high throughput sequence data. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
118. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*. 2002;420:520–62.
119. Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, et al. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol*. 2009;7:1000112.
120. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Ridwan Amode M, et al. Ensembl 2021. *Nucleic Acids Res*. 2021;49:D884–91.
121. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25:1754–60.
122. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25:2078–9.
123. Broad Institute. Picard Toolkit. <http://broadinstitute.github.io/picard/>.
124. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20:1297–303.
125. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43:491–501.
126. Auwera GA, Carneiro MO, Hartl C, Poplin R, Angel G, Levy-Moonshine A. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinforma*. 2013;43:11.10.1–11.10.33.
127. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*. 2017. <https://doi.org/10.1101/201178>.
128. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. 2001;29:308–11.
129. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al. Ensembl 2020. *Nucleic Acids Res*. 2020;48:D682–8.
130. Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, et al. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*. 2011;477:289–94.
131. Avvaru AK, Sharma D, Verma A, Mishra RK, Sowpati DT. MSDB: a comprehensive, annotated database of microsatellites. *Nucleic Acids Res*. 2020;48:D155–9.
132. Cingolani P, Platts A, Le Wang L, Coon M, Nguyen T, Wang LLL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*. 2012;6:80–92.
133. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*. 2003;31:3812–4.
134. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nat Rev Genet*. 2011;12:363–76.
135. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*. 2012;28:3326–8.
136. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*. 2018;35:526–8.
137. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 2009;19:1655–64.
138. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015;4:7.
139. Purcell S, Chang C. PLINK 2. <http://www.cog-genomics.org/plink/2.0/>. Accessed 2 Mai 2019.
140. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27:2156–8.
141. Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-Smith C, Durbin R. BCFTools/RoH: A hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics*. 2016;32:1749–51.
142. Frayer ME, Payseur BA. Demographic history shapes genomic ancestry in hybrid zones. *Ecol Evol*. 2021;11:10290–302.
143. Cox A, Ackert-Bicknell CL, Dumont BL, Yueming D, Bell JT, Brockmann GA, et al. A new standard genetic map for the laboratory mouse. *Genetics*. 2009;182:1335–44.
144. Weir BS, Cockerham CC. Estimating F-Statistics for the analysis of population structure. *Evolution* (NY). 1984;38:1358–70.
145. Lai FN, Zhai HL, Cheng M, Ma JY, Cheng SF, Ge W, et al. Whole-genome scanning for the litter size trait associated genes and SNPs under selection in dairy goat (*Capra hircus*). *Sci Rep*. 2016;6:1–12.
146. Wang GD, Zhai W, Yang HC, Fan RX, Cao X, Zhong L, et al. The genomics of selection in dogs and the parallel evolution between dogs and humans. *Nat Commun*. 2013;4:1860.
147. Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A*. 1979;76:5269–73.
148. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. *PLoS Comput Biol*. 2013;9:e1003118.
149. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000;25:25–9.
150. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res*. 2019;47:D330.
151. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 2000;28:27–30.
152. Kanehisa M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci*. 2019;28:1947–51.
153. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res*. 2021;49:D545.
154. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res*. 2019;47:W199–205.
155. Wang J, Vasaikar S, Shi Z, Greer M, Zhang B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Res*. 2017;45:W130–7.
156. Wang J, Duncan D, Shi Z, Zhang B. WEB-based GENE SeT Analysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res*. 2013;41:W77–83.
157. Zhang B, Kirov S, Snoddy J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res*. 2005;33:W741–8.
158. Mouse Genome Informatics. http://www.informatics.jax.org/downloads/reports/mgi_mrk_coord.rpt. Accessed 22 Feb 2021.
159. Bult CJ, Blake JA, Smith CL, Kadin JA, Richardson JE, Anagnostopoulos A, et al. Mouse Genome Database (MGD) 2019. *Nucleic Acids Res*. 2019;47:D801–6.
160. Core R. Team. R: a language and environment for statistical computing <http://www.r-project.org/>. Vienna, Austria: R Foundation for Statistical Computing. 2020.

161. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the tidyverse. *J Open Source Softw*. 2019;4:1686.
162. Whole Genome Sequencing outbred mouse lines selected for high fertility, body size and endurance. The European Nucleotide Archive. 2021. <http://www.ebi.ac.uk/ena/browser/view/prjeb44248>.
163. Genomic characterization of world's longest selection experiment in mouse reveals the complexity of polygenic traits. The European Variation Archive. 2021. <http://www.ebi.ac.uk/eva/?eva-study=prjeb45961>.
164. WGS analysis of the Dummerstorf mouse lines. GitHub. 2021. http://www.github.com/sosfert/mmu_dummerstorf_wgs.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



Chapter IV

Chromosome-Level Genome Assemblies Expand Capabilities of Genomics for Conservation Biology

Azamat Totikov^{1,†}, Andrey Tomarovsky^{1,†}, Dmitry Prokopov², Aliya Yakupova¹, Tatiana Bulyonkova³, **Lorena Derezanin**⁴, Dmitry Rasskazov⁵, Walter W. Wolfsberger^{6,7}, Klaus-Peter Koepfli^{1,8,9}, Taras K. Oleksyk^{6,7,10,*}, Sergei Kliver^{2,*}

1 Computer Technologies Laboratory, ITMO University, 197101 Saint Petersburg, Russia;

2 Department of the Diversity and Evolution of Genomes, Institute of Molecular and Cellular Biology SB RAS,

630090 Novosibirsk, Russia

3 Laboratory of Mixed Computations, A.P. Ershov Institute of Informatics Systems SB RAS,

630090 Novosibirsk, Russia

4 Department of Evolutionary Genetics, Leibniz Institute for Zoo and Wildlife Research (IZW),

10315 Berlin, Germany

5 Institute of Cytology and Genetics, Siberian Branch of Russian Academy of Sciences,

630090 Novosibirsk, Russia

6 Department of Biological Sciences, Oakland University, Rochester, MI 48307, USA

7 Department of Biology, Uzhhorod National University, 88000 Uzhhorod, Ukraine

8 Smithsonian-Mason School of Conservation, Front Royal, VA 22630, USA

9 Center for Species Survival, Smithsonian Conservation Biology Institute, National Zoological Park, Washington, DC 20008, USA

10 Biology Department, University of Puerto Rico at Mayagüez, Mayagüez, PR 00682, USA



** Corresponding authors*

† These authors contributed equally to this work.

Published in *Genes*, 2021, DOI: 10.3390/genes12091336

Article

Chromosome-Level Genome Assemblies Expand Capabilities of Genomics for Conservation Biology

Azamat Totikov ^{1,†} , Andrey Tomarovsky ^{1,†} , Dmitry Prokopov ² , Aliya Yakupova ¹, Tatiana Bulyonkova ³, Lorena Derezanin ⁴ , Dmitry Rasskazov ⁵, Walter W. Wolfsberger ^{6,7}, Klaus-Peter Koepfli ^{1,8,9} , Taras K. Oleksyk ^{6,7,10,*}  and Sergei Kliver ^{2,*}

- ¹ Computer Technologies Laboratory, ITMO University, 197101 Saint Petersburg, Russia; a.totikov1@gmail.com (A.T.); andrey.tomarovsky@gmail.com (A.T.); yakupovaar@yandex.ru (A.Y.); klauspeter.koepfli527@gmail.com (K.-P.K.)
- ² Department of the Diversity and Evolution of Genomes, Institute of Molecular and Cellular Biology SB RAS, 630090 Novosibirsk, Russia; dprokopov@mcb.nsc.ru
- ³ Laboratory of Mixed Computations, A.P. Ershov Institute of Informatics Systems SB RAS, 630090 Novosibirsk, Russia; ressaure@gmail.com
- ⁴ Department of Evolutionary Genetics, Leibniz Institute for Zoo and Wildlife Research (IZW), 10315 Berlin, Germany; derezanin@izw-berlin.de
- ⁵ Institute of Cytology and Genetics, Siberian Branch of Russian Academy of Sciences, 630090 Novosibirsk, Russia; rassk@bionet.nsc.ru
- ⁶ Department of Biological Sciences, Oakland University, Rochester, MI 48307, USA; wwolfsberger@oakland.edu
- ⁷ Department of Biology, Uzhhorod National University, 88000 Uzhhorod, Ukraine
- ⁸ Smithsonian-Mason School of Conservation, Front Royal, VA 22630, USA
- ⁹ Center for Species Survival, Smithsonian Conservation Biology Institute, National Zoological Park, Washington, DC 20008, USA
- ¹⁰ Biology Department, University of Puerto Rico at Mayagüez, Mayagüez, PR 00682, USA
- * Correspondence: oleksyk@oakland.edu (T.K.O.); mahajrod@gmail.com (S.K.)
- † These authors contributed equally to this work.



Citation: Totikov, A.; Tomarovsky, A.; Prokopov, D.; Yakupova, A.; Bulyonkova, T.; Derezanin, L.; Rasskazov, D.; Wolfsberger, W.W.; Koepfli, K.-P.; Oleksyk, T.K.; et al. Chromosome-Level Genome Assemblies Expand Capabilities of Genomics for Conservation Biology. *Genes* **2021**, *12*, 1336. <https://doi.org/10.3390/genes12091336>

Academic Editor: Jennifer A. Leonard

Received: 29 July 2021

Accepted: 25 August 2021

Published: 28 August 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: Genome assemblies are in the process of becoming an increasingly important tool for understanding genetic diversity in threatened species. Unfortunately, due to limited budgets typical for the area of conservation biology, genome assemblies of threatened species, when available, tend to be highly fragmented, represented by tens of thousands of scaffolds not assigned to chromosomal locations. The recent advent of high-throughput chromosome conformation capture (Hi-C) enables more contiguous assemblies containing scaffolds spanning the length of entire chromosomes for little additional cost. These inexpensive contiguous assemblies can be generated using Hi-C scaffolding of existing short-read draft assemblies, where N50 of the draft contigs is larger than 0.1% of the estimated genome size and can greatly improve analyses and facilitate visualization of genome-wide features including distribution of genetic diversity in markers along chromosomes or chromosome-length scaffolds. We compared distribution of genetic diversity along chromosomes of eight mammalian species, including six listed as threatened by IUCN, where both draft genome assemblies and newer chromosome-level assemblies were available. The chromosome-level assemblies showed marked improvement in localization and visualization of genetic diversity, especially where the distribution of low heterozygosity across the genomes of threatened species was not uniform.

Keywords: genome assemblies; scaffolds; genomes; Hi-C; heterozygosity; mammals; conservation genetics; STR markers

1. Introduction

Each species inhabits a specific environment, a niche, that shapes its unique genome sequence and its expression. Genetic diversity within species is valuable through the existence of the unique combinations of genes and alleles present at a given time in a

population, but it is also valuable as it contributes to ongoing evolutionary processes [1]. As environments continuously change, some species can adapt to this change, while others cannot. Understanding factors that determine survival and adaptive potential in response to the environmental change enables a better and effective design of conservation strategies for each species [2,3].

Genome diversity is formed by a balance of mutation, drift, and gene flow contributing to the distribution of diversity across the loci along the chromosomes of individuals in a population. The resulting patterns of variation provide a backdrop for natural selection, enabling adaptation [4,5], and it is generally thought that preserving genetic diversity is required for adaptability: a species that has lost all its reserves of genetic diversity is doomed to extinction [6–12]. In this context, adaptability is generally understood to depend on the existing genetic variation within each species. Indeed, among many endangered and threatened species, genome-wide genetic diversity has been severely reduced, which is usually seen as a critical sign of vulnerability, as genetically diverse populations should be more resilient to environmental change due to a higher adaptive potential [6,13–15].

Heterozygosity has been routinely used to evaluate the genetic potential of a population faced with extinction, and a significant majority of threatened taxa show lower genetic diversity than taxonomically related but not threatened taxa [16]. Historically, genetic diversity has been estimated as heterozygosity across neutral markers without regard to their chromosomal locations [17,18]. Low heterozygosity points to high levels of inbreeding, non-random mating, population fragmentation, and potential recent bottlenecks. This measure is simple and easy to estimate, even from a relatively small number of individuals, if enough loci are examined [18], which is perhaps the main reason why heterozygosity is still widely used in conservation genetics to make estimates of genetic structure, migration rates, and effective population sizes of endangered species [19–21].

Conservation biology deals with a huge number of species, but genomic studies of non-model organisms usually have significantly smaller budgets than model organisms used in the biomedical or agricultural sciences, forcing conservation scientists into a continuous trade off between quality and quantity of generated data and its cost. Fortunately, the ongoing reduction in genome sequencing costs gradually allows for the increasingly wider adoption of the whole genome resequencing approaches to estimate genetic diversity among as well as within species. However, this usually requires an existing reference genome assembly of sufficient contiguity and quality to be either available or generated for use in aligning reads and calling variants from the resequenced genomes. The price for generating a quality de novo assembly is still a challenge for most conservation genomics teams, depending on the technology used for genome sequencing (e.g., Armstrong et al., [22]).

A temporary solution to this problem (at least in the short-term perspective) is enabled by the recently developed USD 1000 approach for generation of chromosome-level assemblies from one short-insert Illumina paired-end library and an in situ high-throughput chromosome conformation capture (Hi-C) library [23]. An illustration of this approach can help justify a new path towards future studies of genome-wide patterns of diversity across loci in endangered species.

We collected genetic and genomic data from seven threatened mammalian species for which previous highly fragmented scaffold assemblies and recently generated chromosome-level assemblies (including those generated by the USD \$1000 approach) were available. Using these assemblies, we performed a comparison between the analyses based on the traditional genetic data versus the new genomic approach to estimate genetic diversity genome-wide. Our primary objective was to evaluate if the newer, more contiguous assemblies allowed for a better estimation of genetic diversity, localization, and visualization of low heterozygosity regions within genomes.

2. Materials and Methods

2.1. Genomic Data

Draft and chromosome-level assemblies of eight mammal species were downloaded from the NCBI Genome and DNA Zoo databases (Table 1, Table S1). Six of the species examined had a total of 19 pairs of chromosomes ($2n = 38$), one (*Ailurus fulgens*) had 18 pairs ($2n = 36$), and one (*Bison bison*) had 30 pairs ($2n = 60$) (Table 1). Short-read libraries were obtained from NCBI SRA [23–28]; the corresponding SRA accession IDs are listed per species in Supplementary Table S2.

Table 1. Mammalian species and corresponding genome assemblies used in this study. The measures show the length of the genome size (length), the size of gaps in the assembly (Ns), the N50, and the change in the gap size from draft to chromosome-level assembly (dN).

Species	IUCN Red List Category ¹	Common Name	2n	Assembly Source or ID	Assembly Level ²	Length, Gbp	Ns, Mbp	N50, Mbp	dN, %
<i>Aonyx cinereus</i>	VU	Asian small-clawed otter	38	DNA Zoo	Chr	2.44	15.5	130.94	+1048%
				DNA Zoo draft	Draft	2.42	1.35	0.1	
<i>Enhydra lutris</i>	EN	Sea otter	38	DNA Zoo	Chr	2.45	28.94	145.94	−2%
				GCA_002288905.2	Draft	2.46	29.68	38.75	
<i>Lutra lutra</i>	LC	Eurasian river otter	38	DNA Zoo	Chr	2.44	0.1	148.99	n/a
<i>Pteronura brasiliensis</i>	EN	Giant otter	38	DNA Zoo	Chr	2.46	11.89	133.38	+749%
				DNA Zoo draft	Draft	2.45	1.4	0.17	
<i>Ailurus fulgens</i>	EN	Red panda	36	DNA Zoo	Chr	2.34	34.41	143.8	+1%
				GCA_002007465.1	Draft	2.34	34.04	2.98	
<i>Acinonyx jubatus</i>	VU	Cheetah	38	DNA Zoo	Chr	2.37	42.86	144.64	+2%
				GCA_001443585.1	Draft	2.37	42.06	3.12	
<i>Neofelis nebulosa</i>	VU	Clouded leopard	38	DNA Zoo	Chr	2.42	7.94	147.11	+35%
				DNA Zoo draft	Draft	2.41	5.89	1.38	
<i>Bison bison</i>	NT	American bison	60	DNA Zoo	Chr	2.83	199.31	101.69	+2%
				GCF_000754665.1	Draft	2.83	195.77	7.19	

¹ IUCN Red List categories: EN—endangered, VU—vulnerable, NT—near threatened, LC—least concern. ² The assembly levels: draft—initial fragmented assembly, Chr—chromosome-level assembly based on draft and in situ high-throughput chromosome conformation capture (Hi-C).

2.2. Quality Control and Filtration of Data

Raw data quality control was performed using the *FastQC* [29] and *KrATER v1.1* [30] software. Adapter trimming and filtration by quality was performed in two stages with initial kmer-based trimming of large adapter fragments using the *Cookiestrimmer* software [31], followed by additional trimming of small fragments and quality filtering using the *Trimomatic* software, v0.36 [32].

2.3. Alignment and Variant Calling

Filtered reads were aligned to the corresponding reference genome assemblies using the *BWA* tool, v0.7.17 [33]. Read duplicates were marked with the *Samtools* package, v1.9 [34]. Variant calling was performed using *bcftools* v1.10 [35] with the following parameters: “-d 250 -q 30 -Q 30 -adjust-MQ 50 -a AD, INFO/AD, ADF, INFO/ADF, ADR, INFO/ADR, DP, SP, SCR, INFO/SCR” for *bcftools mpileup* and “-m -v -f GQ,GP” for *bcftools call*. Low quality variants (‘QUAL < 20.0 || FORMAT/SP > 60.0 || FORMAT/DP < 5.0 || FORMAT/GQ < 20.0’) were removed using *bcftools* filter. Finally, variants were filtered by coverage. Only variants in regions with 50–250% of whole genome median coverage were retained.

2.4. Identification of X Chromosome, Autosomes, and Pseudoautosomal Region

The position of the pseudoautosomal region (PAR) on the X chromosome was detected in several steps. First, the per-base coverage of the corresponding genome assembly was calculated for each genome library analyzed using the *Genomecov* tool from the *Bedtools* package [36]. Next, median coverage was calculated in stacking windows of 10 kbp. Adjacent windows were merged if their median coverage was at least 70% of the whole genome value, but among the merged windows, only those of 100 kbp or longer were retained. Finally, all the combined windows were merged into final regions if the median coverage of the windows in the gap between them was no lower than 70% of the whole-genome coverage.

Identification of the X chromosome (for all species) and autosomes (for cheetah and red panda) was performed using comparisons of the whole genome alignment (WGA) to the genome assembly (v 9.0) of the reference species (domestic cat, *Felis catus*) and published Zoo-FISH data [37]. The corresponding WGA was generated using *LAST aligner v961* [38]. Synonyms to C-scaffolds of all genome assemblies used in this study are listed in Supplementary Table S4.

2.5. Comparison of Heterozygosity in Autosomes, X Chromosome, and PAR

For male and female individuals of sea otter and American bison, we compared SNP heterozygosity in autosomes and the PAR in 100 kbp stacking windows using Mann–Whitney nonparametric test. To obtain lesser-greater priors and choose the type of test for comparisons between X chromosome and autosome heterozygosity, we selected five subsets of windows and generated boxplots for them using the *Matplotlib* library: windows sampled from the (1) whole genomes (all), (2) autosomes only (noX), (3) the X chromosome only (onlyX), (4) X chromosome without PAR (noPAR), and (5) pseudoautosomal region only (PAR). Based on the distribution plot, we chose one-sided tests for both PAR versus autosomes and autosomes versus noPAR with following alternative hypotheses: “PAR is more heterozygous than autosomes” for the first comparison, and “autosomes are more heterozygous than noPAR” for the second. The first comparison resulted in raw *p*-values of 4.9×10^{-10} for female (SRR8588177) and 1.1×10^{-9} for male (SRR8588180) American bison, and 1.5×10^{-8} , 3.4×10^{-14} , and 2.7×10^{-13} for female (SRR8597300), male 1 (SRR5768046) and male 2 (SRR5768052) sea otters, respectively. The second test was performed only for females and showed a raw *p*-value 2.4×10^{-81} for the female sea otter and a much lower value for female American bison. Even with the Bonferroni correction for multiple comparisons, *p*-values were below the significance level of 0.01, resulting in the acceptance of alternative hypotheses in all cases.

2.6. Heterozygosity Visualization

Filtered genetic variants were split into two categories: (1) single nucleotide polymorphisms (SNPs) and (2) insertion-deletions (indels). All subsequent analyses were based on SNPs only. Indels could not be used in this analysis due to the low-quality calls from short reads. Counts of heterozygous SNPs were calculated in non-overlapping windows of 100 kbp and 1 Mbp and scaled to SNPs per kbp. Heatmaps and boxplots were drawn using custom scripts based on the *Matplotlib 2* library [39].

2.7. Mapping of Known STR Loci on Chromosome-Level Assemblies of Mustelid Species

Primers of 66 previously published STR loci were extracted from seven different mustelid species—stone marten (*Martes foina*), (stone marten), American marten (*Martes americana*), wolverine (*Gulo gulo*), American badger (*Taxidea taxus*), European badger (*Meles meles*), American mink (*Neovison vison*), and ermine (*Mustela erminea*) [40–43] and used for in silico PCR using available chromosome-level genome assemblies of six mustelid species (Asian small-clawed otter, sea otter, North American river otter, Eurasian otter, domestic ferret, giant otter) and draft assembly of the American mink (Table 2). First, the in silico PCR was performed using *Simulate_PCR 1.2* [44] where, for the raw amplicons, no more

than four mismatches with the target sequence were allowed for each primer, and amplicon length was restricted to 50–1000 bp. Next, additional filtration was performed for each primer pair obtained, with raw amplicons ranked (from minimal to maximal values) by length of amplicon and maximum mismatches in pair—MM score = max (forward primer mismatches, reverse primer mismatches), total number of mismatches (TM score = forward primer mismatches + reverse primer mismatches). The top amplicons were then extracted.

Table 2. Mapping of known STR loci onto the six chromosome-level and one draft assemblies of seven mustelid species.

Species	STR Markers			#* Chr**	#* Chr** with Markers	#* Chr** w/o Markers
	Localized (L)	Not Amplified (NA)	Declined (D)			
<i>Aonyx cinereus</i> ¹	31	16	19	19	15	4
<i>Enhydra lutris</i> ²	26	22	18	19	14	5
<i>Lontra canadensis</i> ³	28	17	21	19	15	4
<i>Lutra lutra</i> ⁴	26	22	18	19	13	6
<i>Mustela putorius furo</i> ⁵	28	17	21	20	14	6
<i>Pteronura brasiliensis</i> ⁶	30	18	18	19	13	6
<i>Neovison vison</i> ^{7,***}	36	17	13	15	-	-

¹—Asian small-clawed otter, ²—sea otter, ³—North American river otter, ⁴—Eurasian otter, ⁵—domestic ferret, ⁶—giant otter, ⁷—American mink. *—Number of, **—Chromosomes, ***—there is no chromosome-level assembly for American mink available, but we included this species as control. See Section 3.4 for details.

All primer pairs were divided into three categories: NA—no amplification; D—amplified, but failed filtering criteria and declined; and L—localized (passed filtration) (Table 2). The L category contains primers used for further analysis and includes two groups that were selected. In both groups, the top one ranked amplicon was generated from both forward and reverse primers (RF or FR amplicons) and had an MM score of 3 or less. In addition, the requirements for the first group included existence of the only amplicon for primer pair. For the second group, multiple amplicons were allowed but with additional restrictions: either the difference in MM score between the top and adjacent amplicons had to be 2 or higher, or the difference in TM score had to be 3 or higher, respectively. Location of amplicons on C-scaffolds (sensu Lewin et al., [45]) was visualized using custom scripts based on the *Matplotlib* 2 library [39].

3. Results

3.1. Evaluation of the Genome Assemblies

We analyzed the genomes from eight mammal species representing different IUCN Red List categories (Table 1): sea otter (*Enhydra lutris*), cheetah (*Acinonyx jubatus*), clouded leopard (*Neofelis nebulosa*), giant otter (*Pteronura brasiliensis*), red panda (*Ailurus fulgens*), Asian small-clawed otter (*Aonyx cinereus*), American bison (*Bison bison*), and Eurasian river otter (*Lutra lutra*). Each of these species (except Eurasian river otter) was represented by two genome assemblies: the initial draft assembly and a chromosome-level assembly generated from the draft using Hi-C-scaffolding [23].

The draft assemblies were generated using different sequencing and assembly approaches, resulting in assemblies of different lengths and contiguities (Table 1; Table S1). The scaffold N50 of the draft assemblies ranged from 0.10 Mbp for Asian small-clawed otter to 38.75 Mbp for sea otter (Table 1). The total gap length (Ns, Table 1) also varied considerably among the assemblies, from 1.4 Mbp in giant otter to 195.77 in American bison.

The chromosome-level assemblies included several chromosome-length scaffolds or C-scaffolds [45] that corresponded with the haploid chromosome number (1n) of the species,

along with many smaller scaffolds. Between these categories, the lengths differed by orders of magnitude (from kbp to Mbp). The C-scaffolds were ordered according to length, from longest to shortest, without assignment to species-specific karyotype, except for cheetah and red panda for which such an assignment was performed using the Zoo-FISH data and whole genome alignments (see Section 2.3 for details).

Clearly, with the help Hi-C scaffolding, the N50 of the assemblies increased considerably (N50; Chr. vs. Draft, Table 1), as the fragments were aligned in their respective order along the C-scaffolds (N50, Table 1). The most dramatic improvement was observed for the Asian small-clawed otter ($\times 1309$) and giant otter ($\times 784$), while the smallest was observed for the sea otter ($\times 3.8$).

While the contiguity has been remarkably improved, the total gap size in most cases did not increase (Ns, Table 1) except in two of the eight species we considered: by 14.15 Mbp for Asian small-clawed otter, and by 10.49 Mbp in case of giant otter. Unfortunately, the Hi-C data cannot be used to estimate distances between the ordered scaffolds, because Hi-C scaffolding uses a fixed-length stretches of Ns to fill the gaps. For instance, a 500 bp insertion was used in the case of 3D-DNA pipeline for the analyzed chromosome-length assemblies. In the case of *sea otter*, the gap sizes slightly decreased (by 0.74 Mbp), probably due to an extensive correction of misassemblies or split on long gaps preceding the scaffolding stage.

3.2. Heterozygosity Estimations and Visualization

Heterozygosity is expected to be low in threatened and endangered species [16,19–21]. Heterozygosity is clearly diminished across the genomes of some the endangered and threatened species, but there is a difference in how this measure is distributed. The species we analyzed included those well known for low levels of heterozygosity, such as cheetah (Figure 1A) and sea otter (Figure 2A–C), which clearly showed extended regions of low heterozygosity/SNP density across their chromosomes (Figure 2A,B, dark blue). In other species that are also considered to be threatened, such as the Asian small-clawed otter (Figure 1F), red panda (Figure 1E), and American bison (Figure 2D,E), genetic diversity as reflected by higher SNP densities was still present in many chromosomal regions.

The Eurasian river otter is not considered threatened or endangered and has a global LC (least concern) status [46]. However, high heterozygosity in this species was observed only in a few chromosomal regions (Figure 1D), and 1130 Mbp of its genome was much less diverse. Similar levels of heterozygosity were observed in the clouded leopard ($0.1 < \text{hetSNPs per kbp } 0.75$), which is considered vulnerable (VU), and around 800 Mbp of its genome showed extremely low levels of heterozygosity ($0.1 \text{ hetSNP per kbp}$), similar to the endangered giant otter. The Eurasian river otter genome assembly was sequenced as a part of the 25 Genomes Project by the Wellcome Sanger Institute, but the origin of the individual sample used was not listed in the SRA database. This example emphasizes the necessity of sequencing several wild individuals of each species in a conservation study before making conclusions about genome-wide heterozygosity.

The distributions of the counts of heterozygous SNPs calculated in non-overlapping windows of 100 kbp and 1 Mbp and scaled to SNPs per kbp are presented in Figure 3. We noted that variant counts between the draft and the chromosome-level assemblies were similar for all species in our analysis (Table 3, number of SNPs). However, representing draft assemblies as density plots is challenging due to the high number of short scaffolds that are generally smaller than the window size of 1 Mb. In a typical contiguous 2.5–3.0 Gbp mammalian genome, the number of 100 kbp windows ranges between 25,000 and 30,000, whereas for a window size of 1 Mbp, the number of windows ranges between 2500 and 3000 (Table 3), which enables easier visualization of SNP density and heterozygosity along C-scaffolds (as in Figure 2A,B). Among the eight studied species, giant otter and Asian small-clawed otter had the smallest scaffold N50 values—0.17 and 0.1 Mbp, respectively (Table 1, Figure 3)—and were the most fragmented among the ones we considered. These two draft genome assemblies also had the smallest numbers of SNPs per 1 Mbp and even per 100 kbp windows (Table 3, Figure 3).

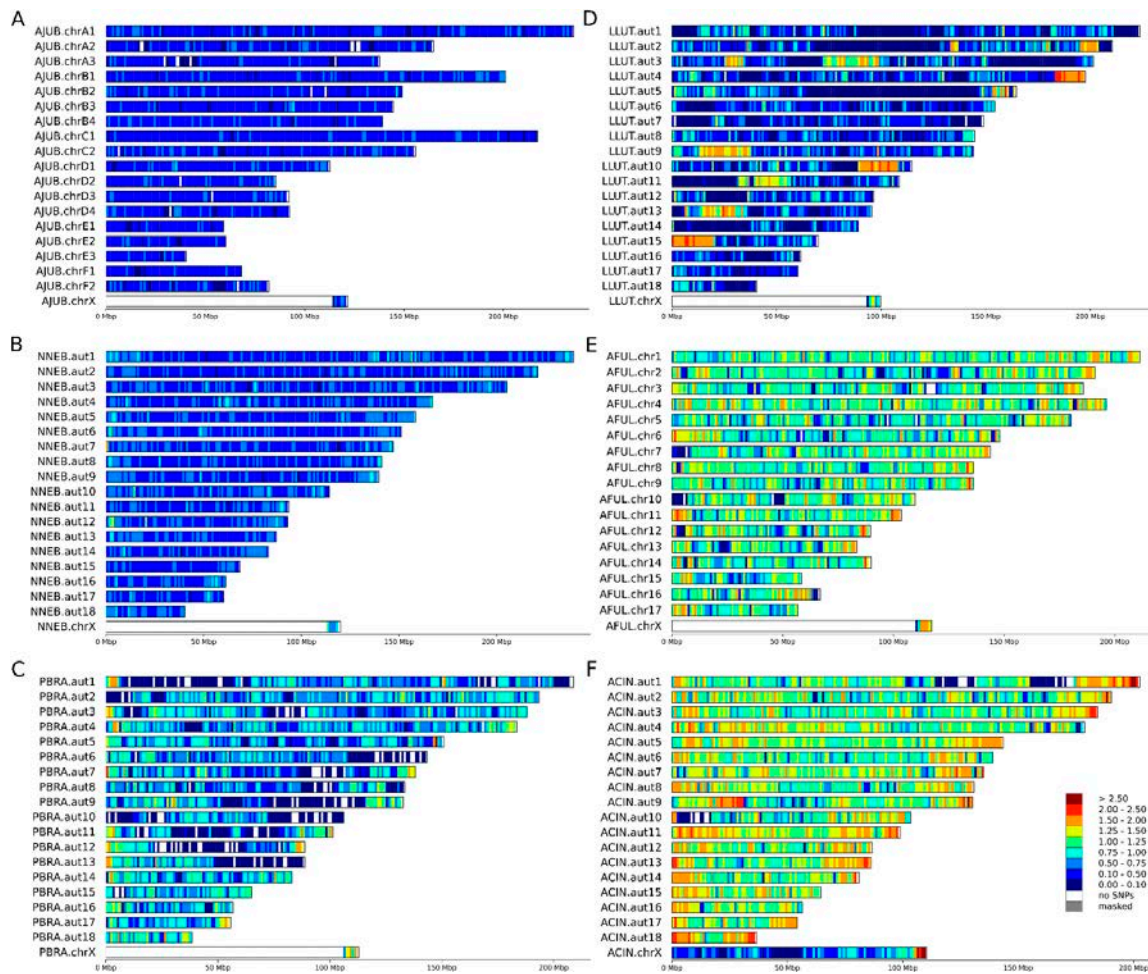


Figure 1. Heatmaps of heterozygous SNP densities for analyzed species based on chromosome-level assemblies for single individuals of six species. Heterozygous SNPs were counted in 1 Mbp windows and scaled to SNP/kbp. (A) male cheetah (*Acinonyx jubatus*), (B) male clouded leopard (*Neofelis nebulosa*), (C) male giant otter (*Pteronura brasiliensis*), (D) male Eurasian river otter (*Lutra lutra*), (E) male red panda (*Ailurus fulgens*), (F) female Asian small-clawed otter (*Aonyx cinereus*).

Table 3. Counts of heterozygous single nucleotide polymorphisms (SNPs), windows and counts, median, and mean heterozygosity in windows of 100 kbp and 1 Mbp for draft and chromosome-level assemblies (Chr.) assemblies of the analyzed genomes. Two species with the lowest window counts are in italic. Bold indicates cases where comparison of mean heterozygosity in windows of 100 kbp and 1 Mbp showed statistically significant difference for significance, level 0.01.

Species	#* Het. SNPs, Millions		Window Size	#*Windows		Median, Het SNPs/kbp		Mean, Het SNPs/kbp		p-Value (Draft vs. Chr.)	
	Draft	Chr.		Draft	Chr.	Draft	Chr.	Draft	Chr.	Raw	Adjusted
<i>Aonyx cinereus</i>	2.73	2.73	100 kbp	9777	22,183	1.100	1.190	1.052	1.144	3.37×10^{-34}	2.36×10^{-33}
			1 Mbp	3	2204	0.001	1.177	0.292	1.146	NA	NA
<i>Enhydra lutris</i>	0.47	0.46	100 kbp	24,146	24,165	0.140	0.140	0.178	0.182	0.98	1
			1 Mbp	2337	2396	0.174	0.176	0.175	0.180	0.79	1
<i>Pteronura brasiliensis</i>	1.25	1.24	100 kbp	13,589	22,819	0.410	0.410	0.488	0.497	0.59	1
			1 Mbp	32	2262	0.699	0.542	0.563	0.497	NA	NA
<i>Ailurus fulgens</i>	2.14	2.14	100 kbp	22,083	23,139	0.920	0.920	0.916	0.912	0.50	1
			1 Mbp	1573	2298	0.980	0.971	0.943	0.914	0.17	1

Table 3. Cont.

Species	#* Het. SNPs, Millions		Window Size	#*Windows		Median, Het SNPs/kbp		Mean, Het SNPs/kbp		<i>p</i> -Value (Draft vs. Chr.)	
	Draft	Chr.		Draft	Chr.	Draft	Chr.	Draft	Chr.	Raw	Adjusted
<i>Acinonyx jubatus</i>	0.75	0.75	100 kbp	22,861	23,609	0.280	0.280	0.314	0.314	0.42	1
			1 Mbp	1757	2350	0.332	0.330	0.322	0.314	0.23	1
<i>Neofelis nebulosa</i>	1.00	1.00	100 kbp	22,004	23,931	0.380	0.370	0.415	0.407	6.62×10^{-04}	0.0046
			1 Mbp	1194	2387	0.427	0.415	0.426	0.407	1.2×10^{-03}	0.0089
<i>Bison bison</i>	3.68	3.68	100 kbp	24,286	26,213	1.160	1.100	1.423	1.378	5.33×10^{-07}	3.73×10^{-06}
			1 Mbp	2181	2604	1.328	1.324	1.414	1.379	0.142	0.9943

*—Number of.

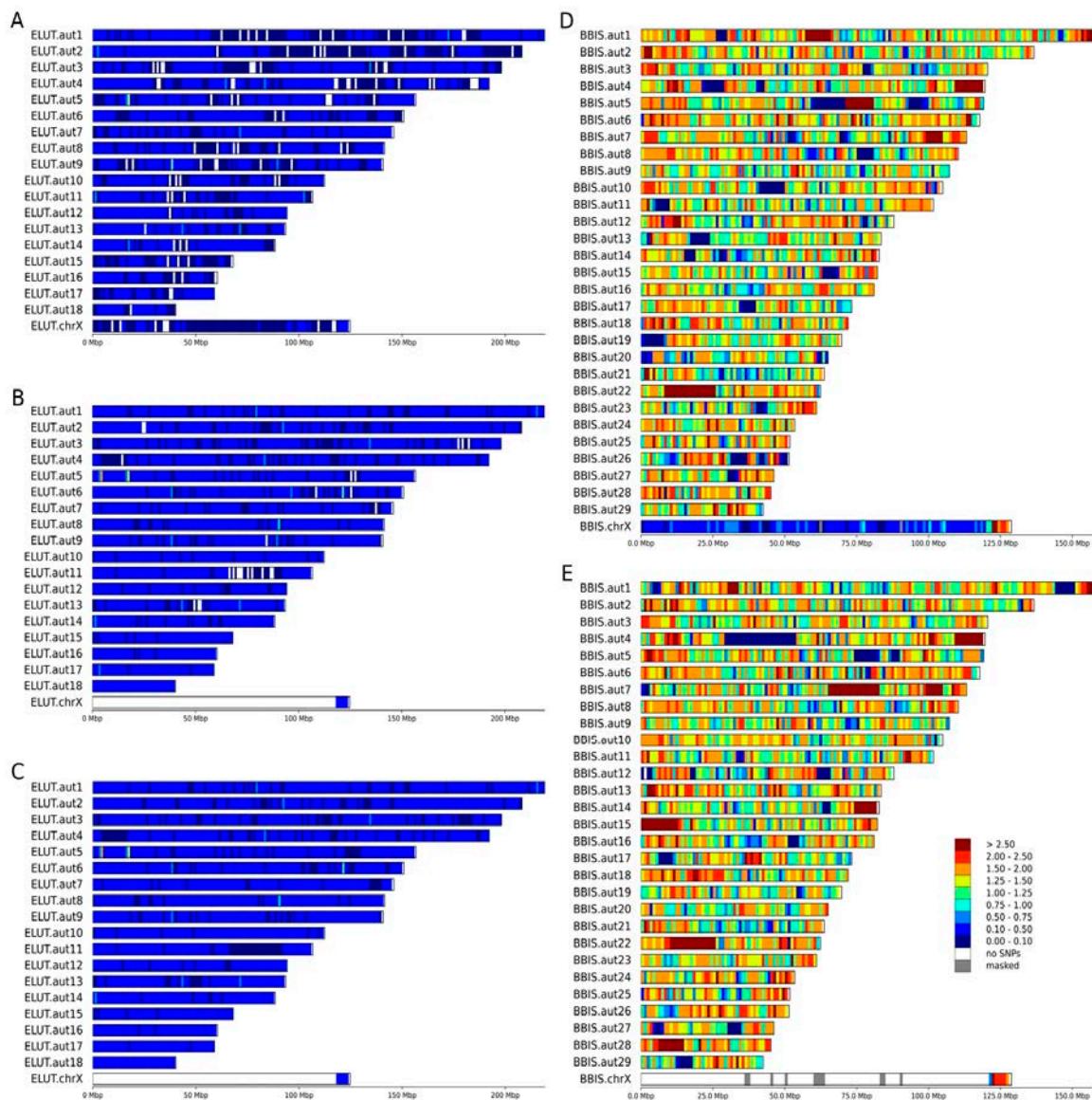


Figure 2. Chromosome-level heatmaps of heterozygosity (density of heterozygous SNP) for two additional species. Individuals of both sexes were available for two the species, sea otters (*Enhydra lutris*) and the American bison (*Bison bison*). Heterozygous SNPs were counted in 1 Mbp windows and scaled to SNP/kbp. (A) female sea otter (SRR8597300), (B,C) male sea otters (SRR5768046, SRR5768052), (D) female bison (SRR8588177), (E) male bison (SRR8588180).

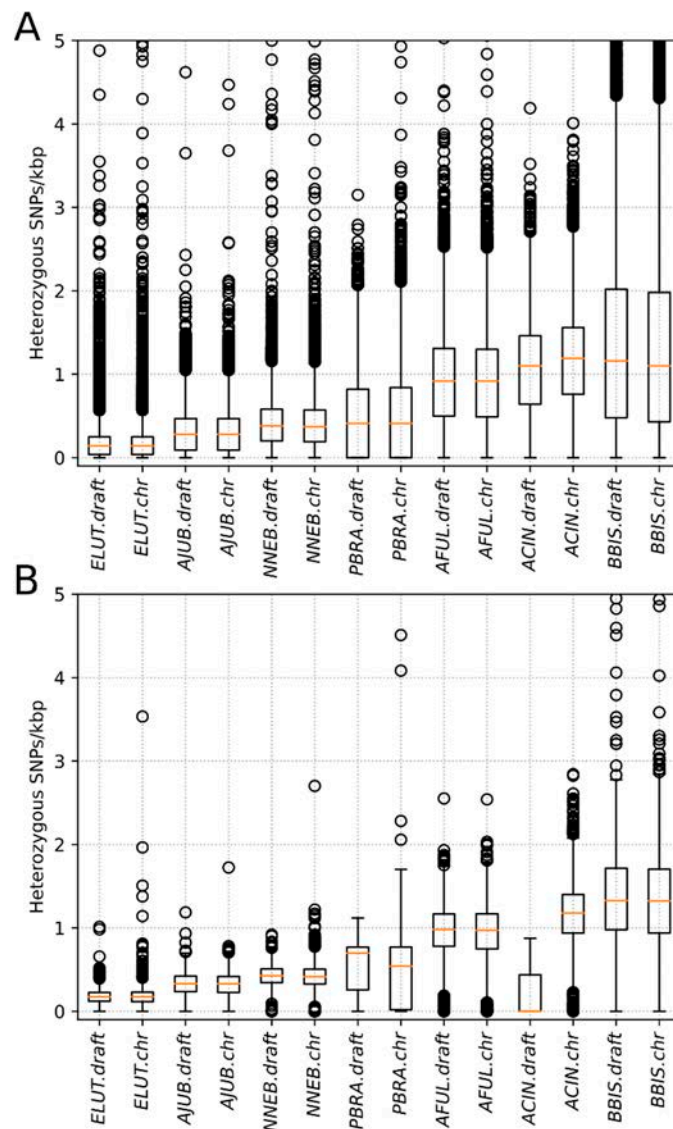


Figure 3. Comparison of the distributions of mean heterozygosity in (A) windows of 100 kbp and (B) windows of 1 Mbp for the short-read assembled draft genomes and the chromosome-level assemblies. The codes in the Figure are: ELUT—sea otter (*Enhydra lutris*), AJUB—cheetah (*Acinonyx jubatus*), NNEB—clouded leopard (*Neofelis nebulosa*), PBRA—giant otter (*Pteronura brasiliensis*), AFUL—red panda (*Ailurus fulgens*), ACIN—Asian small-clawed otter (*Aonyx cinereus*), and BBIS—American bison (*Bison bison*).

3.3. X Chromosome and the Pseudoautosomal Region

C-scaffolds corresponding to the X chromosome were identified in the chromosome-level assemblies using the coverage-based approach and libraries generated from male individuals available for seven of the eight analyzed species. In the case of the Asian small-clawed otter, of which only one female individual was sequenced, the X chromosome was identified from whole genome alignment to domestic cat X chromosome. The depth of coverage counted in 1 Mbp stacking windows for 11 males or females clearly revealed the location of the single pseudoautosomal region on the end of X chromosome as expected (Figure 4). Refinement of PAR borders using 10 kbp windows (see Section 2.3 for details) showed variation in its length among species. The shortest PAR (5.6 Mbp) was observed in cheetah and the longest (7.2 Mbp) in the American bison (Table S6).

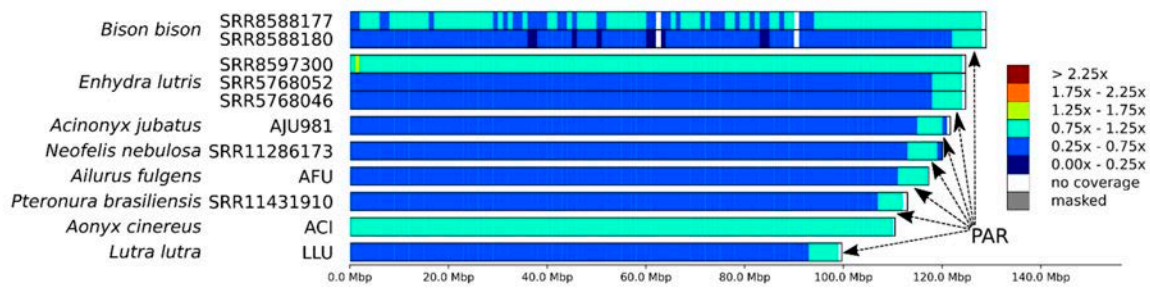


Figure 4. The depth of coverage along the identified X chromosome in eight mammal species (Table 1). The dark blue color corresponds to a half coverage ($0.25\times$ – $0.75\times$ genome coverage, and the teal-colored fragments are covered at $1\times$ ($0.75\times$ – $1.25\times$). Arrows show location of pseudoautosomal region (PAR) in male individuals.

Among the eight species analyzed, whole genome data from both male and female individuals were available only for the sea otter and American bison. For these species, we compared SNP heterozygosity in 100 kbp stacking windows between autosomes and the PAR using the Mann–Whitney nonparametric test. To obtain lesser-greater priors and select the type of test for comparison between X chromosome and autosome heterozygosity, we selected five subsets of windows and generated boxplots for them (Figure 5): windows from the (1) whole genome (all), (2) autosomes only (noX), (3) X chromosome only (onlyX), (4) X chromosome without PAR (noPAR), and (5) pseudoautosomal region only (PAR). For detailed description of comparisons, see Section 2.4. Among tested individuals, we detected statistically significant differences, with heterozygosity of PAR > autosomes (noX) > hemizygous in the male region of X chromosome (noPAR). This pattern is visible in Figure 2D,E for American bison, while for sea otter (Figure 2A–C), it was masked by low heterozygosity and the chosen thresholds for the heatmap.

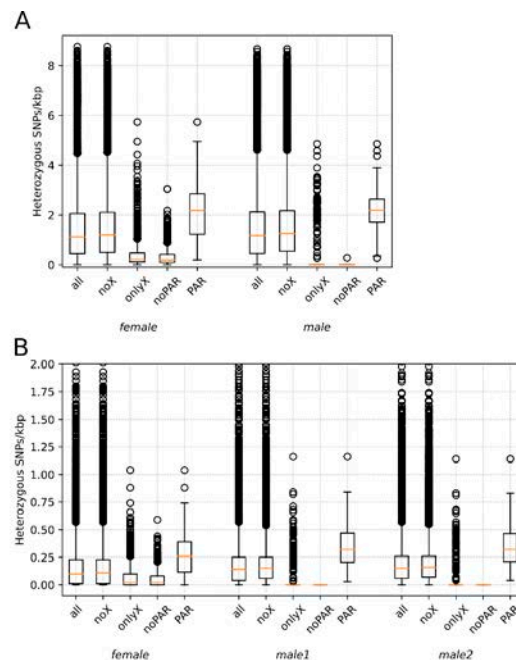


Figure 5. SNP density (SNPs per kbp) in 100 kbp windows inside and outside pseudoautosomal regions in females and males of two mammal species. (A) American bison SRR8588177 (female) and SRR8588180 (male). (B) sea otters SRR8597300 (female), SRR5768046 (male 1), and SRR5768052 (male 2). The abbreviations on the x-axis stand for the following: all—all 100 kbp windows in genome, noX—all without X chromosome, only X—from X chromosome only, noPAR—from X chromosome without PAR, PAR—from pseudoautosomal region only. For the sea otter, part of outlier windows (with more than 2 SNPs per kbp) is not shown.

3.4. STR Marker Localization

We localized 66 STR loci on the chromosomal-level genome assemblies of seven mustelid species (Tables 2, S2 and S4): the sea otter (*E. lutris*), giant otter (*P. brasiliensis*), Asian small-clawed otter (*A. cinereus*), North American river otter (*Lontra canadensis*) and Eurasian (*Lutra lutra*) otter, domestic ferret (*Mustela putorius furo*), and American mink (*Neovison vison*) [40–43]. Among these mustelid species, only the *N. vison* genome assembly has not yet been scaffolded to chromosome level. Nevertheless, it was included in the analysis to serve as a control for our in silico PCR filtering procedure, as described in Section 2.6, because almost one third of all the STR loci in this analysis was originally developed for this particular species [42,43].

The STR markers in this study came from pre-next generation sequencing publications [40–43,47]. To start, we tested 20 different American mink STR markers (Supplementary Table S3). These were denoted according to the source paper either as *Mvi* [42] or *Mvis* [43]. Among these, 18 were successfully mapped to the American mink genome and those also passed our quality criteria, proving the efficiency of our filtration. One locus (*Mvi1272*) was not found in the assembly, and another (*Mvis022*) did not pass the filtration. We further compared our results in a cross-species validation of 7 ermine (*Mustela erminea*) and seven American mink STR loci, denoted in Table S3 as *Mer* and *Mvi*, respectively [42], using the genomes of North American river otter and *E. lutris*. All seven *Mvi* loci and six out of seven ermine loci (all except *Mer041*) were amplified in vitro. Using the same markers and species, we obtained in silico PCR products for seven mink loci—*Mvis* [43] with only one (*Mvis022*) failing the filtering criteria (Table S3). At the same time, three out of seven *M. erminea* markers, *Mer030*, *Mer041*, and *Mer082*, did not amplify. One of these, *Mer082*, did not work in the genomes of any of the analyzed species, while *Mer030* was amplified only in Asian small clawed otter and giant otter, but, in either case, did not pass the filtering criteria. However, *Mer041* resulted in an in silico PCR product for American mink, as well as domestic ferret (not tested in Fleming et al. [42]).

In the six species with chromosome-level genome assemblies, approximately half of the STR loci (between 28 and 31, depending on the species) were mapped (localized), while a quarter failed the quality check, and a quarter were not found (Table 2, Table S2). Overall, the markers originally developed for the American marten (Ma-x), wolverine (Gg-x), and American badger (Tt-x) [41] were the most taxon-specific: the majority of them failed to pass the filtering criteria or did not amplify (Table S3). In contrast, the American mink-derived markers [40,42,43] were the most universal for cross-species usage.

By dividing the number of localized markers by the number of chromosomes, we calculated the approximate density of STR markers to be ~1.5 markers per chromosome. Approximately one quarter of the chromosomes (4–6 depending on the species) did not contain any markers (Figure 6, S1–S5—light grey color), and among the labeled chromosomes, the mean density of markers was only ~2 loci per chromosome.

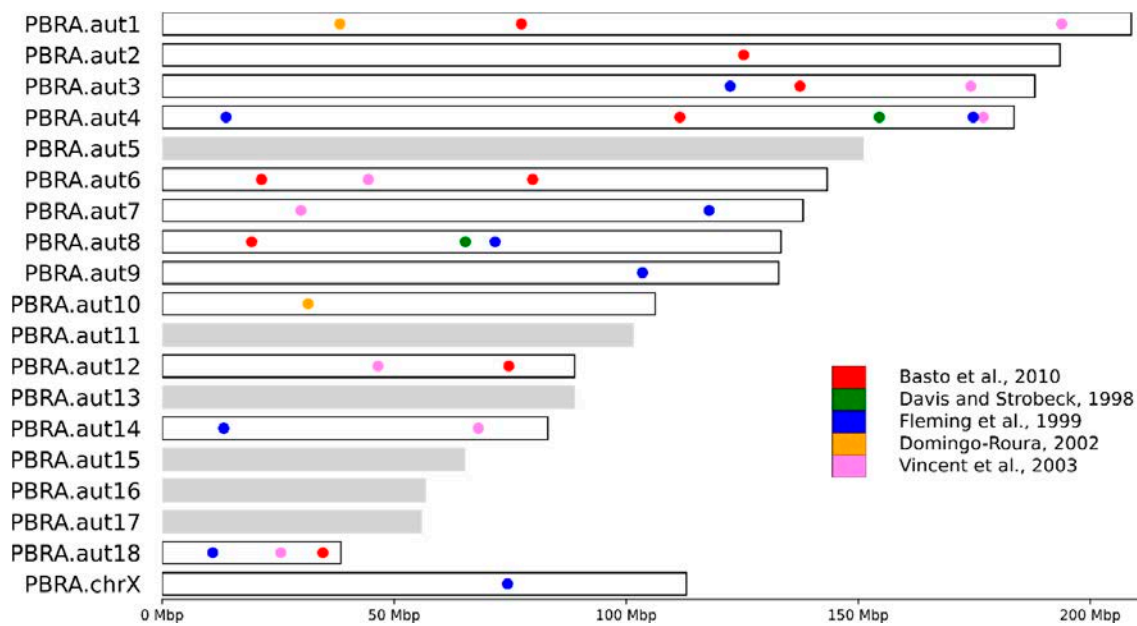


Figure 6. Localization of in silico amplified STR markers on C-scaffolds in the giant otter (*Pteronura brasiliensis*) genome (Basto et al., 2010; Davis and Strobeck, 1998; Fleming et al., 1999; Domingo-Roura, 2002; Vincent et al., 2003) [40–43,47]. The color of the dots indicates the source publication where marker was developed, the light grey bars show unlabeled (no markers) C-scaffolds. Similar localization maps for other mustelids in this study are shown in Supplementary Figures S1–S5.

4. Discussion

4.1. Distribution of Heterozygosity

Genome-wide genetic diversity is usually estimated as heterozygosity—the proportion of sites that contain heterozygous single nucleotide variants across the genome [18]. This yields a single numerical value but does not reveal how variant sites are distributed across the genome, which may be critical for identifying hotspots and cold spots of genetic diversity. A more informative way includes calculation of mean or median heterozygosity in adjacent or overlapping sliding windows of fixed size. The size of the window is a matter of choice depending on the contiguity of the assembly and the questions to be addressed, but a significant part of the genome must be represented in windows to make heterozygosity estimates reliable, especially in the context of runs of homozygosity [48]. For visualization of variant density, a window of 1 Mbp or similar seems to be optimal (Figures 1 and 2), providing clear and easy to understand Figures. However, such window size automatically requires either chromosome-level assemblies or drafts with high N50 to avoid loss of data (Table 3). For the two species with highly fragmented drafts (Asian small clawed otter and giant otter), it was even impossible to perform statistical tests on 1 Mbp windows due to the extremely small numbers of such windows.

Despite significant differences in average genome-wide heterozygosity levels among the species, all eight genomes—of six threatened (three VU and three EN) and two not threatened (one NT and one LC) animals—contained some regions with very low diversity (blue and dark blue regions on Figures 1 and 2). The most striking difference in heterozygosity was observed between different regions in the genome of the giant otter (Figure 2C). Having ~2.5 times higher mean heterozygosity than sea otter [25], the giant otter showed long homozygous stretches (dark blue in Figure 2C) on more than half of its chromosomes.

Assessing and visualizing the distribution of heterozygosity along chromosomes is not the only advantage brought by genomic methods to conservation genetics. Variance in the distribution of diversity and divergence along genomes can be compared between closely related species to detect regions affected by recent and ancient natural selection and introgression [49,50], and annotation of the specific genomic features would help to find

specific functional sites where distribution of genomic changes deviates from that expected from the models based on neutral evolution [51]

4.2. Mapping Sex Chromosomes and PAR

For both coverage-based and diversity-based methods to work correctly in identifying PARs, chromosome-level assemblies are required, as there is no method to distinguish a decrease in X chromosome-dependent coverage from fluctuations in coverage or decreases in X chromosome-dependent heterozygosity from runs of homozygosity in highly fragmented draft assemblies.

Most of the X chromosome in mammals is hemizygous in males and has a lower diversity in females than that along the autosomes, even after the X/A ratio correction [52]. At the same time, its pseudo-autosomal region (PAR) often shows higher levels of heterozygosity [53–55]. Both patterns were observed in the current analysis and were clearly visible on the boxplots (Figure 5) as well as the density maps (Figures 1 and 2). This is because genetic diversity is expected to be higher in the PARs than in the other regions for three different reasons [56]. First, the recombination rate is 5 to 20 times higher in the PAR compared with the genome-wide average [57,58]. If recombination increases the local mutation rate [59–61], this will lead to a higher diversity in PARs than in chromosomal regions that do not recombine. Second, recombination can also unlink alleles affected by selection from nearby sites, lessening the effects of background selection and genetic hitchhiking on decreasing genetic diversity [62,63]. Third, diversity should be higher in PARs due to the larger effective population size compared with the nonrecombining regions of the sex chromosomes, because there are two copies of this region present in both males and females. Therefore, pseudoautosomal regions could be found both in males and females. In males, they could be easily mapped comparing the coverage in windows with the whole genome median coverage (Figure 3). In females (with some exceptions; e.g., Cotter et al., [56]), it could be detected also by examining patterns of heterozygosity (Figure 1F, Figure 2A,D). Our findings (Figure 5) suggest that differences between PARs and hemizygous regions of X chromosome can be observed even in such a homozygous species as the sea otter.

The mammalian X and Y chromosomes are commonly excluded from many types of analysis, such as demographic history reconstruction, because of the complexity of inheritance affecting the localization of genetic diversity [55]. This is easy to do with high contiguity, chromosome-level assemblies, because sex chromosomes can be readily detected using a limited number of linked markers, while with more fragmented short-read draft assemblies, the identification of sex chromosomes requires whole genome alignment of scaffolds to the C-scaffolds corresponding to the X and Y chromosomes (if assembled) of related species followed by checking of sequence coverage.

4.3. STR Marker Localization

Whole genome sequencing of hundreds of individuals is still too expensive for conservation biology studies, resulting in the common usage of low-resolution methods, such as STR panels, or reduced representation approaches, such as restriction site-associated DNA sequencing (RADseq) [64,65]. The issue with markers such as STR loci, developed in the pre-NGS era, is that they are often not localized on chromosomes, and the relatively small number of loci applied in studies (10–50) is used as a proxy for genome-wide heterozygosity and other assessments. The existing datasets, especially in the historic population studies, can now be merged and compared with the new estimated based on the STR loci that are commonly extracted from the whole genome sequencing in the more recent studies.

Localization of markers on chromosomes is crucially important in studies of interspecific hybridization. It is clear that complex structures like these require labeling of each chromosome, or possibly even each arm of each chromosome, in order to identify and classify hybrids correctly. Our analysis demonstrates that even using a large set (66) of mustelid STR markers developed in the five studies of the pre-genomic era with previously

unknown localization, some of the chromosomes were missed in the analysis (Figure 5, chromosomes shaded in grey).

Levels of hybridization, gene flow, and population structure can be very complex, especially where two or more closely related species occur sympatrically. For example, a case of fertile or partly fertile F1 hybrids resulting in backcrosses with parental and maternal species and mosaic F2 hybrids was recently reported for European (*Meles meles*) and Asian (*M. leucurus*) badgers [66]. Both species have 22 pairs of chromosomes ($2n = 44$), but only 9 microsatellite markers were used in this study. Therefore, 40% of chromosomal linkage groups were included in the analysis, and the remaining 60% were not evaluated. This automatically raises another question: were the individuals reported as F1 by Kinoshita et al. [66] really from the F1 generation, or, alternatively, were some of them F2-s or even backcrosses? Unfortunately, at this point, we do not have a definitive answer to this question due to the lack of data. The absence of chromosomal assemblies clearly diminished the certainty of hybridization studies, as mentioned above in the case of badgers and in the investigation of hybrids between sable (*Martes zibellina*) and pine marten (*M. martes*) [67].

5. Conclusions

This study compared highly fragmented draft genome assemblies and recently generated chromosome-level genome assemblies of the eight mammalian species. The analyses of whole genome resequencing data require generation of a reference genome assembly from either the same species or a closely related species. Chromosome-level assemblies can be generated using a combination various long-read or short-read sequencing technologies [23,68]. Inexpensive contiguous assemblies can be generated using Hi-C scaffolding of the existing short-read draft assemblies where N50 of the draft contigs is as low as 0.1% of the estimated length of the genome. Chromosome-length assemblies provide additional benefits, including simplifying the design of STR panels and allowing assessment of previously selected loci. With the help of Hi-C scaffolding, contiguity has been remarkably improved, and we can conclude that chromosome-level genome assemblies provide a more informative way to directly visualize genome-wide genetic diversity. The results of these comparisons illustrate an improvement in representation of genetic diversity, localization, and visualization of heterozygosity across the genomes. The improved understanding of the distribution of heterozygosity and localization of SNP and STR markers afforded by chromosome-level assemblies is particularly applicable for conservation studies of endangered species.

The International Union for Conservation of Nature of Threatened Species now classifies 37,400 species as threatened with extinction in the 2021 edition of the Red List, which is approximately 28% of all assessed species, almost three times the number reported only a decade ago [13,46]. The Earth's biota may be in the middle of a mass extinction event caused by the adverse impact of anthropogenic activities [69]. At the same time, due to the concerted effort of the genomics community, there is an increased accessibility to chromosome-level assemblies, such as through the Vertebrate Genomes Project consortium, which aims to generate highly contiguous, chromosome-scaffolded assemblies for all ~70,000 vertebrate species using a combination of long-read and Hi-C approaches [68]. Therefore, a comprehensive evaluation of the remaining adaptive potential in endangered species may soon become possible. The application of contiguous chromosome-level assemblies allowed us to localize and visualize low heterozygosity regions within genomes. Applied to the conservation genetics research, this can improve our understanding of the factors contributing to the variation in genome-wide diversity and, hence, potentially help us to devise better evidence-based strategies for endangered species. It will allow us to understand how to design effective conservation strategies [2,3] and, hopefully, avoid the worst-case scenario in conservation biology—species extinction.

Supplementary Materials: The following are available online at <https://www.mdpi.com/article/10.3390/genes12091336/s1>. Figures S1–S5. Localization of in silico amplified STR markers on C-scaffolds for four mustelids species (*Aonyx cinereus*, *Enhydra lutris*, *Lontra canadensis*, *Lutra lutra*, *Mustela putorius furo*); Table S1. Assembly parameters for the genomes; Table S2. SRA IDs of short-read libraries used for variant calling; Table S3. Results of in silico PCR for 66 analyzed STR loci; Table S4. C-scaffold IDs from used genome assemblies and their synonyms shown in Figures; Table S5. Coordinates of localized STR loci (L-category) in six mustelid species; Table S6. Coordinates of pseudoautosomal region in eight species analyzed.

Author Contributions: Conceptualization, S.K., T.K.O. and L.D.; methodology, S.K. and T.K.O.; software, A.T. (Andrey Tomarovsky) and D.R.; validation, A.T. (Azamat Totikov), A.T. (Andrey Tomarovsky) and D.P.; formal analysis, A.T. (Azamat Totikov), A.T. (Andrey Tomarovsky), D.P., A.Y. and S.K.; investigation, A.T. (Azamat Totikov), A.T. (Andrey Tomarovsky), A.Y., S.K. and D.P.; writing—original draft preparation, A.T. (Azamat Totikov), A.T. (Andrey Tomarovsky) and S.K.; writing—review and editing, S.K., T.K.O., W.W.W., K.-P.K., T.B. and L.D.; visualization, A.T. (Azamat Totikov), A.T. (Andrey Tomarovsky) and S.K.; supervision, S.K.; project administration, S.K.; funding acquisition, S.K. and T.K.O. All authors have read and agreed to the published version of the manuscript.

Funding: The reported study was funded by the Russian Foundation for Basic Research (RFBR), project number 20-04-00808, T.K.O. and W.W.W. were supported by the startup funds of Oakland University, MI. A.T. (Azamat Totikov), A.T. (Andrey Tomarovsky) and A.Y. were additionally funded from JetBrains Research.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data used in this study are publicly available at NCBI as indicated.

Acknowledgments: We acknowledge Olga Dudchenko, Erez Lieberman-Aiden, and Ruqayya Khan (The Center for Genome Architecture, Baylor College of Medicine and Rice University) for their efforts in building an open resource repository of chromosome-level genome assemblies at dnazoo.org, which made this study possible.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Stange, M.; Barrett, R.D.H.; Hendry, A.P. The importance of genomic variation for biodiversity, ecosystems and people. *Nat. Rev. Genet.* **2021**, *22*, 89–105. [[CrossRef](#)] [[PubMed](#)]
2. Mable, B.K. Conservation of adaptive potential and functional diversity: Integrating old and new approaches. *Conserv. Genet.* **2019**, *20*, 89–100. [[CrossRef](#)]
3. Rus Hoelzel, A.; Bruford, M.W.; Fleischer, R.C. Conservation of adaptive potential and functional diversity. *Conserv. Genet.* **2019**, *20*, 1–5. [[CrossRef](#)]
4. Ellegren, H.; Galtier, N. Determinants of genetic diversity. *Nat. Rev. Genet.* **2016**, *17*, 422–433. [[CrossRef](#)]
5. Oleksyk, T.K.; Smith, M.W.; O'Brien, S.J. Genome-wide scans for footprints of natural selection. *Philos. Trans. R. Soc. B Biol. Sci.* **2010**, *365*, 185–205. [[CrossRef](#)] [[PubMed](#)]
6. Hoffmann, A.; Sgrò, C. Climate change and evolutionary adaptation. *Nature* **2011**, *470*, 479–485. [[CrossRef](#)]
7. Sgrò, C.M.; Lowe, A.J.; Hoffmann, A.A. Building evolutionary resilience for conserving biodiversity under climate change. *Evol. Appl.* **2011**, *4*, 326–337. [[CrossRef](#)] [[PubMed](#)]
8. Reid, N.M.; Proestou, D.A.; Clark, B.W.; Warren, W.C.; Colbourne, J.K.; Shaw, J.R.; Karchner, S.I.; Hahn, M.E.; Nacci, D.; Oleksiak, M.F.; et al. The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. *Science* **2016**, *354*, 1305–1308. [[CrossRef](#)] [[PubMed](#)]
9. Jones, F.C.; Grabherr, M.G.; Chan, Y.F.; Russell, P.; Mauceli, E.; Johnson, J.; Swofford, R.; Pirun, M.; Zody, M.C.; White, S.; et al. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **2012**, *484*, 55–61. [[CrossRef](#)] [[PubMed](#)]
10. Visser, M.E. Keeping up with a warming world; assessing the rate of adaptation to climate change. *Proc. R. Soc. B Biol. Sci.* **2008**, *275*, 649–659. [[CrossRef](#)] [[PubMed](#)]
11. Visser, M.E. Evolution: Adapting to a Warming World. *Current Biology. Curr. Biol.* **2019**, *18*, R1189–R1191. [[CrossRef](#)] [[PubMed](#)]
12. Lai, Y.T.; Yeung, C.K.L.; Omland, K.E.; Pang, E.L.; Hao, Y.; Liao, B.Y.; Cao, H.F.; Zhang, B.W.; Yeh, C.F.; Hung, C.M.; et al. Standing genetic variation as the predominant source for adaptation of a songbird. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 2152–2157. [[CrossRef](#)] [[PubMed](#)]

13. Johnson, W.E.; Koepfli, K. The Role of Genomics in Conservation and Reproductive Sciences. *Reprod. Sci. Anim. Conserv.* **2014**. [CrossRef]
14. Luikart, G.; England, P.R.; Tallmon, D.; Jordan, S.; Taberlet, P. The power and promise of population genomics: From genotyping to genome typing. *Nat. Rev. Genet.* **2003**, *4*, 981–994. [CrossRef]
15. Reed, D.H.; Frankham, R. Correlation between Fitness and Genetic Diversity. *Conserv. Biol.* **2003**, *17*, 230–237. [CrossRef]
16. Spielman, D.; Brook, B.W.; Frankham, R. Most species are not driven to extinction before genetic factors impact them. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 15261–15264. [CrossRef]
17. Kirk, H.; Freeland, J.R. Applications and Implications of Neutral versus Non-neutral Markers in Molecular Ecology. *Int. J. Mol. Sci.* **2011**, *12*, 3966–3988. [CrossRef]
18. Nei, M. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* **1978**, *89*, 583–590. [CrossRef] [PubMed]
19. Amos, W.; Balmford, A. When does conservation genetics matter? *Heredity* **2001**, *87*, 257–265. [CrossRef] [PubMed]
20. Jost, L.; Archer, F.; Flanagan, S.; Gaggiotti, O.; Hoban, S.; Latch, E. Differentiation measures for conservation genetics. *Evol. Appl.* **2018**, *11*, 1139–1148. [CrossRef] [PubMed]
21. McMahon, B.J.; Teeling, E.C.; Höglund, J. How and why should we implement genomics into conservation? *Evol. Appl.* **2014**, *7*, 999–1007. [CrossRef]
22. Armstrong, J.; Fiddes, I.T.; Diekhans, M.; Paten, B. Whole-Genome Alignment and Comparative Annotation. *Annu Rev Anim Biosci.* **2019**, *7*, 41–64. [CrossRef] [PubMed]
23. Dudchenko, O.; Shamim, M.S.; Batra, S.S.; Durand, N.C.; Musial, N.T.; Mostofa, R.; Pham, M.; Glenn St Hilaire, B.; Yao, W.; Stamenova, E.; et al. The juicebox assembly tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under \$1000. *bioRxiv* **2018**. [CrossRef]
24. Hu, Y.; Wu, Q.; Ma, S.; Ma, T.; Shan, L.; Wang, X.; Nie, Y.; Ning, Z.; Yan, L.; Xiu, Y.; et al. Comparative genomics reveals convergent evolution between the bamboo-eating giant and red pandas. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 1081–1086. [CrossRef]
25. Beichman, A.C.; Koepfli, K.P.; Li, G.; Murphy, W.; Dobrynin, P.; Kliver, S.; Tinker, M.T.; Murray, M.J.; Johnson, J.; Lindblad-Toh, K.; et al. Aquatic Adaptation and Depleted Diversity: A Deep Dive into the Genomes of the Sea Otter and Giant Otter. *Mol. Biol. Evol.* **2019**, *36*. [CrossRef] [PubMed]
26. de Manuel, M.; Barnett, R.; Sandoval-Velasco, M.; Yamaguchi, N.; Vieira, F.G.; Lisandra Zepeda Mendoza, M.; Liu, S.; Martin, M.D.; Sinding, M.H.S.; Mak, S.S.T.; et al. The evolutionary history of extinct and living lions. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 10927–10934. [CrossRef] [PubMed]
27. Dobrynin, P.; Liu, S.; Tamazian, G.; Xiong, Z.; Yurchenko, A.A.; Krasheninnikova, K.; Kliver, S.; Schmidt-Küntzel, A.; Koepfli, K.P.; Johnson, W.; et al. Genomic legacy of the African cheetah, *Acinonyx jubatus*. *Genome Biol.* **2015**. [CrossRef] [PubMed]
28. Hoff, J.L.; Decker, J.E.; Schnabel, R.D.; Taylor, J.F. Candidate lethal haplotypes and causal mutations in Angus cattle. *BMC Genom.* **2017**, *18*. [CrossRef] [PubMed]
29. Andrews, S. FastQC A Quality Control tool for High Throughput Sequence Data. *Babraham Bioinform.* **2020**. Available online: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (accessed on 20 January 2021).
30. Kliver, S.F. KrATER (K-Mer Analysis Tool Easy to Run). 2017. Available online: <https://github.com/mahajrod/KrATER> (accessed on 15 January 2021).
31. Starostina, E.; Tamazian, G.; Dobrynin, P.; O'Brien, S.; Komissarov, A. Cookiecutter: A tool for kmer-based read filtering and extraction. *BioRxiv* **2015**. [CrossRef]
32. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef] [PubMed]
33. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef] [PubMed]
34. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef] [PubMed]
35. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **2011**, *27*, 2987–2993. [CrossRef] [PubMed]
36. Quinlan, A.R.; Hall, I.M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **2010**, *26*, 841–842. [CrossRef]
37. Graphodatsky, A.; Perelman, P.; O'Brien, S.J. *Atlas of Mammalian Chromosomes*; John Wiley & Sons, Incorporated: Hoboken, NJ, USA, 2020.
38. Frith, M.C.; Kawaguchi, R. Split-alignment of genomes finds orthologies more accurately. *Genome Biol.* **2015**, *16*. [CrossRef]
39. Hunter, J.D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. [CrossRef]
40. Basto, M.P.; Rodrigues, M.; Santos-Reis, M.; Bruford, M.W.; Fernandes, C.A. Isolation and characterization of 13 tetranucleotide microsatellite loci in the Stone marten (*Martes foina*). *Conserv. Genet. Resour.* **2010**, *2*, 317–319. [CrossRef]
41. Davis, C.S.; Strobeck, C. Isolation, variability, and cross-species amplification of polymorphic microsatellite loci in the family mustelidae. *Mol. Ecol.* **1998**, *7*, 1776–1778. [CrossRef]

42. Fleming, M.A.; Ostrander, E.A.; Cook, J.A. Microsatellite markers for american mink (*Mustela vison*) and ermine (*Mustela erminea*). *Mol. Ecol.* **1999**, *8*, 1352–1355. [[CrossRef](#)]
43. Vincent, I.R.; Farid, A.; Otieno, C.J. Variability of thirteen microsatellite markers in American mink (*Mustela vison*). *Can. J. Anim. Sci.* **2003**, *83*, 597–599. [[CrossRef](#)]
44. Gardner, S.N.; Slezak, T. Simulate_PCR for amplicon prediction and annotation from multiplex, degenerate primers and probes. *BMC Bioinform.* **2014**, *15*, 237. [[CrossRef](#)] [[PubMed](#)]
45. Lewin, H.A.; Graves, J.A.M.; Ryder, O.A.; Graphodatsky, A.S.; O'Brien, S.J. Precision nomenclature for the new genomics. *GigaScience* **2019**. [[CrossRef](#)] [[PubMed](#)]
46. IUCN. The IUCN Red List of Threatened Species [WWW Document]. Version 2021-ISSN 2307-8235. 2021. Available online: <https://www.iucnredlist.org> (accessed on 15 January 2021).
47. Domingo-Roura, X. Genetic distinction of marten species by fixation of a microsatellite region. *J. Mammal.* **2002**, *83*, 907–912. [[CrossRef](#)]
48. Renaud, G.; Hanghøj, K.; Korneliussen, T.S.; Willerslev, E.; Orlando, L. Joint Estimates of Heterozygosity and Runs of Homozygosity for Modern and Ancient Samples. *Genetics* **2019**, *212*, 587–614. [[CrossRef](#)] [[PubMed](#)]
49. Guiblet, W.M.; Zhao, K.; O'Brien, S.J.; Massey, S.E.; Roca, A.L.; Oleksyk, T.K. SmileFinder: A resampling-based approach to evaluate signatures of selection from genome-wide sets of matching allele frequency data in two or more diploid populations. *GigaScience* **2015**. [[CrossRef](#)] [[PubMed](#)]
50. Oleksyk, T.K.; Zhao, K.; de La Vega, F.M.; Gilbert, D.A.; O'Brien, S.J.; Smith, M.W. Identifying selected regions from heterozygosity and divergence using a light-coverage genomic dataset from two human populations. *PLoS ONE* **2008**, *3*, e1712. [[CrossRef](#)] [[PubMed](#)]
51. Volfovsky, N.; Oleksyk, T.K.; Cruz, K.C.; Truelove, A.L.; Stephens, R.M.; Smith, M.W. Genome and gene alterations by insertions and deletions in the evolution of human and chimpanzee chromosome 22. *BMC Genom.* **2009**, *10*. [[CrossRef](#)]
52. Osada, N.; Nakagome, S.; Mano, S.; Kameoka, Y.; Takahashi, I.; Terao, K. Finding the factors of reduced genetic diversity on X chromosomes of *Macaca fascicularis*: Male-driven evolution, demography, and natural selection. *Genetics* **2013**, *195*, 1027–1035. [[CrossRef](#)]
53. Flaquer, A.; Rappold, G.A.; Wienker, T.F.; Fischer, C. The human pseudoautosomal regions: A review for genetic epidemiologists. *Eur. J. Hum. Genet.* **2008**, *16*, 771–779. [[CrossRef](#)] [[PubMed](#)]
54. Otto, S.P.; Pannell, J.R.; Peichel, C.L.; Ashman, T.-L.; Charlesworth, D.; Chippindale, A.K.; Delph, L.F.; Guerrero, R.F.; Scarpino, S.V.; McAllister, B.F. About PAR: The distinct evolutionary dynamics of the pseudoautosomal region. *Trends Genet.* **2011**, *27*, 358–367. [[CrossRef](#)] [[PubMed](#)]
55. Wilson-Sayres, M.A. Genetic Diversity on the Sex Chromosomes. *Genome Biol. Evol.* **2018**, *10*, 1064–1078. [[CrossRef](#)] [[PubMed](#)]
56. Cotter, D.J.; Brotman, S.M.; Wilson Sayres, M.A. Genetic Diversity on the Human X Chromosome Does Not Support a Strict Pseudoautosomal Boundary. *Genetics* **2016**, *203*, 485–492. [[CrossRef](#)] [[PubMed](#)]
57. Filatov, D.A.; Gerrard, D.T. High mutation rates in human and ape pseudoautosomal genes. *Gene* **2003**, *317*, 67–77. [[CrossRef](#)]
58. Lien, S.; Szyda, J.; Schechinger, B.; Rappold, G.; Arnheim, N. Evidence for heterogeneity in recombination in the human pseudoautosomal region: High resolution analysis by sperm typing and radiation-hybrid mapping. *Am. J. Hum. Genet.* **2000**, *66*, 557–566. [[CrossRef](#)] [[PubMed](#)]
59. Hellmann, I.; Ebersberger, I.; Ptak, S.E.; Pääbo, S.; Przeworski, M. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am. J. Hum. Genet.* **2003**, *72*, 1527–1535. [[CrossRef](#)] [[PubMed](#)]
60. Huang, S.W.; Friedman, R.; Yu, N.; Yu, A.; Li, W.H. How strong is the mutagenicity of recombination in mammals? *Mol. Biol. Evol.* **2005**, *22*, 426–431. [[CrossRef](#)] [[PubMed](#)]
61. Perry, J.; Ashworth, A. Evolutionary rate of a gene affected by chromosomal position. *Curr. Biol.* **1999**, *9*. [[CrossRef](#)]
62. Charlesworth, B. The effects of deleterious mutations on evolution at linked sites. *Genetics* **2012**, *190*, 5–22. [[CrossRef](#)]
63. Vicoso, B.; Charlesworth, B. Evolution on the X chromosome: Unusual patterns and processes. *Nat. Rev. Genet.* **2006**, *7*, 645–653. [[CrossRef](#)]
64. Zimmerman, S.J.; Aldridge, C.L.; Oyler-McCance, S.J. An Empirical Comparison of Population Genetic Analyses Using Microsatellite and SNP Data for a Species of Conservation Concern. *BMC Genom.* **2020**, *21*, 382. [[CrossRef](#)] [[PubMed](#)]
65. Andrews, K.R.; Good, J.M.; Miller, M.R.; Luikart, G.; Hohenlohe, P.A. Harnessing the Power of RADseq for Ecological and Evolutionary Genomics. *Nat. Rev. Genet.* **2016**, *17*, 81–92. [[CrossRef](#)]
66. Kinoshita, E.; Abramov, A.V.; Soloviev, V.A.; Saveljev, A.P.; Nishita, Y.; Kaneko, Y.; Masuda, R. Hybridization between the European and Asian Badgers (*Meles, Carnivora*) in the Volga-Kama Region, Revealed by Analyses of Maternally, Paternally and Biparentally Inherited Genes. *Mamm. Biol.* **2019**, *94*, 140–148. [[CrossRef](#)]
67. Rozhnov, V.V.; Pishchulina, S.L.; Meschersky, I.G.; Simakin, L.V. On the ratio of phenotype and genotype of sable and pine marten in sympatry zone in the Northern Urals. *Mosc. Univ. Biol. Sci. Bull.* **2013**, *68*, 178–181. [[CrossRef](#)]
68. Rhie, A.; McCarthy, S.A.; Fedrigo, O.; Damas, J.; Formenti, G.; Koren, S.; Uliano-Silva, M.; Chow, W.; Fungtammasan, A.; Kim, J.; et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **2021**, *592*, 737–746. [[CrossRef](#)]
69. Ceballos, G.; Ehrlich, P.R.; Barnosky, A.D.; García, A.; Pringle, R.M.; Palmer, T.M. Accelerated modern human-induced species losses: Entering the sixth mass extinction. *Sci. Adv.* **2015**, *1*, e1400253. [[CrossRef](#)]

General discussion

Aims and importance of this dissertation

Comparative genomics represents a means to understand how biological diversity arises, and thus also a first step in safeguarding it for the future (Alföldi and Lindblad-Toh 2013; Capilla et al. 2016). This approach enables the identification of different types of genomic variation, including single nucleotide differences, changes in gene family sizes, and structural variants potentially impacting several to hundreds of genes.

In my dissertation, I advocate the inclusion of structural variants as a crucial piece of information in our understanding of the full extent of genomic variation within and between species, and its effect on adaptive evolution. By cataloguing the whole spectrum of genomic variation, we can gain insight into the mechanisms that create genomic diversity. Therefore, when investigating biological diversity and its genomic roots, it is important to take into account multiple types of variation in order to obtain a more comprehensive insight into an impact on genome evolution.

The novel findings presented in the four chapters of this thesis, along with the contribution of the bioinformatics pipeline and genome assembly, represent valuable scientific resources that can be freely used by the research community to build upon and expand the future research of these species, interaction of their genomes and environment, and the adaptive potential their genomic differences may harbour.

Association of structural variation with trait-related genes

To detect structural variants associated with relevant traits, we need to focus on genic SVs, as these represent genomic variation that contributes to phenotypic variation and may subsequently affect adaptive phenotypes. This also has implications for conservation genomics, a field where the contribution of SVs to genetic diversity has largely gone uncharacterized. For this reason, I examined structural variation in genomes of several mustelid species, including the endangered black-footed ferret. SVs can impact RNA splicing patterns, mRNA translational efficiency, and subsequently protein functions, leading to the production of protein isoforms with different structural and functional properties, that create genomic and biological diversity (Xing and Lee 2006; Wang et al. 2015). Additionally, the overall gene repertoire of genomes can change through SVs, by the loss (deletion) or gain (duplication) of genes.

Inversions in particular have a strong impact on genome evolution by suppressing recombination and protecting favourable allele combinations in heterozygotes (Dobigny, Britton-Davidian, and Robinson 2017; Hammer, Schimenti, and Silver 1989). Remarkably, a 900-kb long inversion polymorphism showing strong linkage disequilibrium was found to be associated with an increased number of children in the Icelandic population, representing a notable example of natural selection in the human population (Stefansson et al. 2005). Studies of the evolution of primate lineages have found that such polymorphic rearrangements are implicated in population diversification (Stefansson et al. 2005; Alves et al. 2012) and speciation (Feuk et al. 2005; Porubsky et al. 2020). I suggest that this aspect of structural variation is one of the likely facilitators of adaptive radiation and the establishment of the Mustelidae as the most ecologically and taxonomically diverse family within the mammalian order Carnivora.

In the first chapter, I investigated genomic variation in species of the mustelid subfamily Guloninae that differ in feeding ecologies, reproductive strategies and morphology with the goal of associating candidate loci with adaptations to their respective environments. This was achieved through comparative genomic analyses of genomes of tayra (*Eira barbara*), wolverine (*Gulo gulo*) and sable (*Martes zibellina*). I found SVs overlapping numerous candidate genes associated with species-specific traits related to diet, body condition and reproduction. These SVs associated with candidate loci are primarily deletions or duplications of exons and introns of protein-coding genes, implying the possibility of both regulatory and functional gene modifications. Unlike other gulonine species, tayras do not exhibit embryonic diapause and are aseasonal breeders with multiple oestrous cycles per year (Poglayen-Neuwall et al. 1989; Proulx and Aubry 2017). Notably, I observed species-specific SVs in pregnancy-related genes in tayra, involved in placental development, implantation and embryogenesis, while fewer SNPs have been detected in these Gene Ontology categories.

In the wolverine and sable, species that must cope with seasonal food scarcity in the northern Palaearctic (Inman et al. 2012; Lukacs et al. 2020), besides the SVs primarily associated with cell cycle genes, SVs were identified in genes associated with diet, body condition and development. I observed deletions in a gene related to thermoregulation in both species, indicating independent modification of the gene in gulonines inhabiting colder environments. Likewise, another independent deletion was detected in sable and wolverine in a gene associated with the regulation of insulin homeostasis. I suggest that this gene modification may impact “adaptive fasting” in these species, an adaptation to prolonged periods of nutrient deprivation observed in several carnivorans (Viscarra et al. 2013; Martinez and Ortiz 2017). Furthermore, in the hypercarnivorous wolverine, I did not observe candidate loci impacted by structural variation associated with carbohydrate metabolism and omnivorous diet, while several were detected in the omnivorous tayra, most likely due to fruits and honey making a considerable portion of this species’ diet throughout the year (Soley and Alvarado-Díaz 2011; Heldstab et al. 2018).

In the second chapter, I investigated SVs in the genus *Mustela*, the largest group within the family Mustelidae. SVs were characterized in genomes of the least weasel (*Mustela nivalis*), steppe polecat (*Mustela eversmannii*), European polecat (*Mustela putorius*) and three black-footed ferrets (*Mustela nigripes*), in relation to the domestic ferret (*Mustela putorius furo*) as a reference. The highest number of SVs was detected in *M.nivalis* (> 7000), and the lowest number in *M. putorius* (~10% of the *M.nivalis* count). These findings suggest a relationship between the number of SVs detected and the relatedness of a given species to the domestic ferret, and to some extent biases due to the chosen reference. An increasing number of SVs was detected in more distantly related species in accordance with the known phylogenies (Law, Slater, and Mehta 2018; Koepfli et al. 2008) showing evidence for the accumulation of SVs over time. Functions of genes overlapped by SVs in the mustelid species are primarily associated with the cell cycle, nervous system (mainly sensory perception), metabolism and immune system.

Moreover, I have examined SVs in three black-footed ferret individuals in more detail. Black-footed ferrets are one of the most endangered mammalian species of North America, affected by low population size and inbreeding (Wisely et al. 2002; Santymire et al. 2019). Not surprisingly, a low number of SVs was detected in these individuals. Around 400 SVs are shared among all three individuals in relation to the domestic ferret genome, with 2.6% of these SVs overlapping protein-coding regions. Trait-relevant genes affected by SVs follow similar patterns in black-footed ferrets as previously noted for all four mustelid species. Overall, most of the structural variants were flagged as heterozygous in all species, with deletions and inversions being the most frequent SV types. Correspondingly, I observed a higher number of heterozygous sample-specific SVs in black-footed ferret individuals compared to homozygous ones. This represents a possible source of variation, as large parts of their genome consist of runs of homozygosity (ROH). These heterozygous SVs include high-impact variants that are not shared among the three individuals. There is

important genomic diversity in the black-footed ferret gene pool that is not detected by examining the SNP diversity alone.

In these two chapters, I have demonstrated that structural variation impacts trait-related genes and is potentially involved in the adaptive genomic evolution of mustelids. However, the patterns of selection and inheritance of these variants are currently unknown. Studying the dynamics of these evolutionary aspects is complex in wild populations due to the strong effect of the environment influencing phenotypic variation. Alas, I did not have transcriptome data to directly infer if the changes in the genome observed affect gene expression and thus phenotype. While I was able to use predictive software to address this, future studies of these species should strive to include such types of data, to strengthen the link between SVs and trait evolution.

The four concluding remarks of this section are:

1. If genetic diversity is only investigated at the nucleotide level (SNPs), a large part of genomic variation may be missed; including structural variation that can impact gene function and expression.
2. All species were polymorphic for some SVs. In fact, in some species, the majority of detected SVs were heterozygous, strongly suggesting that there is standing variation present as SV in the gene pools of these species.
3. Even in genomes of species with low SNP variation, such as the black-footed ferret, I detected many polymorphic SVs, including some in genic regions.
4. While most SVs accumulated over time in the studied mustelid lineages are species-specific, some are shared among even relatively distantly related species.

Artificial selection and accumulation of structural variants in trait-related genes

It was previously recognized that structural variation associated with notable phenotypic differences is subjected to selection during domestication (Paudel et al. 2013). For example, a duplication of the agouti signalling protein gene results in a white coat colouration in sheep (Norris and Whan 2008). Furthermore, genes related to metabolic activity and production traits were shown to be affected by SVs during the selection of other domesticated species, such as milk composition in cattle (Gao et al. 2017), high fertility in goats (Zhang et al. 2019), and several traits in pigs (Chen et al. 2012).

The artificial selection represents a valuable method to study adaptive genetic responses within populations. Here I examined if as a result of selection, there is an accumulation of SVs associated with trait-linked genes. To investigate this, I analyzed genome sequences of mouse lines (125 individuals) that were artificially selected for high fertility, increased body mass and high endurance. These traits were selected for over the course of more than 140 generations and are represented with distinct phenotypes in each of five lines relative to the unselected control line (only exposed to genetic drift).

During this long-lasting experiment, the reproductive performance has doubled in both fertility lines, individuals of the high body mass lines (mice with “obese” and muscular phenotype) have become considerably larger and heavier, and mice selected for endurance covered on average three times more distance compared to the control line. With the exception of the mouse line displaying the obese phenotype, each one of the trait-selected lines has developed an extreme phenotype without obvious detrimental effects on their general health and longevity.

In relation to the reference mouse genome, I found that structural variants were on average twice as abundant in the trait-selected lines compared to the control line, which is the least divergent from the reference. Most SNPs were fixed in all trait-selected mouse lines. This occurred most likely due to small population sizes, reproductive isolation and genetic drift, along with the positive selection on trait-related SNPs (Foote et al. 2016). Contrary to this, I found that the majority of SVs in all mouse lines were polymorphic. For some of these variants, this may imply heterosis or heterozygote advantage, where heterozygotes for a given locus have higher fitness than homozygotes. Observed in several domesticated species, heterozygotes were found at high frequencies in populations, with the same variant having detrimental effects, often quite severe, when present in a homozygous state (Hedrick 2012; Leffler et al. 2013; Hedrick 2015). The advantage of SV polymorphism is associated with fecundity in sheep (Gemmell and Slate 2006), litter size in pigs (Sironen et al. 2012),

muscle milk yields in cattle (Kadri et al. 2014), coat colouration in horses (Bellone et al. 2013) and hair morphology in dogs (Karlsson et al. 2007), respectively.

I found that line-specific SVs that overlap protein-coding genes mostly comprised deletions and inversions. Duplications were detected in lower numbers, likely due to the reduced detection power caused by low and variable sequencing coverage within the dataset. Insertions were found in a low number of individuals, below the detection confidence threshold, and therefore not included in further analysis. The most gene-rich functional groups were those associated with sensory perception, predominantly olfaction (detected in the fertility lines), followed by the cell cycle, metabolism and body condition, reproduction, immunity, and others.

The composition of detected SVs linked to trait-related genes in all five selected lines is the following:

Some of the notable findings pertain to the two fertility lines. These mice share the same evolutionary history, they were bred according to the same criteria, and have achieved comparable litter sizes, with more than double the number of offspring per litter since the beginning of the selection. Despite these shared characteristics, these lines achieve improved fertility via different physiological pathways (Langhammer et al. 2014).

High fertility line 1 - females from this fertility line showed an increased progesterone level compared to both the control line and the fertility line 2. Progesterone is a critical steroid hormone that regulates pregnancy maintenance and mammary gland development (Arck et al. 2007; Taraborrelli 2015). In line with this observation, I found that a gene involved in the preparation of the endometrium for implantation and pregnancy, and progesterone signalling (Lei et al. 2012) is affected by a short heterozygous deletion.

High fertility line 2 - in contrast to fertility line 1, females from fertility line 2 exhibit follicles containing a higher number of oocytes compared to the control and another fertility line. The formation of multiple oocytes was related to a higher number of offspring per litter in sheep (McNatty et al. 2017). Similarly, follicles with increased numbers of oocytes have been observed in dogs (Payan-Carreira and Pires 2008), cats (Bristol-Gould and Woodruff 2006) and mice (Alm et al. 2010) associated with offspring number. Among the genes related to oocyte development, the majority of candidate loci harboured single nucleotide changes.

Several SVs I detected were related to broader aspects of reproduction. A considerably high number of SVs overlapped genes encoding olfactory receptors and the vomeronasal chemosensory system in fertility lines 1 and 2, respectively. Olfactory receptors, among other important functions, have been found to have a role in fertilisation. They are involved in the detection of chemical cues by spermatozoa when locating the oocyte in the female reproductive tract (Eisenbach and Giojalas 2006; Flegel et al. 2015). The vomeronasal organ is found in the nose of most mammals and is involved with olfaction that initiates innate behavioural responses. Such chemical communication, for example, is critical for

learning the smell of a mother by offspring to guide suckling interactions in mice and rats (Logan et al. 2012; Ibarra-Soria, Levitin, and Logan 2014). I found a high number of these olfaction-related genes duplicated or inverted in individuals in fertility lines, which may imply their importance in chemical perception and recognition in the case of large litters. Similarly, olfaction-related responses associated with litter size have been found in California mice (Wilson, Wagner, and Saltzman 2022).

Both lines selected for increased body mass, exhibit structural variation in genomic regions containing genes related to metabolism, energy conversion and body condition in agreement with their differing phenotypes.

Muscular phenotype line - candidate SNPs in mice with muscular phenotype conform with growth-related major quantitative trait loci found in sheep (Xu et al. 2020). These loci are also known to influence stature and body size in cattle (Taye et al. 2018) and pigs (Jiao et al. 2014). In line with this, I found a heterozygous deletion overlapping a gene involved in limb development (Schrauwen et al. 2019). Species-specific exon loss is recognized as one of the important evolutionary mechanisms and was shown to have a predominantly regulatory function in humans (Wang et al. 2015).

“Obese” phenotype line - it was observed that mice selected for the ‘obese’ phenotype have substantially larger bones than the control (Müller-Eigner et al. 2022). I detected a complex variant involving heterozygous deletion and inversion overlapping a gene associated with modulation of bone formation (Liu and Mao 2004).

Increased endurance line - besides several SVs found spanning metabolism-related genes in mice selected for high endurance, I detected an inversion in a gene that when overexpressed, leads to a reduction of the effect of exercise-induced cardiac hypertrophy in mice (Sagara et al. 2012). Cardiac hypertrophy is seen as an adaptive physiological response to pressure or volume stress in cardiac tissue (Nakamura and Sadoshima 2018). This inversion encompasses the third and fourth exons and introns of a gene, putatively affecting splicing and expression patterns.

These results show that structural variants do indeed contribute to the evolution of the selected phenotypic traits. Through selective breeding of mouse lines, desired traits were enhanced within these populations, while maintaining and passing on structural variants specific to each line. Unfortunately, genetic material from the founders of these lines and offspring from the following generations is not available. This and the incomplete pedigree information hamper the detection of signatures of selection. However, the sequencing data analysed here still allows deriving biological interpretations based on the line-specific patterns of genetic differentiation.

The three concluding remarks of this section are:

1. Besides the genetic variation at the SNP level, a substantial portion of variation is also harbored at the structural level in artificially selected lines and should be investigated as part of the comprehensive assessment of genomic variation within and between populations.
2. Structural variation accumulates in artificially selected lines within genic regions, also encompassing genes related to phenotypic traits specific to each mouse line.
3. The SVs were twice as abundant in the selected lines compared to the control line, with the majority detected as polymorphic, and fewer in a fixed state. This suggests that despite the reproductive isolation and genetic drift, there is standing genetic variation in a form of SV present within the selected mouse lines.

Impact of assembly quality and discovery methods on variation detection

Comparative assessment of variation among genomes of closely related species strongly relies on the contiguity and completeness of assemblies and rigorous variant-calling analysis. As part of this dissertation, I generated the first reference genome assembly of tayra (*Eira barbara*), using linked-read technology paired with short-read sequencing to achieve a high level of genome completeness. Furthermore, due to the lack of a ‘best-practice’ SV calling pipeline, I set up and employed an SV detection and annotation pipeline with an ensemble of three SV calling methods based on evidence from the assembly (AS), read depth (RD), read pair (RP) and split read (SR) mapping, along with supporting bioinformatics tools for WGS data preparation. This approach enables SV detection with higher specificity and sensitivity compared to using a single detection algorithm (Kosugi et al. 2019; Moreno-Cabrera et al. 2019).

However, even when combining multiple tools, variant detection may to a certain extent be influenced by underlying biases of each of the SV discovery methods. To minimize this effect I used tools relying on different detection algorithms. Several other studies have used a similar approach, combining multiple algorithms to call SVs, followed by merging the outputs to increase the specificity and sensitivity (Mills et al. 2011; Lin et al. 2015; Sudmant et al. 2015).

In the IV chapter, an assessment of single nucleotide polymorphism diversity was conducted to estimate the effect of the contiguity and completeness of available genome assemblies on variant discovery. We investigated the distribution of genetic diversity along chromosomes of eight mammals (sea otter, cheetah, clouded leopard, giant otter, red panda, Asian small-clawed otter, American bison, and Eurasian river otter), including six species

listed in IUCN categories as vulnerable or endangered, for which both fragmented draft genome assemblies and recently generated chromosome-level assemblies were available.

Each of these species, except the Eurasian river otter (only chromosome-length genome analysed), was represented by two genome assemblies: the initial draft assembly and a chromosome-level assembly generated from that draft using Hi-C-scaffolding (Lieberman-Aiden et al. 2009). In general, assemblies are built from short and long adjacent DNA fragments referred to as contigs and scaffolds, respectively. With Hi-C mapping, the frequency (averaged over a cell population) at which two DNA fragments physically associate in 3D space can be measured, leading to the linking of the chromosomal structure directly to the genomic sequence.

We inspected scaffold N50, a metric widely used in the assessment and comparison of the contiguity among assemblies within the uniform size range. It represents the length of the shortest scaffold for which longer and equal length scaffolds cover at least 50 % of the assembly (Mäkinen, Salmela, and Ylinen 2012). With the application of Hi-C scaffolding, the N50 of the assemblies increased considerably. The most dramatic improvement was observed for the Asian small-clawed otter ($\times 1309$) and giant otter ($\times 784$), while the smallest was observed for the sea otter ($\times 3.8$).

Heterozygosity is usually low in threatened and endangered species due to smaller population size and a more prominent effect of genetic drift (Spielman, Brook, and Frankham 2004; McMahon, Teeling, and Höglund 2014; Jost et al. 2018). We observed the diminished heterozygosity in the genomes of some species, with the most extended regions of low heterozygosity/SNP density across the chromosomes of cheetah and sea otter.

The distributions of the heterozygous SNPs calculated in non-overlapping windows of 100 kbp and 1 Mbp between the draft and the chromosome-level assemblies were similar for all species. However, representing the distribution of the heterozygous SNPs of draft assemblies as density plots was challenging due to the high number of short scaffolds that are generally smaller than the window size of 1 Mb. The chromosome-level assemblies showed notable improvement in localization and visualization of genetic diversity. Among the eight studied species, giant otter and Asian small-clawed otter had the lowest scaffold N50 values - 0.17 and 0.1 Mbp, respectively, and were the most fragmented among the ones we considered. These two draft genome assemblies also had the lowest numbers of SNPs per 1 Mbp and even per 100 kbp windows.

We demonstrated that the more contiguous assemblies can be generated using Hi-C scaffolding of the existing short-read draft assemblies, where N50 of the draft contigs is as low as 0.1% of the estimated length of the genome. With the contiguity markedly improved in chromosome-length assemblies, a more complete overview of heterozygosity distribution across the genomes is enabled.

Similarly, a thorough assessment of structural variation distribution along genomes is dependent on the sequencing technology and assembly approach. The SV discovery from

short-read WGS data is determined by indirect inferences (e.g. read-depth and discordant read-pair mapping) (Chaisson et al. 2019). Analyses from the Human Genome Structural Variation Consortium (HGSVC) of three families captured ~11,000 SVs per genome from short-read WGS and ~25,000 SVs per genome from long-read WGS assembly. The 9.7% of the GRCh38 reference is defined by segmental duplication (SD) and simple repeat (SR), and 91.4% of deletions were specifically discovered by long-read WGS localized to these regions. Across the remaining 90.3% of the reference sequence, a high (93.8%) concordance was observed between technologies for deletions in these datasets. In contrast, long-read WGS performed better in the detection of insertions across all genomic contexts (Zhao et al. 2021). Additionally, 32 fully phased genome assemblies from diverse human populations were recently assembled using long-read technology. Following characterization of the structural variation, it was found that 68% of SVs detected were not discovered in short-read sequencing data (Ebert et al. 2021).

It is necessary to be aware of potential technological biases and shortcomings as well as genome assembly incompleteness during an SV project design. **Which metrics are important to consider for the structural variation discovery?**

Scaffold N50 is one of the metrics that should be taken into account in the initial screening of the datasets prior to SV analysis if the assembled genomes of highly similar size are available. In Chapter I, the scaffold N50 values of assemblies of three investigated species varied from 0.2 Mb for wolverine to 42 Mb for tayra genome assembly. Moreover, the total number of scaffolds ranged from 47,417 in wolverine to 14,579 in tayra assembly, implying a high fragmentation level of the wolverine genome, partly affecting the SV detectability (~87x fewer SVs detected compared to sable, and ~33x fewer SVs compared to tayra).

BUSCO (*Benchmarking Universally Single-Copy Orthologs*) metric is based on evolutionarily-informed expectations of the gene content of near-universal single-copy orthologs. The BUSCO metric is used to compare the status (complete, duplicated, fragmented, missing) of the gene content of a genome assembly, providing an insight into the level of assembly completeness. In our dataset, the BUSCO values for fragmented and missing orthologs ranged from 104 and 582 for tayra to 374 and 1143 for wolverine, respectively, providing evidence for a difference in assembly completeness.

Read length - comprehensive detection of SVs requires highly contiguous genome assemblies covering the repetitive fraction of genomes. This is particularly problematic when the read length is notably shorter than the repetitive element, in which case it is difficult to anchor the reads to unique genomic regions. The limited length of next-generation sequencing reads (≤ 300 bp) impedes the detection of SVs, especially insertions. The majority (~83%) of insertions are being missed by common short-read variant calling algorithms (Chaisson et al. 2019). These technical limitations can be amended with long-read sequencing reads (10–50 kbp) able to span over longer genomic segments (Sedlazeck et al. 2018; Chaisson et al. 2019).

Reference genome - conventionally, read alignment is performed using a single reference genome that usually comprises the genome sequence of one individual and it does not capture the genomic diversity of a population. This results in a reference bias that can have effects on downstream population genomic analysis where heterozygous sites can be falsely considered homozygous for the reference allele (Günther and Nettelblad 2019). To diminish the reference genome bias, we aligned the preprocessed reads from closely-related species to an equidistant reference genome, which is a common approach for non-model species (Weissensteiner et al. 2020; Prasad, Lorenzen, and Westbury 2022). Preferably, a contiguous reference assembly with a low number of gaps (<1000/ 100kbp) should be used (Peona et al. 2021).

Library and insert size - from our experience in Chapter I, sequencing a library with a large number of reads (> 1 billion, 150 PE) with one uniform insert size (e.g. 500 bp) provided more reliable SV detection, compared to multiple libraries with fewer reads per library and with different insert sizes (Ekblom et al. 2018). The latter leads to lower confidence in supporting evidence, e.g. fewer paired and split reads to accurately localize breakpoints of candidate SVs, and subsequently results in lower detection as observed in the case of the wolverine genome. To achieve uniformity among the samples, we compared paired-end reads of the same length and trimmed them where needed to retain the consistent read length (150 PE).

Multiple SV detection algorithms - as previously noted, a frequently used strategy to make SV detection more accurate is to run multiple algorithms simultaneously and use a subset of candidate SVs predicted by at least two or more algorithms (Gokcumen et al. 2013; Zichner et al. 2013; Weissensteiner et al. 2020). Although a number of existing SV detection algorithms can detect many types of SVs from the WGS data, no single computational algorithm can detect all types and all sizes of SVs with high sensitivity and high specificity (Lin et al. 2015). Until the cost and throughput of long-read sequencing can feasibly support large-scale comparative genomic studies, a triaged application of multiple methods for SV detection from WGS data should be considered to gain a ~3% increase in sensitivity over individual methods while decreasing FDR from 7% to 3% (current standard 5%) (Chaisson et al. 2019).

Multiple samples - besides the application of multiple callers, another approach to increase both the specificity and sensitivity of SV detection is to integrate multiple samples. SV detection in population-scale datasets allows the more reliable discovery of SVs and hence increases the power for linking functional variation to phenotype (Lin et al. 2015). We chose this approach in Chapter III where 150 samples originating from six mice lines were analyzed.

Genome coverage - this is a critical metric to consider prior to SV detection when using the read depth algorithm (Medvedev et al. 2010; Sims et al. 2014). In cases where samples vary greatly in their depth of coverage, downsampling to the lowest coverage within the sample set should be performed. However, the suggested coverage for the SV detection from short-read data is 15 - 30x (Sedlazeck et al. 2018), with lower detection power

observed below this threshold (Zhao et al. 2021). We took a cautious approach in Chapter III, where the coverage among samples varied from 5 - 30x, by splitting samples into low and high coverage sets and focusing our analysis on the latter as the principal source of candidate SVs.

Future outlook

As the impact of structural variants on the adaptive evolution of mammalian genomes becomes ever more apparent, I am confident that the research of structural genomic variation will steadily gain ever more relevance across the scientific domains, from biomedical fields to conservation biology. Already a number of studies demonstrate the dramatic impact SVs have in the development of certain human diseases and disorders, such as Duchenne muscular dystrophy (Barseghyan et al. 2017), Crohn's disease (McCarroll et al. 2008), and increased CNV in different types of cancer (Shlien and Malkin 2009; Shao et al. 2019). However, the role of structural variants in the adaptive evolution of humans (Radke and Lee 2015; Giner-Delgado et al. 2019) and other mammalian species (Maggiolini et al. 2020; Porubsky et al. 2020) is somewhat lagging behind.

Still, with the rising number of high-quality genomes and a variety of open-source detection tools being available, I expect the investigation of SVs in a wide range of nonmodel species to follow. Systematic integration of SV analysis into population studies will certainly help answer questions about the complex interaction of selection, drift and gene flow with SV distribution on intra- and interspecies levels. Additionally, the development of novel sequencing methods such as Strand-Seq (Sanders et al. 2017), enabled efficient haplotype phasing of entire chromosomes and reliable detection of different SV types, especially important for the study of the copy-neutral SVs (inversions and translocations). The cost and availability of new technologies will be crucial for their universal application across research facilities. Moreover, establishing the 'best-practice' computational tools for SV analysis, generalized for different sequencing data inputs, similar to ones available for SNP variant calling (e.g. GATK developed by Broad Institute), will be beneficial for reproducibility and uniformity of SV analyses across the studies.

For many wildlife species, linking genetic variation to phenotype, demonstrating that phenotype impacts fitness and showing allele frequency changes as a result of the selection will be next to impossible. However, it is indeed possible to assess the functional aspect of variation, using e.g. transcriptomics and proteomics approaches (Alvarez et al. 2015; Xie et al. 2020). Moreover, the predictions could be potentially tested using CRISPR tools to examine the expected effect on phenotype, as in a recent study of variation in coat colouration in mouse lines that linked fitness measurements in the wild to genes involved in fitness-related phenotypes and changes in their allele frequencies (Barret et al. 2019).

Besides improvements in technical aspects of structural variant characterization, I anticipate a substantial focus to be drawn on structural variants affecting non-coding and regulatory regions of the genomes, to better understand their roles in the evolution of adaptive and maladaptive phenotypes. Recently, it was found that the SVs also impact long-range chromatin structure, by disrupting or re-establishing chromatin contacts. The 3D chromatin structure is characterized by topologically associated domains (TADs) and chromatin loops that create physical interactions between genes and distant regulatory elements. Notably, large SVs, such as deletions, duplications and inversions, causing TAD disruption and/or fusion, are associated with rare developmental disorders (e.g. human limb malformations) and cancers (Franke et al. 2016; Huynh and Hormozdiari 2019; Sadowski et al. 2019).

Furthermore, there is a notable potential for applying SV assessment to assist with decision-making in plant and animal breeding programs, especially for the purpose of species survival (Mable 2019; Wold et al. 2021). Thus, it is fundamental to develop tools applicable across different studies and organisms and establish best practices in order to ensure that comparable insights can be obtained from joint analysis of genome sequences.

References

- Adalsteinsson, S. 1980. "Establishment of Equilibrium for the Dominant Lethal Gene for Manx Taillessness in Cats." *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik* 57 (4): 49–53.
- Aitman, Timothy J., Rong Dong, Timothy J. Vyse, Penny J. Norsworthy, Michelle D. Johnson, Jennifer Smith, Jonathan Mangion, et al. 2006. "Copy Number Polymorphism in Fcgr3 Predisposes to Glomerulonephritis in Rats and Humans." *Nature* 439 (7078): 851–55.
- Albalat, Ricard, and Cristian Cañestro. 2016. "Evolution by Gene Loss." *Nature Reviews. Genetics* 17 (7): 379–91.
- Alföldi, Jessica, and Kerstin Lindblad-Toh. 2013. "Comparative Genomics as a Tool to Understand Evolution and Disease." *Genome Research* 23 (7): 1063–68.
- Alkan, Can, Bradley P. Coe, and Evan E. Eichler. 2011. "Genome Structural Variation Discovery and Genotyping." *Nature Reviews. Genetics* 12 (5): 363–76.
- Alkan, Can, Jeffrey M. Kidd, Tomas Marques-Bonet, Gozde Aksay, Francesca Antonacci, Fereydoun Hormozdiari, Jacob O. Kitzman, et al. 2009. "Personalized Copy Number and Segmental Duplication Maps Using Next-Generation Sequencing." *Nature Genetics* 41 (10): 1061–67.
- Alm, Hannelore, Simone Kuhlmann, Martina Langhammer, Armin Tuchscherer, Helmut Torner, and Norbert Reinsch. 2010. "Occurrence of Polyovular Follicles in Mouse Lines Selected for High Fecundity." *The Journal of Reproduction and Development* 56 (4): 449–53.
- Alvarez, Mariano, Aaron W. Schrey, and Christina L. Richards. 2015. "Ten Years of Transcriptomics in Wild Populations: What Have We Learned about Their Ecology and Evolution?" *Molecular Ecology* 24 (4): 710–25.
- Alves, Joao M., Alexandra M. Lopes, Lounès Chikhi, and António Amorim. 2012. "On the Structural Plasticity of the Human Genome: Chromosomal Inversions Revisited." *Current Genomics* 13 (8): 623–32.
- Annis, Angela M., Jim Apostolopoulos, Sebastian Dworkin, Louise E. Purton, and Rosemary L. Sparrow. 2002. "An Oxysterol-Binding Protein Family Identified in the Mouse." *DNA and Cell Biology* 21 (8): 571–80.
- Amarasinghe, Shanika L., Shian Su, Xueyi Dong, Luke Zappia, Matthew E. Ritchie, and Quentin Gouil. 2020. "Opportunities and Challenges in Long-Read Sequencing Data Analysis." *Genome Biology* 21 (1): 30.
- Arck, Petra, Peter J. Hansen, Biserka Mulac Jericevic, Marie-Pierre Piccinni, and Julia Szekeres-Bartho. 2007. "Progesterone during Pregnancy: Endocrine-Immune Cross Talk in Mammalian Species and the Role of Stress." *American Journal of Reproductive Immunology* 58 (3): 268–79.
- Arendt, Maja, Tove Fall, Kerstin Lindblad-Toh, and Erik Axelsson. 2014. "Amylase Activity Is Associated with AMY2B Copy Numbers in Dog: Implications for Dog Domestication, Diet and Diabetes." *Animal Genetics* 45 (5): 716–22.
- Armstrong, Ellie E., Ryan W. Taylor, Stefan Prost, Peter Blinston, Esther van der Meer, Hillary Madzikanda, Olivia Mufute, et al. 2019. "Cost-Effective Assembly of the African Wild Dog (*Lycaon pictus*) Genome Using Linked Reads." *GigaScience* 8 (2). <https://doi.org/10.1093/gigascience/giy124>.
- Axelsson, Erik, Abhirami Ratnakumar, Maja-Louise Arendt, Khurram Maqbool, Matthew T. Webster, Michele Perloski, Olof Liberg, Jon M. Arnemo, Ake Hedhammar, and Kerstin Lindblad-Toh. 2013. "The Genomic Signature of Dog Domestication Reveals Adaptation to a Starch-Rich Diet." *Nature* 495 (7441): 360–64.
- Babcock, Melanie, Adam Pavlicek, Elizabeth Spiteri, Catherine D. Kashork, Ilya Ioshikhes, Lisa G. Shaffer, Jerzy Jurka, and Bernice E. Morrow. 2003. "Shuffling of Genes within Low-Copy Repeats on 22q11 (LCR22) by Alu-Mediated Recombination Events during Evolution." *Genome Research* 13 (12): 2519–32.
- Bailey, Jeffrey A., Ge Liu, and Evan E. Eichler. 2003. "An Alu Transposition Model for the Origin and Expansion of Human Segmental Duplications." *American Journal of Human Genetics* 73 (4): 823–34.
- Balachandran, Parithi, and Christine R. Beck. 2020. "Structural Variant Identification and Characterization." *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology* 28 (1): 31–47.

- Barrett, Rowan D. H., Stefan Laurent, Ricardo Mallarino, Susanne P. Pfeifer, Charles C. Y. Xu, Matthieu Foll, Kazumasa Wakamatsu, Jonathan S. Duke-Cohan, Jeffrey D. Jensen, and Hopi E. Hoekstra. 2019. "Linking a Mutation to Survival in Wild Mice." *Science* 363 (6426): 499–504.
- Barseghyan, Hayk, Wilson Tang, Richard T. Wang, Miguel Almalvez, Eva Segura, Matthew S. Bramble, Allen Lipson, et al. 2017. "Next-Generation Mapping: A Novel Approach for Detection of Pathogenic Structural Variants with a Potential Utility in Clinical Diagnosis." *Genome Medicine* 9 (1): 90.
- Bellone, Rebecca R., Heather Holl, Vijayasradhi Setaluri, Sulochana Devi, Nityanand Maddodi, Sheila Archer, Lynne Sandmeyer, et al. 2013. "Evidence for a Retroviral Insertion in TRPM1 as the Cause of Congenital Stationary Night Blindness and Leopard Complex Spotting in the Horse." *PloS One* 8 (10): e78280.
- Benazzo, Andrea, Emiliano Trucchi, James A. Cahill, Pierpaolo Maisano Delsler, Stefano Mona, Matteo Fumagalli, Lynsey Bunnefeld, et al. 2017. "Survival and Divergence in a Small Group: The Extraordinary Genomic History of the Endangered Apennine Brown Bear Stragglers." *Proceedings of the National Academy of Sciences of the United States of America* 114 (45): E9589–97.
- Bickhart, Derek M., and George E. Liu. 2014. "The Challenges and Importance of Structural Variation Detection in Livestock." *Frontiers in Genetics* 5 (February): 37.
- Boeke, J. D., D. J. Garfinkel, C. A. Styles, and G. R. Fink. 1985. "Ty Elements Transpose through an RNA Intermediate." *Cell* 40 (3): 491–500.
- Bornstein, Eran, Erez Lenchner, Alan Donnenfeld, Cristiano Jodicke, Sean M. Keeler, Sara Kapp, and Michael Y. Divon. 2010. "Complete Trisomy 21 vs Translocation Down Syndrome: A Comparison of Modes of Ascertainment." *American Journal of Obstetrics and Gynecology* 203 (4): 391.e1–5.
- Bourque, Guillaume, Kathleen H. Burns, Mary Gehring, Vera Gorbunova, Andrei Seluanov, Molly Hammell, Michaël Imbeault, et al. 2018. "Ten Things You Should Know about Transposable Elements." *Genome Biology* 19 (1): 199.
- Bristol-Gould, Sarah, and Teresa K. Woodruff. 2006. "Folliculogenesis in the Domestic Cat (*Felis Catus*)." *Theriogenology* 66 (1): 5–13.
- Burgoyne, Paul S., Shantha K. Mahadevaiah, and James M. A. Turner. 2009. "The Consequences of Asynapsis for Mammalian Meiosis." *Nature Reviews. Genetics* 10 (3): 207–16.
- Cao, Xiaolong, Yeting Zhang, Lindsay M. Payer, Hannah Lords, Jared P. Steranka, Kathleen H. Burns, and Jinchuan Xing. 2020. "Polymorphic Mobile Element Insertions Contribute to Gene Expression and Alternative Splicing in Human Tissues." *Genome Biology* 21 (1): 185.
- Capilla, Laia, Rosa Ana Sánchez-Guillén, Marta Farré, Andreu Paytuví-Gallart, Roberto Malinverni, Jacint Ventura, Denis M. Larkin, and Aurora Ruiz-Herrera. 2016. "Mammalian Comparative Genomics Reveals Genetic and Epigenetic Features Associated with Genome Reshuffling in Rodentia." *Genome Biology and Evolution* 8 (12): 3703–17.
- Catanach, Andrew, Ross Crowhurst, Cecilia Deng, Charles David, Louis Bernatchez, and Maren Wellenreuther. 2019. "The Genomic Pool of Standing Structural Variation Outnumbers Single Nucleotide Polymorphism by Threefold in the Marine Teleost *Chrysophrys Auratus*." *Molecular Ecology* 28 (6): 1210–23.
- Chain, Frédéric J. J., and Philine G. D. Feulner. 2014. "Ecological and Evolutionary Implications of Genomic Structural Variations." *Frontiers in Genetics* 5 (September): 326.
- Chaisson, Mark J. P., Ashley D. Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J. Gardner, et al. 2019. "Multi-Platform Discovery of Haplotype-Resolved Structural Variation in Human Genomes." *Nature Communications* 10 (1): 1784.
- Chakraborty, Mahul, J. J. Emerson, Stuart J. Macdonald, and Anthony D. Long. 2019. "Structural Variants Exhibit Widespread Allelic Heterogeneity and Shape Variation in Complex Traits." *Nature Communications* 10 (1): 4872.
- Chandana, Ediriweera P. S., Yasuhiro Maeda, Akihiko Ueda, Hiroshi Kiyonari, Naoko Oshima, Mako Yamamoto, Shunya Kondo, et al. 2010. "Involvement of the Reck Tumor Suppressor Protein in Maternal and Embryonic Vascular Remodeling in Mice." *BMC Developmental Biology*. <https://doi.org/10.1186/1471-213x-10-84>.
- Chen, Congying, Ruimin Qiao, Rongxing Wei, Yuanmei Guo, Huashui Ai, Junwu Ma, Jun Ren, and Lusheng Huang. 2012. "A Comprehensive Survey of Copy Number Variation in 18 Diverse Pig Populations and Identification of Candidate Copy Number Variable Genes Associated with Complex Traits." *BMC Genomics* 13 (December): 733.
- Chen, Pingping, Hexige Saiyin, Ruona Shi, Bin Liu, Xu Han, Yuping Gao, Xiantao Ye, Xiaofei Zhang, and Yu Sun. 2021. "Loss of SPACA1 Function Causes Autosomal Recessive Globozoospermia by Damaging the Acrosome-Acroplaxome Complex." *Human Reproduction* 36 (9): 2587–96.
- Chen, Xiaoyu, Ole Schulz-Trieglaff, Richard Shaw, Bret Barnes, Felix Schlesinger, Morten Källberg, Anthony

- J. Cox, Semyon Kruglyak, and Christopher T. Saunders. 2016. "Manta: Rapid Detection of Structural Variants and Indels for Germline and Cancer Sequencing Applications." *Bioinformatics* 32 (8): 1220–22.
- Chiang, Colby, Ryan M. Layer, Gregory G. Faust, Michael R. Lindberg, David B. Rose, Erik P. Garrison, Gabor T. Marth, Aaron R. Quinlan, and Ira M. Hall. 2015. "SpeedSeq: Ultra-Fast Personal Genome Analysis and Interpretation." *Nature Methods* 12 (10): 966–68.
- Chiang, Colby, Alexandra J. Scott, Joe R. Davis, Emily K. Tsang, Xin Li, Yungil Kim, Tarik Hadzic, et al. 2017. "The Impact of Structural Variation on Human Gene Expression." *Nature Genetics* 49 (5): 692–99.
- Conner, Jeffrey K. 2003. "Artificial Selection: A Powerful Tool for Ecologists." *Ecology* 84 (7): 1650–60.
- Conrad, Donald F., and Matthew E. Hurler. 2007. "The Population Genetics of Structural Variation." *Nature Genetics* 39 (7 Suppl): S30–36.
- Conrad, Donald F., Dalila Pinto, Richard Redon, Lars Feuk, Omer Gokcumen, Yujun Zhang, Jan Aerts, et al. 2010. "Origins and Functional Impact of Copy Number Variation in the Human Genome." *Nature* 464 (7289): 704–12.
- Cooper, Gregory M., Troy Zerr, Jeffrey M. Kidd, Evan E. Eichler, and Deborah A. Nickerson. 2008. "Systematic Assessment of Copy Number Variant Detection via Genome-Wide SNP Genotyping." *Nature Genetics* 40 (10): 1199–1203.
- Copes, Lynn E., Heidi Schutz, Elizabeth M. Dlugosz, Wendy Acosta, Mark A. Chappell, and Theodore Garland Jr. 2015. "Effects of Voluntary Exercise on Spontaneous Physical Activity and Food Consumption in Mice: Results from an Artificial Selection Experiment." *Physiology & Behavior* 149 (October): 86–94.
- Cui, Chenghua, Wei Shu, and Peining Li. 2016. "Fluorescence In Situ Hybridization: Cell-Based Genetic Diagnostic and Research Applications." *Frontiers in Cell and Developmental Biology* 4 (September): 89.
- Danchin, Etienne G. J., Philippe Gouret, and Pierre Pontarotti. 2006. "Eleven Ancestral Gene Families Lost in Mammals and Vertebrates While Otherwise Universally Conserved in Animals." *BMC Evolutionary Biology* 6 (January): 5.
- Deng, Wulan, Xinghua Shi, Robert Tjian, Timothée Lionnet, and Robert H. Singer. 2015. "CASFiSH: CRISPR/Cas9-Mediated in Situ Labeling of Genomic Loci in Fixed Cells." *Proceedings of the National Academy of Sciences of the United States of America* 112 (38): 11870–75.
- Dhar, Riddhiman, Tobias Bergmiller, and Andreas Wagner. 2014. "Increased Gene Dosage Plays a Predominant Role in the Initial Stages of Evolution of Duplicate TEM-1 Beta Lactamase Genes." *Evolution; International Journal of Organic Evolution* 68 (6): 1775–91.
- Dobigny, Gauthier, Janice Britton-Davidian, and Terence J. Robinson. 2017. "Chromosomal Polymorphism in Mammals: An Evolutionary Perspective." *Biological Reviews of the Cambridge Philosophical Society* 92 (1): 1–21.
- Dobzhansky, T., and A. H. Sturtevant. 1938. "Inversions in the Chromosomes of *Drosophila Pseudoobscura*." *Genetics* 23 (1): 28–64.
- Dudchenko, Olga, Sanjit S. Batra, Arina D. Omer, Sarah K. Nyquist, Marie Hoeger, Neva C. Durand, Muhammad S. Shamim, et al. 2017. "De Novo Assembly of the *Aedes Aegypti* Genome Using Hi-C Yields Chromosome-Length Scaffolds." *Science* 356 (6333): 92–95.
- Dudchenko, Olga, Muhammad S. Shamim, Sanjit S. Batra, Neva C. Durand, Nathaniel T. Musial, Ragib Mostofa, Melanie Pham, et al. 2018. "The Juicebox Assembly Tools Module Facilitates de Novo Assembly of Mammalian Genomes with Chromosome-Length Scaffolds for under \$1000." <https://doi.org/10.1101/254797>.
- Ebert, Peter, Peter A. Audano, Qihui Zhu, Bernardo Rodriguez-Martin, David Porubsky, Marc Jan Bonder, Arvis Sulovari, et al. 2021. "Haplotype-Resolved Diverse Human Genomes and Integrated Analysis of Structural Variation." *Science* 372 (6537). <https://doi.org/10.1126/science.abf7117>.
- Ehrenreich, Ian M., Noorossadat Torabi, Yue Jia, Jonathan Kent, Stephen Martis, Joshua A. Shapiro, David Gresham, Amy A. Caudy, and Leonid Kruglyak. 2010. "Dissection of Genetically Complex Traits with Extremely Large Pools of Yeast Segregants." *Nature* 464 (7291): 1039–42.
- Eichler, Evan E., Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M. Leal, Jason H. Moore, and Joseph H. Nadeau. 2010. "Missing Heritability and Strategies for Finding the Underlying Causes of Complex Disease." *Nature Reviews. Genetics* 11 (6): 446–50.
- Eisenbach, Michael, and Laura C. Giojalas. 2006. "Sperm Guidance in Mammals - an Unpaved Road to the Egg." *Nature Reviews. Molecular Cell Biology* 7 (4): 276–85.
- Ekblom, Robert, Birte Brechlin, Jens Persson, Linnéa Smeds, Malin Johansson, Jessica Magnusson, Øystein Flagstad, and Hans Ellegren. 2018. "Genome Sequencing and Conservation Genomics in the Scandinavian Wolverine Population." *Conservation Biology: The Journal of the Society for Conservation Biology* 32 (6): 1301–12.

- Ellis, T. H. Noel, Julie M. I. Hofer, Gail M. Timmerman-Vaughan, Clarice J. Coyne, and Roger P. Hellens. 2011. "Mendel, 150 Years on." *Trends in Plant Science* 16 (11): 590–96.
- Elshire, Robert J., Jeffrey C. Glaubitz, Qi Sun, Jesse A. Poland, Ken Kawamoto, Edward S. Buckler, and Sharon E. Mitchell. 2011. "A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species." *PLoS One* 6 (5): e19379.
- Eom, Gwang Hyeon, Kee-Beom Kim, Jin Hee Kim, Ji-Young Kim, Ju-Ryung Kim, Hae Jin Kee, Dong-Wook Kim, et al. 2011. "Histone Methyltransferase SETD3 Regulates Muscle Differentiation." *The Journal of Biological Chemistry* 286 (40): 34733–42.
- Escaramís, Geòrgia, Elisa Docampo, and Raquel Rabionet. 2015. "A Decade of Structural Variants: Description, History and Methods to Detect Structural Variation." *Briefings in Functional Genomics* 14 (5): 305–14.
- Espregueira Themudo, Gonçalo, Luís Q. Alves, André M. Machado, Mónica Lopes-Marques, Rute R. da Fonseca, Miguel Fonseca, Raquel Ruivo, and L. Filipe C. Castro. 2020. "Losing Genes: The Evolutionary Remodeling of Cetacea Skin." *Frontiers in Marine Science* 7. <https://doi.org/10.3389/fmars.2020.592375>.
- Etherington, Graham J., Darren Heavens, David Baker, Ashleigh Lister, Rose McNelly, Gonzalo Garcia, Bernardo Clavijo, Iain Macaulay, Wilfried Haerty, and Federica Di Palma. 2020. "Sequencing Smart: De Novo Sequencing and Assembly Approaches for a Non-Model Mammal." *GigaScience* 9 (5). <https://doi.org/10.1093/gigascience/giaa045>.
- Fan, Huizhong, Qi Wu, Fuwen Wei, Fengtang Yang, Bee Ling Ng, and Yibo Hu. 2019. "Chromosome-Level Genome Assembly for Giant Panda Provides Novel Insights into Carnivora Chromosome Evolution." *Genome Biology* 20 (1): 267.
- Fan, Xian, Travis E. Abbott, David Larson, and Ken Chen. 2014. "BreakDancer: Identification of Genomic Structural Variation from Paired-End Read Mapping." *Current Protocols in Bioinformatics / Editorial Board, Andreas D. Baxevanis ... [et Al.]* 45: 15.6.1–11.
- Fellermann, Klaus, Daniel E. Stange, Elke Schaeffeler, Hartmut Schmalzl, Jan Wehkamp, Charles L. Bevins, Walter Reinisch, et al. 2006. "A Chromosome 8 Gene-Cluster Polymorphism with Low Human Beta-Defensin 2 Gene Copy Number Predisposes to Crohn Disease of the Colon." *American Journal of Human Genetics* 79 (3): 439–48.
- Feuk, Lars, Jeffrey R. MacDonald, Terence Tang, Andrew R. Carson, Martin Li, Girish Rao, Razi Khaja, and Stephen W. Scherer. 2005. "Discovery of Human Inversion Polymorphisms by Comparative Analysis of Human and Chimpanzee DNA Sequence Assemblies." *PLoS Genetics* 1 (4): e56.
- Flegel, Caroline, Felix Vogel, Adrian Hofreuter, Benjamin S. P. Schreiner, Sandra Osthold, Sophie Veitinger, Christian Becker, et al. 2015. "Characterization of the Olfactory Receptors Expressed in Human Spermatozoa." *Frontiers in Molecular Biosciences* 2: 73.
- Foote, Andrew D., Nagarjun Vijay, María C. Ávila-Arcos, Robin W. Baird, John W. Durban, Matteo Fumagalli, Richard A. Gibbs, et al. 2016. "Genome-Culture Coevolution Promotes Rapid Divergence of Killer Whale Ecotypes." *Nature Communications* 7 (May): 11693.
- Franke, Martin, Daniel M. Ibrahim, Guillaume Andrey, Wibke Schwarzer, Verena Heinrich, Robert Schöpflin, Katerina Kraft, et al. 2016. "Formation of New Chromatin Domains Determines Pathogenicity of Genomic Duplications." *Nature* 538 (7624): 265–69.
- Frazer, Kelly A., Sarah S. Murray, Nicholas J. Schork, and Eric J. Topol. 2009. "Human Genetic Variation and Its Contribution to Complex Traits." *Nature Reviews. Genetics* 10 (4): 241–51.
- Fujihara, Yoshitaka, Yuhkoh Satouh, Naokazu Inoue, Ayako Isotani, Masahito Ikawa, and Masaru Okabe. 2012. "SPACA1-Deficient Male Mice Are Infertile with Abnormally Shaped Sperm Heads Reminiscent of Globozoospermia." *Development* 139 (19): 3583–89.
- Gao, Yahui, Jianping Jiang, Shaohua Yang, Yali Hou, George E. Liu, Shengli Zhang, Qin Zhang, and Dongxiao Sun. 2017. "CNV Discovery for Milk Composition Traits in Dairy Cattle Using Whole Genome Resequencing." *BMC Genomics* 18 (1): 265.
- Gardner, Eugene J., Vincent K. Lam, Daniel N. Harris, Nelson T. Chuang, Emma C. Scott, W. Stephen Pittard, Ryan E. Mills, 1000 Genomes Project Consortium, and Scott E. Devine. 2017. "The Mobile Element Locator Tool (MELT): Population-Scale Mobile Element Discovery and Biology." *Genome Research* 27 (11): 1916–29.
- Gemmell, Neil J., and Jon Slate. 2006. "Heterozygote Advantage for Fecundity." *PLoS One* 1 (December): e125.
- Ge, Steven Xijin, Dongmin Jung, and Runan Yao. 2020. "ShinyGO: A Graphical Gene-Set Enrichment Tool for Animals and Plants." *Bioinformatics* 36 (8): 2628–29.
- Giani, Alice Maria, Guido Roberto Gallo, Luca Gianfranceschi, and Giulio Formenti. 2020. "Long Walk to Genomics: History and Current Approaches to Genome Sequencing and Assembly." *Computational and*

- Giner-Delgado, Carla, Sergi Villatoro, Jon Lerga-Jaso, Magdalena Gayà-Vidal, Meritxell Oliva, David Castellano, Lorena Pantano, et al. 2019. “Evolutionary and Functional Impact of Common Polymorphic Inversions in the Human Genome.” *Nature Communications* 10 (1): 4222.
- Goddard, Michael E., and Ben J. Hayes. 2009. “Mapping Genes for Complex Traits in Domestic Animals and Their Use in Breeding Programmes.” *Nature Reviews. Genetics* 10 (6): 381–91.
- Goidts, Violaine, David N. Cooper, Lluís Armengol, Werner Schempp, Jeffrey Conroy, Xavier Estivill, Norma Nowak, Horst Hameister, and Hildegard Kehrer-Sawatzki. 2006. “Complex Patterns of Copy Number Variation at Sites of Segmental Duplications: An Important Category of Structural Variation in the Human Genome.” *Human Genetics* 120 (2): 270–84.
- Gokcumen, Omer, Verena Tischler, Jelena Tica, Qihui Zhu, Rebecca C. Iskow, Eunjung Lee, Markus Hsi-Yang Fritz, et al. 2013. “Primate Genome Architecture Influences Structural Variation Mechanisms and Functional Consequences.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (39): 15764–69.
- Gomez-Raya, Luis, Hanne Gro Olsen, Frode Lingaas, Helge Klungland, Dag Inge Våge, Ingrid Olsaker, Seblewengel Bekele Talle, Monica Aasland, and Sigbjørn Lien. 2002. “The Use of Genetic Markers to Measure Genomic Response to Selection in Livestock.” *Genetics* 162 (3): 1381–88.
- Guijarro-Clarke, Cristina, Peter W. H. Holland, and Jordi Paps. 2020. “Widespread Patterns of Gene Loss in the Evolution of the Animal Kingdom.” *Nature Ecology & Evolution* 4 (4): 519–23.
- Günther, Torsten, and Carl Nettelblad. 2019. “The Presence and Impact of Reference Bias on Population Genomic Studies of Prehistoric Human Populations.” *PLoS Genetics* 15 (7): e1008302.
- Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. 2013. “QUAST: Quality Assessment Tool for Genome Assemblies.” *Bioinformatics* 29 (8): 1072–75.
- Gu, Wenli, Feng Zhang, and James R. Lupski. 2008. “Mechanisms for Human Genomic Rearrangements.” *PathoGenetics* 1 (1): 4.
- Hall, Ira M., and Aaron R. Quinlan. 2012. “Detection and Interpretation of Genomic Structural Variation in Mammals.” In *Genomic Structural Variants: Methods and Protocols*, edited by Lars Feuk, 225–48. New York, NY: Springer New York.
- Hämälä, Tuomas, Eric K. Wafula, Mark J. Guiltinan, Paula E. Ralph, Claude W. dePamphilis, and Peter Tiffin. 2021. “Genomic Structural Variants Constrain and Facilitate Adaptation in Natural Populations of *Theobroma Cacao*, the Chocolate Tree.” *Proceedings of the National Academy of Sciences of the United States of America* 118 (35). <https://doi.org/10.1073/pnas.2102914118>.
- Ha, Misook, Eun-Deok Kim, and Z. Jeffrey Chen. 2009. “Duplicate Genes Increase Expression Diversity in Closely Related Species and Allopolyploids.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (7): 2295–2300.
- Hammer, M. F., J. Schimenti, and L. M. Silver. 1989. “Evolution of Mouse Chromosome 17 and the Origin of Inversions Associated with T Haplotypes.” *Proceedings of the National Academy of Sciences of the United States of America* 86 (9): 3261–65.
- Han, Kyudong, Jungnam Lee, Thomas J. Meyer, Paul Remedios, Lindsey Goodwin, and Mark A. Batzer. 2008. “L1 Recombination-Associated Deletions Generate Human Genomic Variation.” *Proceedings of the National Academy of Sciences of the United States of America* 105 (49): 19366–71.
- Han, Mira V., Jeffery P. Demuth, Casey L. McGrath, Claudio Casola, and Matthew W. Hahn. 2009. “Adaptive Evolution of Young Gene Duplicates in Mammals.” *Genome Research* 19 (5): 859–67.
- Han, Seungnam, Jungyong Nam, Yan Li, Seho Kim, Suk-Hee Cho, Yi Sul Cho, So-Yeon Choi, et al. 2010. “Regulation of Dendritic Spines, Spatial Memory, and Embryonic Development by the TANC Family of PSD-95-Interacting Proteins.” *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 30 (45): 15102–12.
- Hedrick, Philip W. 2012. “What Is the Evidence for Heterozygote Advantage Selection?” *Trends in Ecology & Evolution* 27 (12): 698–704.
- Hedrick, Philip W. 2015. “Heterozygote Advantage: The Effect of Artificial Selection in Livestock and Pets.” *The Journal of Heredity* 106 (2): 141–54.
- Heldstab, Sandra A., Dennis W. H. Müller, Sereina M. Graber, Laurie Bingaman Lackey, Eberhard Rensch, Jean-Michel Hatt, Philipp Zerbe, and Marcus Clauss. 2018. “Geographical Origin, Delayed Implantation, and Induced Ovulation Explain Reproductive Seasonality in the Carnivora.” *Journal of Biological Rhythms* 33 (4): 402–19.
- Hoekstra, Hopi E., Rachel J. Hirschmann, Richard A. Bunde, Paul A. Insel, and Janet P. Crossland. 2006. “A Single Amino Acid Mutation Contributes to Adaptive Beach Mouse Color Pattern.” *Science* 313 (5783): 101–4.
- Hoffmann, Ary A., and Loren H. Rieseberg. 2008. “Revisiting the Impact of Inversions in Evolution: From

- Population Genetic Markers to Drivers of Adaptive Shifts and Speciation?" *Annual Review of Ecology, Evolution, and Systematics* 39 (December): 21–42.
- Ho, Steve S., Alexander E. Urban, and Ryan E. Mills. 2020. "Structural Variation in the Sequencing Era." *Nature Reviews. Genetics* 21 (3): 171–89.
- Hughes, Graham M., Emma M. Boston, John A. Finarelli, William J. Murphy, Desmond G. Higgins, and Emma C. Teeling. 2018. "The Birth and Death of Olfactory Receptor Gene Families in Mammalian Niche Adaptation." *Molecular Biology and Evolution*, March. <https://doi.org/10.1093/molbev/msy028>.
- Hu, Linping, Kun Ru, Li Zhang, Yuting Huang, Xiaofan Zhu, Hanzhi Liu, Anders Zetterberg, Tao Cheng, and Weimin Miao. 2014. "Fluorescence in Situ Hybridization (FISH): An Increasingly Demanded Tool for Biomarker Research and Personalized Medicine." *Biomarker Research* 2 (1): 3.
- Huynh, Linh, and Fereydoun Hormozdiari. 2019. "TAD Fusion Score: Discovery and Ranking the Contribution of Deletions to Genome Structure." *Genome Biology* 20 (1): 60.
- Iafraite, A. John, Lars Feuk, Miguel N. Rivera, Marc L. Listewnik, Patricia K. Donahoe, Ying Qi, Stephen W. Scherer, and Charles Lee. 2004. "Detection of Large-Scale Variation in the Human Genome." *Nature Genetics* 36 (9): 949–51.
- Ibarra-Soria, Ximena, Maria O. Levitin, and Darren W. Logan. 2014. "The Genomic Basis of Vomeronasal-Mediated Behaviour." *Mammalian Genome: Official Journal of the International Mammalian Genome Society* 25 (1-2): 75–86.
- Inman, Robert M., Audrey J. Magoun, Jens Persson, and Jenny Mattisson. 2012. "The Wolverine's Niche: Linking Reproductive Chronology, Caching, Competition, and Climate." *Journal of Mammalogy* 93 (3): 634–44.
- Jain, Miten, Sergey Koren, Karen H. Miga, Josh Quick, Arthur C. Rand, Thomas A. Sasani, John R. Tyson, et al. 2018. "Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads." *Nature Biotechnology* 36 (4): 338–45.
- Jain, Miten, Hugh E. Olsen, Benedict Paten, and Mark Akeson. 2016. "The Oxford Nanopore MinION: Delivery of Nanopore Sequencing to the Genomics Community." *Genome Biology* 17 (1): 239.
- Jeffares, Daniel C., Clemency Jolly, Mimoza Hoti, Doug Speed, Liam Shaw, Charalampos Rallis, Francois Balloux, Christophe Dessimoz, Jürg Bähler, and Fritz J. Sedlazeck. 2017. "Transient Structural Variations Have Strong Effects on Quantitative Traits and Reproductive Isolation in Fission Yeast." *Nature Communications* 8 (January): 14061.
- Jiao, S., C. Maltecca, K. A. Gray, and J. P. Cassady. 2014. "Feed Intake, Average Daily Gain, Feed Efficiency, and Real-Time Ultrasound Traits in Duroc Pigs: II. Genomewide Association." *Journal of Animal Science* 92 (7): 2846–60.
- Jones, Felicity C., Manfred G. Grabherr, Yingguang Frank Chan, Pamela Russell, Evan Mauceli, Jeremy Johnson, Ross Swofford, et al. 2012. "The Genomic Basis of Adaptive Evolution in Threespine Sticklebacks." *Nature* 484 (7392): 55–61.
- Jost, Lou, Frederick Archer, Sarah Flanagan, Oscar Gaggiotti, Sean Hoban, and Emily Latch. 2018. "Differentiation Measures for Conservation Genetics." *Evolutionary Applications* 11 (7): 1139–48.
- Kadri, Naveen Kumar, Goutam Sahana, Carole Charlier, Terhi Iso-Touru, Bernt Guldbbrandtsen, Latifa Karim, Ulrik Sander Nielsen, et al. 2014. "A 660-Kb Deletion with Antagonistic Effects on Fertility and Milk Production Segregates at High Frequency in Nordic Red Cattle: Additional Evidence for the Common Occurrence of Balancing Selection in Livestock." *PLoS Genetics* 10 (1): e1004049.
- Karaođlanođlu, Fatih, Camir Ricketts, Ezgi E布伦, Marzieh Eslami Rasekh, Iman Hajirasouliha, and Can Alkan. 2020. "VALOR2: Characterization of Large-Scale Structural Variants Using Linked-Reads." *Genome Biology* 21 (1): 72.
- Karlsson, Elinor K., Izabella Baranowska, Claire M. Wade, Nicolette H. C. Salmon Hillbertz, Michael C. Zody, Nathan Anderson, Tara M. Biagi, et al. 2007. "Efficient Mapping of Mendelian Traits in Dogs through Genome-Wide Association." *Nature Genetics* 39 (11): 1321–28.
- Kazazian, H. H., Jr, and J. V. Moran. 1998. "The Impact of L1 Retrotransposons on the Human Genome." *Nature Genetics* 19 (1): 19–24.
- Kessner, Darren, and John Novembre. 2015. "Power Analysis of Artificial Selection Experiments Using Efficient Whole Genome Simulation of Quantitative Traits." *Genetics* 199 (4): 991–1005.
- Kidd, Jeffrey M., Gregory M. Cooper, William F. Donahue, Hillary S. Hayden, Nick Sampas, Tina Graves, Nancy Hansen, et al. 2008. "Mapping and Sequencing of Structural Variation from Eight Human Genomes." *Nature* 453 (7191): 56–64.
- Kidd, Jeffrey M., Tina Graves, Tera L. Newman, Robert Fulton, Hillary S. Hayden, Maika Malig, Joelle Kallicki, Rajinder Kaul, Richard K. Wilson, and Evan E. Eichler. 2010. "A Human Genome Structural Variation Sequencing Resource Reveals Insights into Mutational Mechanisms." *Cell* 143 (5): 837–47.
- Kim, Philip M., Hugo Y. K. Lam, Alexander E. Urban, Jan O. Korb, Jason Affourtit, Fabian Grubert,

- Xueying Chen, Sherman Weissman, Michael Snyder, and Mark B. Gerstein. 2008. "Analysis of Copy Number Variants and Segmental Duplications in the Human Genome: Evidence for a Change in the Process of Formation in Recent Evolutionary History." *Genome Research* 18 (12): 1865–74.
- Kim, Sobin, and Ashish Misra. 2007. "SNP Genotyping: Technologies and Biomedical Applications." *Annual Review of Biomedical Engineering* 9: 289–320.
- Koepfli, Klaus-Peter, Kerry A. Deere, Graham J. Slater, Colleen Begg, Keith Begg, Lon Grassman, Mauro Lucherini, Geraldine Veron, and Robert K. Wayne. 2008. "Multigene Phylogeny of the Mustelidae: Resolving Relationships, Tempo and Biogeographic History of a Mammalian Adaptive Radiation." *BMC Biology* 6 (February): 10.
- Korbel, Jan O., Alexander Eckehart Urban, Jason P. Affourtit, Brian Godwin, Fabian Grubert, Jan Fredrik Simons, Philip M. Kim, et al. 2007. "Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome." *Science* 318 (5849): 420–26.
- Kosugi, Shunichi, Yukihide Momozawa, Xiaoxi Liu, Chikashi Terao, Michiaki Kubo, and Yoichiro Kamatani. 2019. "Comprehensive Evaluation of Structural Variation Detection Algorithms for Whole Genome Sequencing." *Genome Biology* 20 (1): 117.
- Kriventseva, Evgenia V., Dmitry Kuznetsov, Fredrik Tegenfeldt, Mosè Manni, Renata Dias, Felipe A. Simão, and Evgeny M. Zdobnov. 2019. "OrthoDB v10: Sampling the Diversity of Animal, Plant, Fungal, Protist, Bacterial and Viral Genomes for Evolutionary and Functional Annotations of Orthologs." *Nucleic Acids Research* 47 (D1): D807–11.
- Kronenberg, Zev N., Edward J. Osborne, Kelsey R. Cone, Brett J. Kennedy, Eric T. Domyan, Michael D. Shapiro, Nels C. Elde, and Mark Yandell. 2015. "Wham: Identifying Structural Variants of Biological Consequence." *PLoS Computational Biology* 11 (12): e1004572.
- Langhammer, Martina, Marten Michaelis, Andreas Hoeflich, Alexander Sobczak, Jennifer Schoen, and Joachim M. Weitzel. 2014. "High-Fertility Phenotypes: Two Outbred Mouse Models Exhibit Substantially Different Molecular and Physiological Strategies Warranting Improved Fertility." *Reproduction* 147 (4): 427–33.
- Langmead, Ben, and Steven L. Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59.
- Larson, Greger, Dolores R. Piperno, Robin G. Allaby, Michael D. Purugganan, Leif Andersson, Manuel Arroyo-Kalin, Loukas Barton, et al. 2014. "Current Perspectives and the Future of Domestication Studies." *Proceedings of the National Academy of Sciences of the United States of America* 111 (17): 6139–46.
- Law, Chris J., Graham J. Slater, and Rita S. Mehta. 2018. "Lineage Diversity and Size Disparity in Musteloidea: Testing Patterns of Adaptive Radiation Using Molecular and Fossil-Based Methods." *Systematic Biology* 67 (1): 127–44.
- Layer, Ryan M., Colby Chiang, Aaron R. Quinlan, and Ira M. Hall. 2014. "LUMPY: A Probabilistic Framework for Structural Variant Discovery." *Genome Biology* 15 (6): R84.
- Lee, Jennifer A., Claudia M. B. Carvalho, and James R. Lupski. 2007. "A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders." *Cell* 131 (7): 1235–47.
- Lee, Jungnam, Kyudong Han, Thomas J. Meyer, Heui-Soo Kim, and Mark A. Batzer. 2008. "Chromosomal Inversions between Human and Chimpanzee Lineages Caused by Retrotransposons." *PLoS One* 3 (12): e4047.
- Leffler, Ellen M., Ziyue Gao, Susanne Pfeifer, Laure Ségurel, Adam Auton, Oliver Venn, Rory Bowden, et al. 2013. "Multiple Instances of Ancient Balancing Selection Shared between Humans and Chimpanzees." *Science* 339 (6127): 1578–82.
- Lei, Wei, Xu-Hui Feng, Wen-Bo Deng, Hua Ni, Zhi-Rong Zhang, Bo Jia, Xin-Ling Yang, et al. 2012. "Progesterone and DNA Damage Encourage Uterine Cell Proliferation and Decidualization through up-Regulating Ribonucleotide Reductase 2 Expression during Early Pregnancy in Mice." *The Journal of Biological Chemistry* 287 (19): 15174–92.
- Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." *Science* 326 (5950): 289–93.
- Lin, Ke, Sandra Smit, Guusje Bonnema, Gabino Sanchez-Perez, and Dick de Ridder. 2015. "Making the Difference: Integrating Structural Variation Detection Tools." *Briefings in Bioinformatics* 16 (5): 852–64.
- Liu, Bing, and Ning Mao. 2004. "Smad5: Signaling Roles in Hematopoiesis and Osteogenesis." *The International Journal of Biochemistry & Cell Biology* 36 (5): 766–70.
- Liu, George E., Yali Hou, Bin Zhu, Maria Francesca Cardone, Lu Jiang, Angelo Cellamare, Apratim Mitra, et al. 2010. "Analysis of Copy Number Variations among Diverse Cattle Breeds." *Genome Research* 20 (5): 693–703.

- Liu, Mei, Yang Zhou, Benjamin D. Rosen, Curtis P. Van Tassell, Alessandra Stella, Gwenola Tosser-Klopp, Rachel Rupp, et al. 2019. "Diversity of Copy Number Variation in the Worldwide Goat Population." *Heredity* 122 (5): 636–46.
- Logan, Darren W., Lisa J. Brunet, William R. Webb, Tyler Cutforth, John Ngai, and Lisa Stowers. 2012. "Learned Recognition of Maternal Signature Odors Mediates the First Suckling Episode in Mice." *Current Biology: CB* 22 (21): 1998–2007.
- Lukacs, Paul M., Diane Evans Mack, Robert Inman, Justin A. Gude, Jacob S. Ivan, Robert P. Lanka, Jeffrey C. Lewis, et al. 2020. "Wolverine Occupancy, Spatial Distribution, and Monitoring Design." *The Journal of Wildlife Management* 14 (March): 17.
- Lye, Zoe N., and Michael D. Purugganan. 2019. "Copy Number Variation in Domestication." *Trends in Plant Science* 24 (4): 352–65.
- Lynch, M., and A. Force. 2000. "The Probability of Duplicate Gene Preservation by Subfunctionalization." *Genetics* 154 (1): 459–73.
- Lyon, Mary F. 2003. "Transmission Ratio Distortion in Mice." *Annual Review of Genetics* 37: 393–408.
- Mable, Barbara K. 2019. "Conservation of Adaptive Potential and Functional Diversity: Integrating Old and New Approaches." *Conservation Genetics* 20 (1): 89–100.
- Maggiolini, Flavia Angela Maria, Ashley D. Sanders, Colin James Shew, Arvis Sulovari, Yafei Mao, Marta Puig, Claudia Rita Catacchio, et al. 2020. "Single-Cell Strand Sequencing of a Macaque Genome Reveals Multiple Nested Inversions and Breakpoint Reuse during Primate Evolution." *Genome Research* 30 (11): 1680–93.
- Maglott, Donna, Jim Ostell, Kim D. Pruitt, and Tatiana Tatusova. 2011. "Entrez Gene: Gene-Centered Information at NCBI." *Nucleic Acids Research* 39 (Database issue): D52–57.
- Mahmoud, Medhat, Nastassia Gobet, Diana Ivette Cruz-Dávalos, Ninon Mounier, Christophe Dessimoz, and Fritz J. Sedlazeck. 2019. "Structural Variant Calling: The Long and the Short of It." *Genome Biology* 20 (1): 246.
- Mäkinen, Veli, Leena Salmela, and Johannes Ylinen. 2012. "Normalized N50 Assembly Metric Using Gap-Restricted Co-Linear Chaining." *BMC Bioinformatics* 13 (October): 255.
- Marks, Patrick, Sarah Garcia, Alvaro Martinez Barrio, Kamila Belhocine, Jorge Bernate, Rajiv Bharadwaj, Keith Bjornson, et al. 2019. "Resolving the Full Spectrum of Human Genome Variation Using Linked-Reads." *Genome Research* 29 (4): 635–45.
- Martinez, Bridget, and Rudy M. Ortiz. 2017. "Thyroid Hormone Regulation and Insulin Resistance: Insights From Animals Naturally Adapted to Fasting." *Physiology*, February. <https://doi.org/10.1152/physiol.00018.2016>.
- McCarroll, Steven A., Alan Huett, Petric Kuballa, Shannon D. Chilewski, Aimee Landry, Philippe Goyette, Michael C. Zody, et al. 2008. "Deletion Polymorphism Upstream of IRGM Associated with Altered IRGM Expression and Crohn's Disease." *Nature Genetics* 40 (9): 1107–12.
- McClintock, Barbara. 1931. "Cytological Observations of Deficiencies Involving Known Genes, Translocations and an Inversion in Zea Mays." *University of Missouri Agricultural Experiment Station Research Bulletin* 163: 3–30.
- McClintock, Barbara. 1950. "The Origin and Behavior of Mutable Loci in Maize." *Proceedings of the National Academy of Sciences of the United States of America* 36 (6): 344–55.
- McLaren, William, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. 2016. "The Ensembl Variant Effect Predictor." *Genome Biology* 17 (1): 122.
- McMahon, Barry J., Emma C. Teeling, and Jacob Höglund. 2014. "How and Why Should We Implement Genomics into Conservation?" *Evolutionary Applications* 7 (9): 999–1007.
- McNatty, Kenneth P., Derek A. Heath, Zaramasina Clark, Karen Reader, Jennifer L. Juengel, and Janet L. Pitman. 2017. "Ovarian Characteristics in Sheep with Multiple Fecundity Genes." *Reproduction* 153 (2): 233–40.
- Medvedev, Paul, Marc Fiume, Misko Dzamba, Tim Smith, and Michael Brudno. 2010. "Detecting Copy Number Variation with Mated Short Reads." *Genome Research* 20 (11): 1613–22.
- Meltz Steinberg, Karyn, Valerie A. Schneider, Can Alkan, Michael J. Montague, Wesley C. Warren, Deanna M. Church, and Richard K. Wilson. 2017. "Building and Improving Reference Genome Assemblies." *Proceedings of the IEEE* 105 (3): 422–35.
- Mendes, Tito, Liliana Silva, Daniela Almeida, and Agostinho Antunes. 2020. "Neofunctionalization of the UCP1 Mediated the Non-Shivering Thermogenesis in the Evolution of Small-Sized Placental Mammals." *Genomics* 112 (3): 2489–98.
- Mérot, Claire, Rebekah A. Oomen, Anna Tigano, and Maren Wellenreuther. 2020. "A Roadmap for Understanding the Evolutionary Significance of Structural Genomic Variation." *Trends in Ecology &*

- Evolution*, April. <https://doi.org/10.1016/j.tree.2020.03.002>.
- Mills, Ryan E., Klaudia Walter, Chip Stewart, Robert E. Handsaker, Ken Chen, Can Alkan, Alexej Abyzov, et al. 2011. "Mapping Copy Number Variation by Population-Scale Genome Sequencing." *Nature* 470 (7332): 59–65.
- Moran, J. V., R. J. DeBerardinis, and H. H. Kazazian Jr. 1999. "Exon Shuffling by L1 Retrotransposition." *Science* 283 (5407): 1530–34.
- Moreno-Cabrera, José Marcos, Jesús del Valle, Elisabeth Castellanos, Lidia Feliubadaló, Marta Pineda, Joan Brunet, Eduard Serra, Gabriel Capellà, Conxi Lázaro, and Bernat Gel. 2019. "Benchmark of Tools for CNV Detection from NGS Panel Data in a Genetic Diagnostics Context." *bioRxiv*. <https://doi.org/10.1101/850958>.
- Müller-Eigner, Annika, Adrián Sanz-Moreno, Irene de-Diego, Anuroop Venkateswaran Venkatasubramani, Martina Langhammer, Raffaele Gerlini, Birgit Rathkolb, et al. 2022. "Dietary Intervention Improves Health Metrics and Life Expectancy of the Genetically Obese Titan Mouse." *Communications Biology* 5 (1): 408.
- Murigneux, Valentine, Subash Kumar Rai, Agnelo Furtado, Timothy J. C. Bruxner, Wei Tian, Ivon Harliwong, Hanmin Wei, et al. 2020. "Comparison of Long-Read Methods for Sequencing and Assembly of a Plant Genome." *GigaScience* 9 (12). <https://doi.org/10.1093/gigascience/giaa146>.
- Nachman, Michael W., Hopi E. Hoekstra, and Susan L. D'Agostino. 2003. "The Genetic Basis of Adaptive Melanism in Pocket Mice." *Proceedings of the National Academy of Sciences of the United States of America* 100 (9): 5268–73.
- Nakamura, Michinari, and Junichi Sadoshima. 2018. "Mechanisms of Physiological and Pathological Cardiac Hypertrophy." *Nature Reviews. Cardiology* 15 (7): 387–407.
- Norris, Belinda J., and Vicki A. Whan. 2008. "A Gene Duplication Affecting Expression of the Ovine ASIP Gene Is Responsible for White and Black Sheep." *Genome Research* 18 (8): 1282–93.
- Ohno, Susumu. 1970. *Evolution by Gene Duplication*. Springer, Berlin, Heidelberg.
- Ott, Alina, James C. Schnable, Cheng-Ting Yeh, Linjiang Wu, Chao Liu, Heng-Cheng Hu, Clifton L. Dalgard, Soumik Sarkar, and Patrick S. Schnable. 2018. "Linked Read Technology for Assembling Large Complex and Polyploid Genomes." *BMC Genomics* 19 (1): 1–15.
- Pang, Andy W., Jeffrey R. MacDonald, Dalila Pinto, John Wei, Muhammad A. Rafiq, Donald F. Conrad, Hansoo Park, et al. 2010. "Towards a Comprehensive Structural Variation Map of an Individual Human Genome." *Genome Biology* 11 (5): R52.
- Paudel, Yogesh, Ole Madsen, Hendrik-Jan Megens, Laurent A. F. Frantz, Mirte Bosse, John W. M. Bastiaansen, Richard P. M. A. Crooijmans, and Martien A. M. Groenen. 2013. "Evolutionary Dynamics of Copy Number Variation in Pig Genomes in the Context of Adaptation and Domestication." *BMC Genomics* 14 (July): 449.
- Payan-Carreira, R., and M. A. Pires. 2008. "Multiocyte Follicles in Domestic Dogs: A Survey of Frequency of Occurrence." *Theriogenology* 69 (8): 977–82.
- Payer, Lindsay M., Jared P. Steranka, Wan Rou Yang, Maria Kryatova, Sibyl Medabalimi, Daniel Ardeljan, Chunhong Liu, Jef D. Boeke, Dimitri Avramopoulos, and Kathleen H. Burns. 2017. "Structural Variants Caused by Alu Insertions Are Associated with Risks for Many Human Diseases." *Proceedings of the National Academy of Sciences of the United States of America* 114 (20): E3984–92.
- Peona, Valentina, Mozes P. K. Blom, Luohao Xu, Reto Burri, Shawn Sullivan, Ignas Bunikis, Ivan Liachko, et al. 2021. "Identifying the Causes and Consequences of Assembly Gaps Using a Multiplatform Genome Assembly of a Bird-of-Paradise." *Molecular Ecology Resources* 21 (1): 263–86.
- Perry, George H., Amir Ben-Dor, Anya Tsalenko, Nick Sampas, Laia Rodriguez-Revenga, Charles W. Tran, Alicia Scheffer, et al. 2008. "The Fine-Scale and Complex Architecture of Human Copy-Number Variation." *American Journal of Human Genetics* 82 (3): 685–95.
- Poglayen-Neuwall, Ivo, Barbara S. Durrant, Marda L. Swansen, Robert C. Williams, and Roy A. Barnes. 1989. "Estrous Cycle of the tayra, Eira Barbara." *Zoo Biology* 8 (2): 171–77.
- Porubsky, David, Ashley D. Sanders, Wolfram Höps, Pingsun Hsieh, Arvis Sulovari, Ruiyang Li, Ludovica Mercuri, et al. 2020. "Recurrent Inversion Toggling and Great Ape Genome Evolution." *Nature Genetics* 52 (8): 849–58.
- Prasad, Aparna, Eline D. Lorenzen, and Michael V. Westbury. 2022. "Evaluating the Role of Reference-Genome Phylogenetic Distance on Evolutionary Inference." *Molecular Ecology Resources* 22 (1): 45–55.
- Proulx, G., & Aubry, K. B. (2017). *The Martes complex: A monophyletic clade that shares many life-history traits and conservation challenges*. In *The Martes Complex in the 21st Century: ecology and conservation*, Proulx, Gilbert, Eds. Mammal Research Institute, Polish Academy of Sciences, Białowieża, Poland: 3-24.

- Quinlan, Aaron R., and Ira M. Hall. 2012. "Characterizing Complex Structural Variation in Germline and Somatic Genomes." *Trends in Genetics: TIG* 28 (1): 43–53.
- Radke, David W., and Charles Lee. 2015. "Adaptive Potential of Genomic Structural Variation in Human and Mammalian Evolution." *Briefings in Functional Genomics* 14 (5): 358–68.
- Redon, Richard, Shumpei Ishikawa, Karen R. Fitch, Lars Feuk, George H. Perry, T. Daniel Andrews, Heike Fiegler, et al. 2006. "Global Variation in Copy Number in the Human Genome." *Nature* 444 (7118): 444–54.
- Rhie, Arang, Shane A. McCarthy, Olivier Fedrigo, Joana Damas, Giulio Formenti, Sergey Koren, Marcela Uliano-Silva, et al. 2021. "Towards Complete and Error-Free Genome Assemblies of All Vertebrate Species." *Nature* 592 (7856): 737–46.
- Rhoads, Anthony, and Kin Fai Au. 2015. "PacBio Sequencing and Its Applications." *Genomics, Proteomics & Bioinformatics* 13 (5): 278–89.
- Rinker, David C., Natalya K. Specian, Shu Zhao, and John G. Gibbons. 2019. "Polar Bear Evolution Is Marked by Rapid Changes in Gene Copy Number in Response to Dietary Shift." *Proceedings of the National Academy of Sciences of the United States of America* 116 (27): 13446–51.
- Sadowski, Michal, Agnieszka Kraft, Przemyslaw Szalaj, Michal Wlasnowolski, Zhonghui Tang, Yijun Ruan, and Dariusz Plewczynski. 2019. "Spatial Chromatin Architecture Alteration by Structural Variations in Human Genomes at the Population Scale." *Genome Biology* 20 (1): 148.
- Sagara, Shigeki, Tomohiro Osanai, Taihei Itoh, Kei Izumiyama, Shuji Shibutani, Kenji Hanada, Hiroaki Yokoyama, et al. 2012. "Overexpression of Coupling Factor 6 Attenuates Exercise-Induced Physiological Cardiac Hypertrophy by Inhibiting PI3K/Akt Signaling in Mice." *Journal of Hypertension* 30 (4): 778–86.
- Samonte, Rhea Vallente, and Evan E. Eichler. 2002. "Segmental Duplications and the Evolution of the Primate Genome." *Nature Reviews. Genetics* 3 (1): 65–72.
- Sanders, Ashley D., Ester Falconer, Mark Hills, Diana C. J. Spierings, and Peter M. Lansdorp. 2017. "Single-Cell Template Strand Sequencing by Strand-Seq Enables the Characterization of Individual Homologs." *Nature Protocols* 12 (6): 1151–76.
- Sanders, Ashley D., Mark Hills, David Porubský, Victor Guryev, Ester Falconer, and Peter M. Lansdorp. 2016. "Characterizing Polymorphic Inversions in Human Genomes by Single-Cell Sequencing." *Genome Research* 26 (11): 1575–87.
- Santymire, R. M., E. V. Lonsdorf, C. M. Lynch, D. E. Wildt, P. E. Marinari, J. S. Kreeger, and J. G. Howard. 2019. "Inbreeding Causes Decreased Seminal Quality Affecting Pregnancy and Litter Size in the Endangered Black-footed Ferret." *Animal Conservation* 22 (4): 331–40.
- Schrader, Lukas, and Jürgen Schmitz. 2019. "The Impact of Transposable Elements in Adaptive Evolution." *Molecular Ecology* 28 (6): 1537–49.
- Schrauwen, Isabelle, Arnaud Pj Giese, Abdul Aziz, David Tino Lafont, Imen Chakchouk, Regie Lyn P. Santos-Cortez, Kwanghyuk Lee, et al. 2019. "FAM92A Underlies Nonsyndromic Postaxial Polydactyly in Humans and an Abnormal Limb and Digit Skeletal Phenotype in Mice." *Journal of Bone and Mineral Research: The Official Journal of the American Society for Bone and Mineral Research* 34 (2): 375–86.
- Sedlazeck, Fritz J., Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt von Haeseler, and Michael C. Schatz. 2018. "Accurate Detection of Complex Structural Variations Using Single-Molecule Sequencing." *Nature Methods* 15 (6): 461–68.
- Shao, Xin, Ning Lv, Jie Liao, Jinbo Long, Rui Xue, Ni Ai, Donghang Xu, and Xiaohui Fan. 2019. "Copy Number Variation Is Highly Correlated with Differential Gene Expression: A Pan-Cancer Study." *BMC Medical Genetics* 20 (1): 175.
- Sharma, Virag, Nikolai Hecker, Juliana G. Roscito, Leo Foerster, Bjoern E. Langer, and Michael Hiller. 2018. "A Genomics Approach Reveals Insights into the Importance of Gene Losses for Mammalian Adaptations." *Nature Communications* 9 (1): 1215.
- Sharma, Virag, and Michael Hiller. 2020. "Losses of Human Disease-Associated Genes in Placental Mammals." *NAR Genomics and Bioinformatics* 2 (1): lqz012.
- Sharp, Andrew J., Ze Cheng, and Evan E. Eichler. 2006. "Structural Variation of the Human Genome." *Annual Review of Genomics and Human Genetics* 7: 407–42.
- Shastri, Barkur S. 2009. "SNPs: Impact on Gene Function and Phenotype." In *Single Nucleotide Polymorphisms: Methods and Protocols*, edited by Anton A. Komar, 3–22. Totowa, NJ: Humana Press.
- Shaw, Christine J., and James R. Lupski. 2004. "Implications of Human Genome Architecture for Rearrangement-Based Disorders: The Genomic Basis of Disease." *Human Molecular Genetics* 13 Spec No 1 (April): R57–64.
- Shlien, Adam, and David Malkin. 2009. "Copy Number Variations and Cancer." *Genome Medicine* 1 (6): 62.
- Shumate, Alaina, and Steven L. Salzberg. 2020. "Liftoff: An Accurate Gene Annotation Mapping Tool."

<https://doi.org/10.1101/2020.06.24.169680>.

- Sims, David, Ian Sudbery, Nicholas E. Illott, Andreas Heger, and Chris P. Ponting. 2014. "Sequencing Depth and Coverage: Key Considerations in Genomic Analyses." *Nature Reviews. Genetics* 15 (2): 121–32.
- Sironen, A., P. Uimari, T. Iso-Touru, and J. Vilkki. 2012. "L1 Insertion within SPEF2 Gene Is Associated with Increased Litter Size in the Finnish Yorkshire Population." *Journal of Animal Breeding and Genetics = Zeitschrift Fur Tierzucht Und Zuchtungsbiologie* 129 (2): 92–97.
- Sohn, Jang-Il, and Jin-Wu Nam. 2018. "The Present and Future of de Novo Whole-Genome Assembly." *Briefings in Bioinformatics* 19 (1): 23–40.
- Soley, Fernando G., and Isaías Alvarado-Díaz. 2011. "Prospective Thinking in a Mustelid? Eira Barbara (Carnivora) Cache Unripe Fruits to Consume Them Once Ripened." *Die Naturwissenschaften* 98 (8): 693–98.
- Spielman, Derek, Barry W. Brook, and Richard Frankham. 2004. "Most Species Are Not Driven to Extinction before Genetic Factors Impact Them." *Proceedings of the National Academy of Sciences of the United States of America* 101 (42): 15261–64.
- Stankiewicz, Paweł, and James R. Lupski. 2002. "Genome Architecture, Rearrangements and Genomic Disorders." *Trends in Genetics: TIG* 18 (2): 74–82.
- Stanzione, Marcello, Marek Baumann, Frantzeskos Papanikos, Ihsan Dereli, Julian Lange, Angelique Ramlal, Daniel Tränkner, et al. 2016. "Meiotic DNA Break Formation Requires the Unsynapsed Chromosome Axis-Binding Protein IHO1 (CCDC36) in Mice." *Nature Cell Biology* 18 (11): 1208–20.
- Stefansson, Hreinn, Agnar Helgason, Gudmar Thorleifsson, Valgerdur Steinthorsdottir, Gisli Masson, John Barnard, Adam Baker, et al. 2005. "A Common Inversion under Selection in Europeans." *Nature Genetics* 37 (2): 129–37.
- Steinmann, Katharina, David N. Cooper, Lan Kluwe, Nadia A. Chuzhanova, Cornelia Senger, Eduard Serra, Conxi Lazaro, et al. 2007. "Type 2 NF1 Deletions Are Highly Unusual by Virtue of the Absence of Nonallelic Homologous Recombination Hotspots and an Apparent Preference for Female Mitotic Recombination." *American Journal of Human Genetics* 81 (6): 1201–20.
- Stroud, James T., and Jonathan B. Losos. 2016. "Ecological Opportunity and Adaptive Radiation." *Annual Review of Ecology, Evolution, and Systematics* 47 (1): 507–32.
- Sturtevant, A. H. 1921. "A Case of Rearrangement of Genes in *Drosophila*." *Proceedings of the National Academy of Sciences of the United States of America* 7 (8): 235–37.
- Sturtevant, A. H. 1925. "The Effects of Unequal Crossing over at the Bar Locus in *Drosophila*." *Genetics* 10 (2): 117–47.
- Sturtevant, A. H., and K. Mather. 1938. "The Interrelations of Inversions, Heterosis and Recombination." *The American Naturalist* 72 (742): 447–52.
- Sudmant, Peter H., Tobias Rausch, Eugene J. Gardner, Robert E. Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, et al. 2015. "An Integrated Map of Structural Variation in 2,504 Human Genomes." *Nature* 526 (7571): 75–81.
- Taraborrelli, Stefania. 2015. "Physiology, Production and Action of Progesterone." *Acta Obstetricia et Gynecologica Scandinavica* 94 Suppl 161 (November): 8–16.
- Taye, Mengistie, Joon Yoon, Tadelle Dessie, Seoae Cho, Sung Jong Oh, Hak-Kyo Lee, and Heebal Kim. 2018. "Deciphering Signature of Selection Affecting Beef Quality Traits in Angus Cattle." *Genes & Genomics* 40 (1): 63–75.
- The UniProt Consortium. 2017. "UniProt: The Universal Protein Knowledgebase." *Nucleic Acids Research* 45 (D1): D158–69.
- Trask, Barbara J. 2002. "Human Cytogenetics: 46 Chromosomes, 46 Years and Counting." *Nature Reviews. Genetics* 3 (10): 769–78.
- Turner, Daniel J., Marcos Miretti, Diana Rajan, Heike Fiegler, Nigel P. Carter, Martyn L. Blayney, Stephan Beck, and Matthew E. Hurler. 2008. "Germline Rates of de Novo Meiotic Deletions and Duplications Causing Several Genomic Disorders." *Nature Genetics* 40 (1): 90–95.
- Tuzun, Eray, Andrew J. Sharp, Jeffrey A. Bailey, Rajinder Kaul, V. Anne Morrison, Lisa M. Pertz, Eric Haugen, et al. 2005. "Fine-Scale Structural Variation of the Human Genome." *Nature Genetics* 37 (7): 727–32.
- Udagawa, Osamu, Chizuru Ito, Narumi Ogonuki, Hiroyasu Sato, Shoken Lee, Pearlta Tripvanuntakul, Ikuyo Ichi, et al. 2014. "Oligo-Astheno-Teratozoospermia in Mice Lacking ORP4, a Sterol-Binding Protein in the OSBP-Related Protein Family." *Genes to Cells: Devoted to Molecular & Cellular Mechanisms* 19 (1): 13–27.
- Van't Hof, Arjen E., Pascal Campagne, Daniel J. Rigden, Carl J. Yung, Jessica Lingley, Michael A. Quail, Neil Hall, Alistair C. Darby, and Ilik J. Saccheri. 2016. "The Industrial Melanism Mutation in British Peppered Moths Is a Transposable Element." *Nature* 534 (7605): 102–5.

- Viscarra, Jose A., Ruben Rodriguez, Jose Pablo Vazquez-Medina, Andrew Lee, Michael S. Tift, Stephen K. Tavoni, Daniel E. Crocker, and Rudy M. Ortiz. 2013. "Insulin and GLP-1 Infusions Demonstrate the Onset of Adipose-Specific Insulin Resistance in a Large Fasting Mammal: Potential Glucogenic Role for GLP-1." *Physiological Reports* 1 (2): e00023.
- Wang, Guo-Dong, Hai-Bing Xie, Min-Sheng Peng, David Irwin, and Ya-Ping Zhang. 2014. "Domestication Genomics: Evidence from Animals." *Annual Review of Animal Biosciences* 2 (February): 65–84.
- Wang, Jinkai, Zhi-Xiang Lu, Collin J. Tokheim, Sara E. Miller, and Yi Xing. 2015. "Species-Specific Exon Loss in Human Transcriptomes." *Molecular Biology and Evolution* 32 (2): 481–94.
- Wang, Kai, Mingyao Li, Dexter Hadley, Rui Liu, Joseph Glessner, Struan F. A. Grant, Hakon Hakonarson, and Maja Bucan. 2007. "PennCNV: An Integrated Hidden Markov Model Designed for High-Resolution Copy Number Variation Detection in Whole-Genome SNP Genotyping Data." *Genome Research* 17 (11): 1665–74.
- Weischenfeldt, Joachim, Orsolya Symmons, François Spitz, and Jan O. Korbel. 2013. "Phenotypic Impact of Genomic Structural Variation: Insights from and for Human Disease." *Nature Reviews. Genetics* 14 (2): 125–38.
- Weisenfeld, Neil I., Vijay Kumar, Preyas Shah, Deanna M. Church, and David B. Jaffe. 2017. "Direct Determination of Diploid Genome Sequences." *Genome Research* 27 (5): 757–67.
- Weissensteiner, Matthias H., Ignas Bunikis, Ana Catalán, Kees-Jan Francoijs, Ulrich Knief, Wieland Heim, Valentina Peona, et al. 2020. "Discovery and Population Genomics of Structural Variation in a Songbird Genus." *Nature Communications* 11 (1): 3403.
- Wellenreuther, Maren, Claire Mérot, Emma Berdan, and Louis Bernatchez. 2019. "Going beyond SNPs: The Role of Structural Genomic Variants in Adaptive Evolution and Species Diversification." *Molecular Ecology* 28 (6): 1203–9.
- Welm, Bryan, Joni Mott, and Zena Werb. 2002. "Developmental Biology: Vasculogenesis Is a Wreck without RECK." *Current Biology: CB* 12 (6): R209–11.
- Wenger, Aaron M., Paul Peluso, William J. Rowell, Pi-Chuan Chang, Richard J. Hall, Gregory T. Concepcion, Jana Ebler, et al. 2019. "Accurate Circular Consensus Long-Read Sequencing Improves Variant Detection and Assembly of a Human Genome." *Nature Biotechnology* 37 (10): 1155–62.
- Weterings, Eric, and Dik C. van Gent. 2004. "The Mechanism of Non-Homologous End-Joining: A Synopsis of Synapsis." *DNA Repair* 3 (11): 1425–35.
- Wilkinson, Alex W., Jonathan Diep, Shaobo Dai, Shuo Liu, Yaw Shin Ooi, Dan Song, Tie-Mei Li, et al. 2019. "SETD3 Is an Actin Histidine Methyltransferase That Prevents Primary Dystocia." *Nature* 565 (7739): 372–76.
- Wilson, Kerianne M., Victoria A. Wagner, and Wendy Saltzman. 2022. "Specificity of California Mouse Pup Vocalizations in Response to Olfactory Stimuli." *Developmental Psychobiology* 64 (4): e22261.
- Wisely, S. M., S. W. Buskirk, M. A. Fleming, D. B. McDonald, and E. A. Ostrander. 2002. "Genetic Diversity and Fitness in Black-Footed Ferrets before and during a Bottleneck." *The Journal of Heredity* 93 (4): 231–37.
- Wiszniewska, Joanna, Weimin Bi, Chad Shaw, Pawel Stankiewicz, Sung-Hae L. Kang, Amber N. Pursley, Seema Lalani, et al. 2014. "Combined Array CGH plus SNP Genome Analyses in a Single Assay for Optimized Clinical Testing." *European Journal of Human Genetics: EJHG* 22 (1): 79–87.
- Wold, Jana, Klaus-Peter Koepfli, Stephanie J. Galla, David Eccles, Carolyn J. Hogg, Marissa F. Le Lec, Joseph Guhlin, Anna W. Santure, and Tammy E. Steeves. 2021. "Expanding the Conservation Genomics Toolbox: Incorporating Structural Variants to Enhance Genomic Studies for Species of Conservation Concern." *Molecular Ecology* 30 (23): 5949–65.
- Wolfner, Mariana F., and Danny E. Miller. 2016. "Alfred Sturtevant Walks into a Bar: Gene Dosage, Gene Position, and Unequal Crossing Over in *Drosophila*." *Genetics* 204 (3): 833–35.
- Wood, Andrew R., Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H. Pers, Stefan Gustafsson, Audrey Y. Chu, et al. 2014. "Defining the Role of Common Variation in the Genomic and Biological Architecture of Adult Human Height." *Nature Genetics* 46 (11): 1173–86.
- Xie, Chen, Cemalettin Bekpen, Sven Künzel, Maryam Keshavarz, Rebecca Krebs-Wheaton, Neva Skrabar, Kristian K. Ullrich, Wenyu Zhang, and Diethard Tautz. 2020. "Dedicated Transcriptomics Combined with Power Analysis Lead to Functional Understanding of Genes with Weak Phenotypic Changes in Knockout Lines." *PLoS Computational Biology* 16 (11): e1008354.
- Xing, Jinchuan, Yuhua Zhang, Kyudong Han, Abdel Halim Salem, Shurjo K. Sen, Chad D. Huff, Qiong Zhou, et al. 2009. "Mobile Elements Create Structural Variation: Analysis of a Complete Human Genome." *Genome Research* 19 (9): 1516–26.
- Xing, Yi, and Christopher Lee. 2006. "Alternative Splicing and RNA Selection Pressure--Evolutionary Consequences for Eukaryotic Genomes." *Nature Reviews. Genetics* 7 (7): 499–509.

- Xu, Hongwei, Haixia Li, Zhen Wang, Ayimuguli Abudureyimu, Jutian Yang, Xin Cao, Xianyong Lan, Rongxin Zang, and Yong Cai. 2020. "A Deletion Downstream of the CHCHD7 Gene Is Associated with Growth Traits in Sheep." *Animals: An Open Access Journal from MDPI* 10 (9). <https://doi.org/10.3390/ani10091472>.
- Young, Alexander I., Stefania Benonisdottir, Molly Przeworski, and Augustine Kong. 2019. "Deconstructing the Sources of Genotype-Phenotype Associations in Humans." *Science* 365 (6460): 1396–1400.
- Yuan, Yuan, Yaolei Zhang, Peijun Zhang, Chang Liu, Jiahao Wang, Haiyu Gao, A. Rus Hoelzel, et al. 2021. "Comparative Genomics Provides Insights into the Aquatic Adaptations of Mammals." *Proceedings of the National Academy of Sciences of the United States of America* 118 (37). <https://doi.org/10.1073/pnas.2106080118>.
- Zarrei, Mehdi, Jeffrey R. MacDonald, Daniele Merico, and Stephen W. Scherer. 2015. "A Copy Number Variation Map of the Human Genome." *Nature Reviews. Genetics* 16 (3): 172–83.
- Zhang, Feng, Mehrdad Khajavi, Anne M. Connolly, Charles F. Towne, Sat Dev Batish, and James R. Lupski. 2009. "The DNA Replication FoSTeS/MMBIR Mechanism Can Generate Genomic, Genic and Exonic Complex Rearrangements in Humans." *Nature Genetics* 41 (7): 849–53.
- Zhang, Rui-Qian, Jun-Jie Wang, Teng Zhang, Hong-Li Zhai, and Wei Shen. 2019. "Copy-Number Variation in Goat Genome Sequence: A Comparative Analysis of the Different Litter Size Trait Groups." *Gene* 696 (May): 40–46.
- Zhang, Yaran, Yan Hu, Xiuge Wang, Qiang Jiang, Han Zhao, Jinpeng Wang, Zhihua Ju, et al. 2019. "Population Structure, and Selection Signatures Underlying High-Altitude Adaptation Inferred From Genome-Wide Copy Number Variations in Chinese Indigenous Cattle." *Frontiers in Genetics* 10: 1404.
- Zhao, Pengju, Junhui Li, Huimin Kang, Haifei Wang, Ziyao Fan, Zongjun Yin, Jiafu Wang, Qin Zhang, Zhiquan Wang, and Jian-Feng Liu. 2016. "Structural Variant Detection by Large-Scale Sequencing Reveals New Evolutionary Evidence on Breed Divergence between Chinese and European Pigs." *Scientific Reports* 6 (January): 18501.
- Zhao, Xuefang, Ryan L. Collins, Wan-Ping Lee, Alexandra M. Weber, Yukyung Jun, Qihui Zhu, Ben Weisburd, et al. 2021. "Expectations and Blind Spots for Structural Variation Detection from Long-Read Assemblies and Short-Read Genome Sequencing Technologies." *American Journal of Human Genetics* 108 (5): 919–28.
- Zhou, Xuming, Di Sun, Xuanmin Guang, Siming Ma, Xiaodong Fang, Marco Mariotti, Rasmus Nielsen, Vadim N. Gladyshev, and Guang Yang. 2018. "Molecular Footprints of Aquatic Adaptation Including Bone Mass Changes in Cetaceans." *Genome Biology and Evolution* 10 (3): 967–75.
- Zichner, Thomas, David A. Garfield, Tobias Rausch, Adrian M. Stütz, Enrico Cannavó, Martina Braun, Eileen E. M. Furlong, and Jan O. Korb. 2013. "Impact of Genomic Structural Variation in *Drosophila melanogaster* Based on Population-Scale Sequencing." *Genome Research* 23 (3): 568–79.

Appendix A.

Supporting information for Chapter I, Derežanin et al., 2022
**Multiple types of genomic variation contribute to adaptive traits
in the mustelid subfamily Guloninae**

Supporting tables S4 - S6A can be accessed at:

<https://onlinelibrary.wiley.com/doi/10.1111/mec.16443>

Supplemental information for:

Multiple types of genomic variation contribute to adaptive traits in the mustelid subfamily Guloninae

Lorena Derežanin, Asta Blažytė, Pavel Dobrynin, David A. Duchêne, José H. Grau, Sungwon Jeon, Sergei Kliver, Klaus-Peter Koepfli, Dorina Meneghini, Michaela Preick, Andrey Tomarovsky, Azamat Totikov, Jörns Fickel, Daniel W. Förster

Table of contents:

Multiple types of genomic variation contribute to adaptive traits in the mustelid subfamily Guloninae	
SUPPLEMENTAL FIGURES AND TABLES	1
Figure S1. Genome alignment between domestic ferret and tayra assemblies.	1
Table S1. Tayra genome assembly metrics.	2
Figure S2. Genome completeness metrics.	3
Table S2. Major types of transposable elements found in the genomes of four mustelid species.	4
Figure S3. Repeat landscape of the tayra genome assembly.	4
Figure S4. Historical demography of three gulonine species.	5
Figure S5. Coverage plots for (A) the tayra, (B) the sable, and (C) the wolverine.	6
Table S3. Statistics for counts of heterozygous SNPs.	7
Figure S6. Genome-wide heterozygosity for tayra (<i>Eira barbara</i>), sable (<i>Martes zibellina</i>) and wolverine (<i>Gulo gulo</i>).	8
Figure S7. Gene family expansions and contractions for eight carnivoran species.	8
Figure S8. Structural variants detected in gulonine species.	9
Figure S9. Heterozygous and homozygous SVs detected in each gulonine species.	10
SUPPLEMENTAL INFORMATION	11
Genome assembly and processing of sequencing data	11
Alignment to pseudochromosome assemblies and sex verification	11
Phylogenomic data preparation, analysis and dating	12
Nucleotide diversity	13
Gene copies completeness assessment	14
Structural variants	14
REFERENCES	15
Supplemental Tables S4-S6 - References	18

SUPPLEMENTAL FIGURES AND TABLES

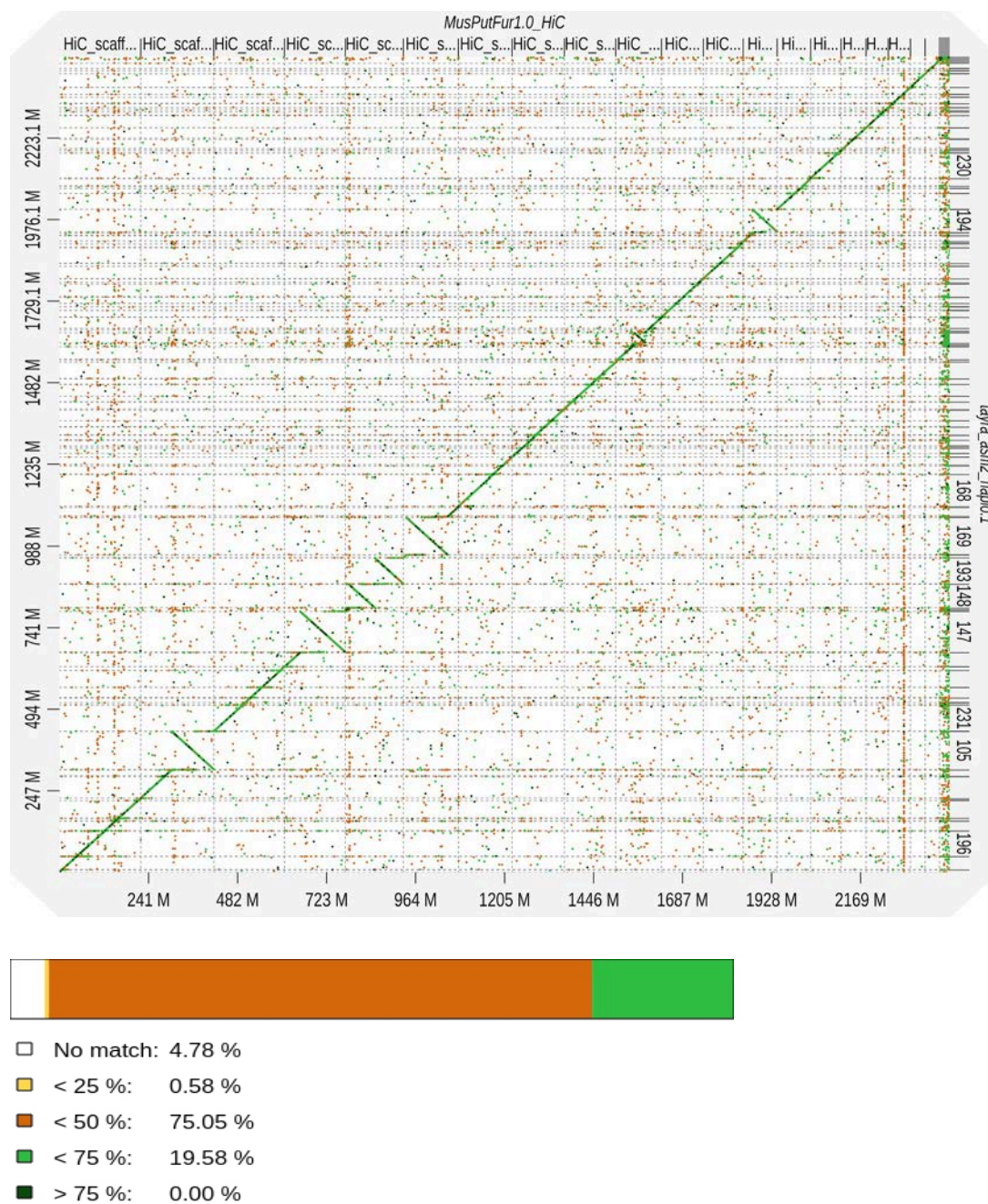


Figure S1. Genome alignment between domestic ferret and tayra assemblies. Synteny analysis of domestic ferret Hi-C genome assembly (upper X-axis) and tayra assembly (right Y-axis) with overall match indicating 95% identity (with > 50% similarity).

Table S1. Tayra genome assembly metrics.

Combined assembly metrics obtained from Supernova assembler and Quality Assessment Tool for Genome Assemblies (QUAST).

Supernova assembly metrics (v2.1.1)	
Sequencing strategy	Illumina NovaSeq + 10x Genomics
Assembly method	Supernova assembler v2.1.1
Sequencing reads	1.3 billion reads, 150 PE
Mean read length	139.5 bases
Raw coverage	75.63x
Effective coverage	55.34x
Proper read pairs	88.78 %
Median insert size	344 bases
GC content	41.76 %
Molecule length	50.75 kb
Contig N50	289.96 kb
Scaffold N50	42.07 Mb
Number of scaffolds	14579
Number of scaffolds (≥ 5 kb)	3773
Longest scaffold	123.17 Mb
Phaseblock N50	5.77 Mb
Assembly size	2.447 Gb (scaffolds ≥ 5 kb)
QUAST assembly metrics (v5.0.2)	
Scaffolds > 10 kb	1621
Scaffolds > 25 kb	469
Scaffolds > 50 kb	254
Scaffold L50	17
N's per 100 kbp	1145.00

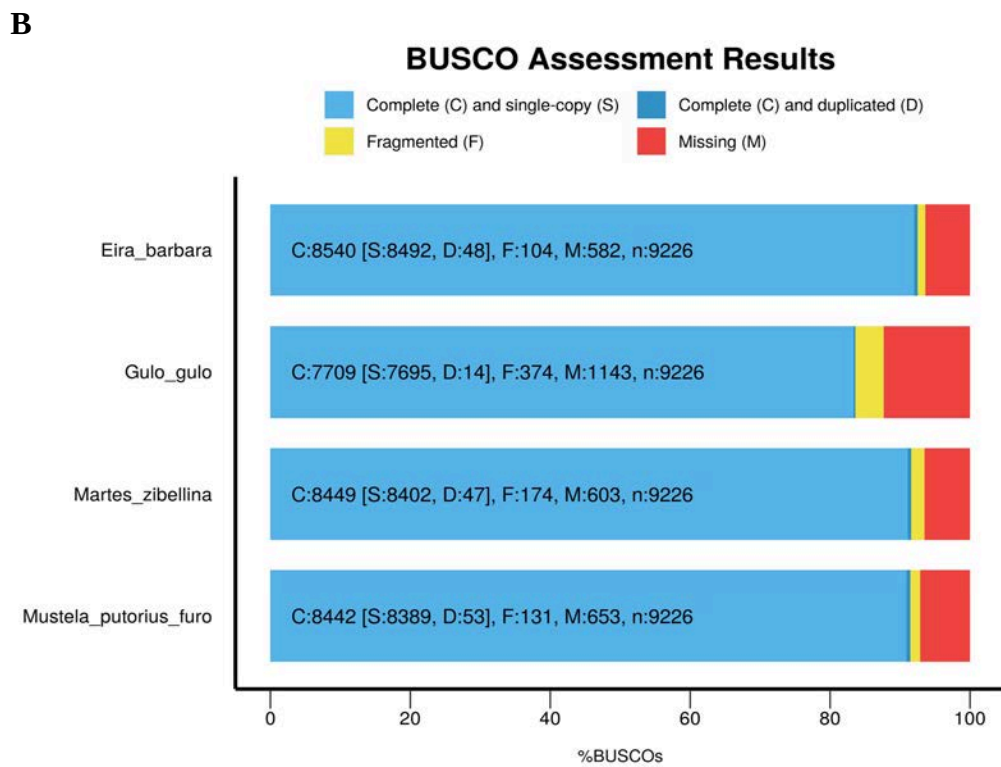
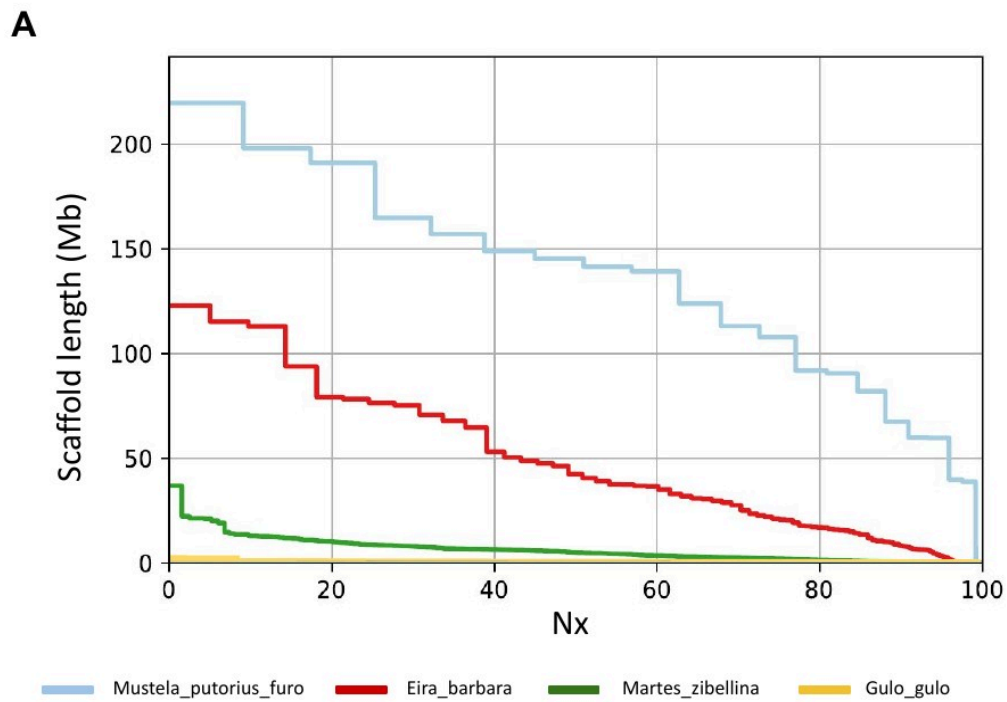


Figure S2. Genome completeness metrics.

A) Summary of QUAST assembly completeness analysis (number of contigs is represented on the X-axis) and (B) comparison of the BUSCO gene completeness assessment for four mustelid species.

Table S2. Major types of transposable elements found in the genomes of four mustelid species.

Type	Tayra (<i>Eira barbara</i>)	Sable (<i>Martes zibellina</i>)	Wolverine (<i>Gulo gulo</i>)	Domestic ferret (<i>M. putorius furo</i>)
SINE	1.91 % (47 Mb)	1.5 % (36 Mb)	1.56 % (37 Mb)	1.36 % (32 Mb)
LINE	23.17 % (570 Mb)	19 % (465 Mb)	15.73 % (381 Mb)	17.13 % (417 Mb)
LTR	4.05 % (99 Mb)	3.5 % (84 Mb)	3.28 % (79 Mb)	4.4 % (106 Mb)
DNA	1.63 % (40 Mb)	2.03 % (49 Mb)	2.05 % (49 Mb)	3.13 % (75 Mb)
Total	32.96 % (814 Mb)	28.15 % (681 Mb)	25.19 % (610 Mb)	28.86 % (695 Mb)

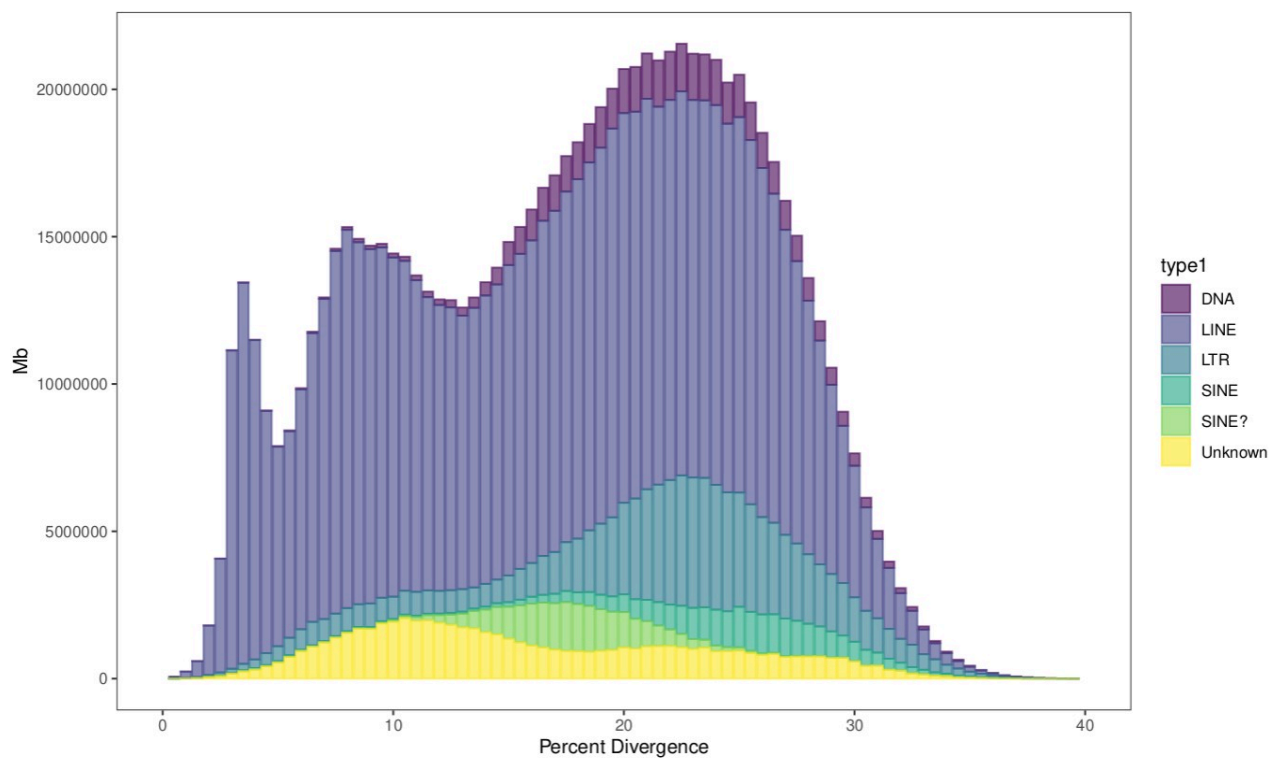


Figure S3. Repeat landscape of the tayra genome assembly.

Coloring scheme refers to different proportions of different repeat classes and x-axis indicates percent divergence within different repeat classes.

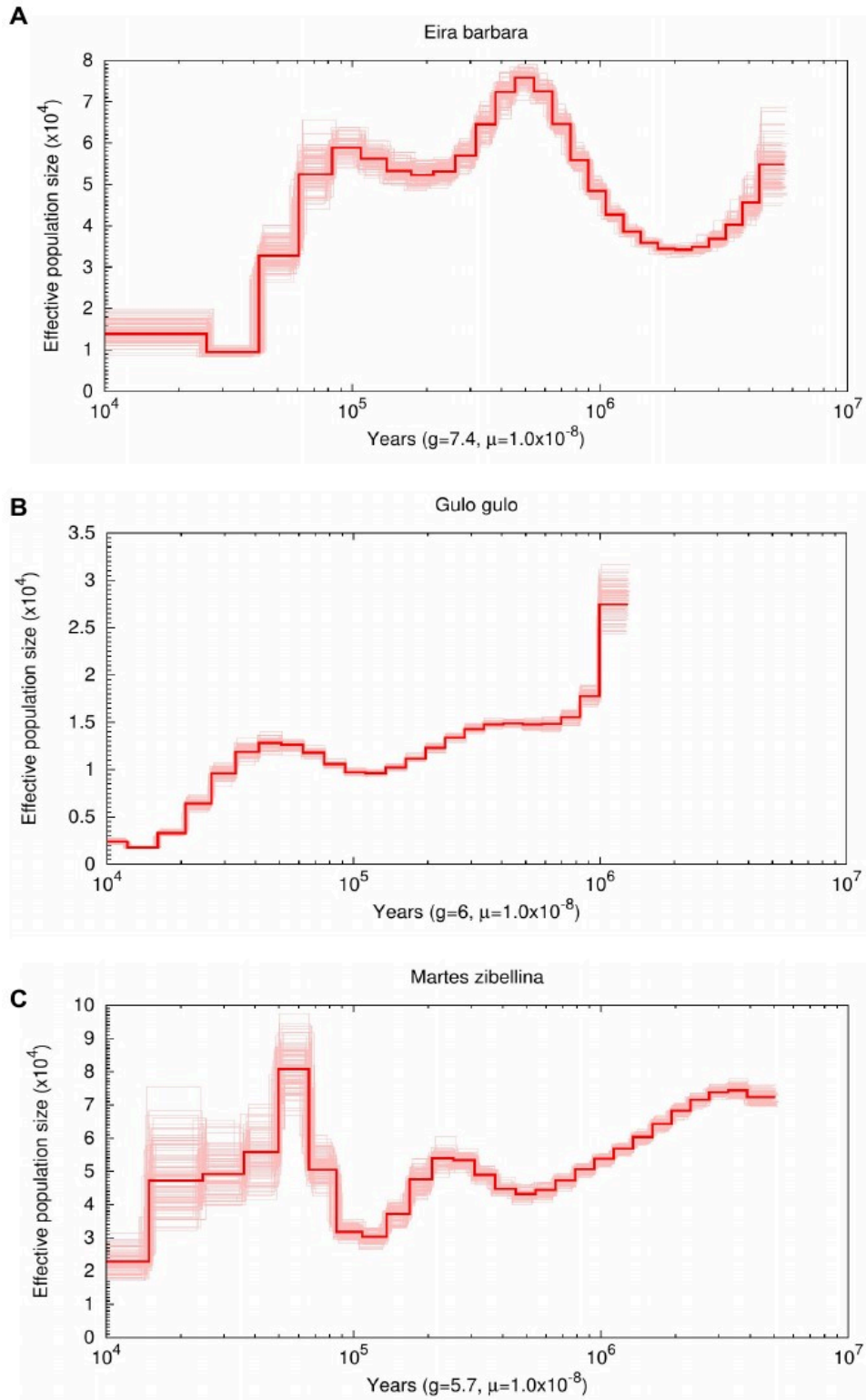


Figure S4. Historical demography of three gulonine species.

Bootstrapped inference of effective population size change over time for (A) tayra, (B) wolverine, and (C) sable. Historical demography for all three species was inferred under different generation times (*E.barbara* = 7.4y, *G.gulo* = 6y, *M.zibellina* = 5.7y). Time scale on the x-axis is calculated assuming a mutation rate of 1.0×10^{-8} .

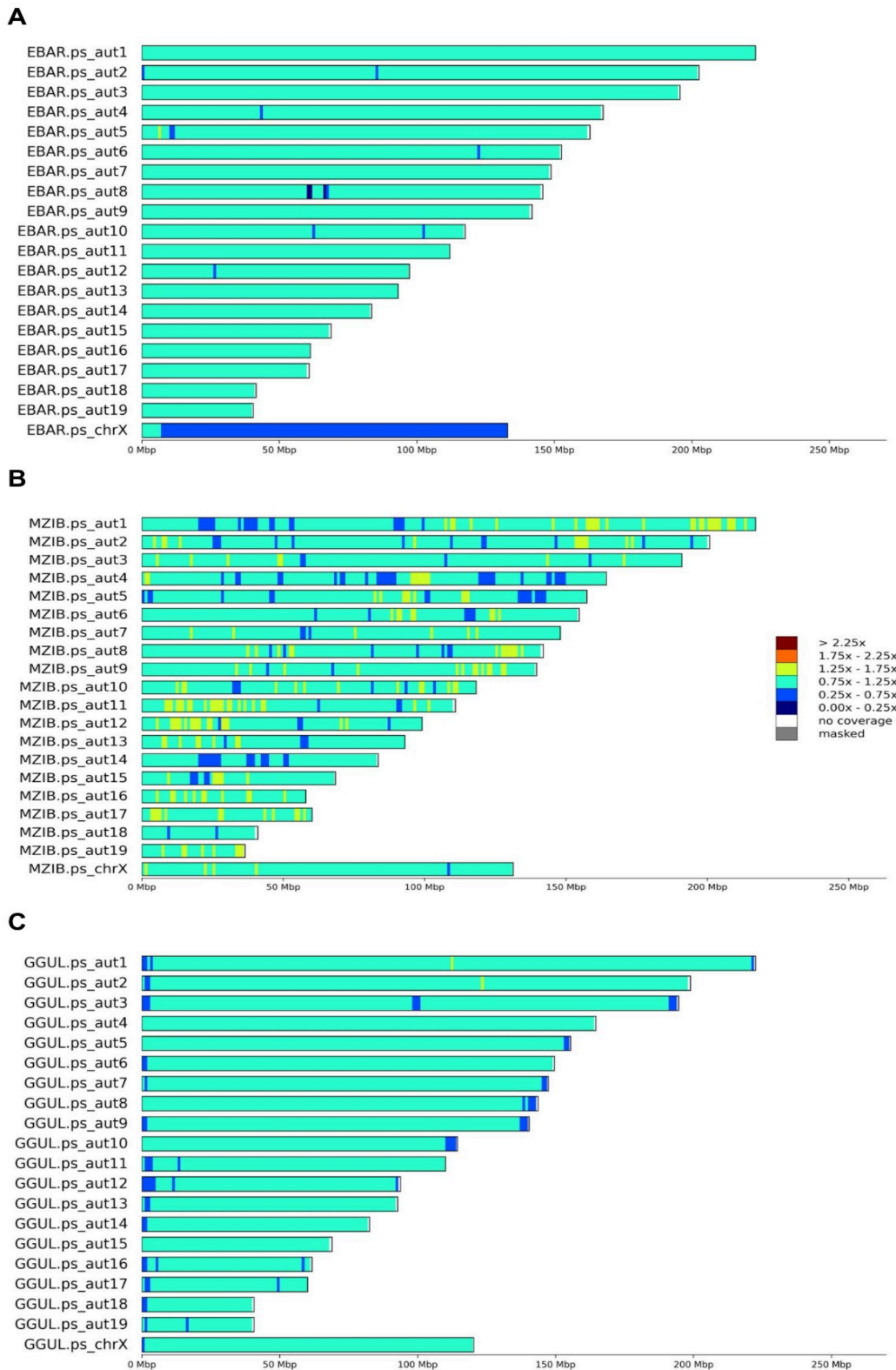


Figure S5. Coverage plots for (A) the tayra, (B) the sable, and (C) the wolverine. Coverage was calculated in non-overlapping sliding windows of 1 Mbp and divided by whole genome median coverage.

Table S3. Statistics for counts of heterozygous SNPs.

SNPs are counted in 1 Mbp non-overlapping sliding windows for tayra, sable and wolverine scaled to SNPs per kbp. Bold marks the median values.

Species	All scaffolds, heterozygous SNPs/kbp				Without X pseudo-chromosome, heterozygous SNPs/kbp			
	Min	Max	Median	Mean	Min	Max	Median	Mean
<i>Eira barbara</i>	0	5.99	1.89	1.81	0	5.99	1.93	1.90
<i>Martes zibellina</i>	0	4.45	1.44	1.51	0	4.45	1.47	1.56
<i>Gulo gulo</i>	0	1.24	0.28	0.27	0	1.24	0.29	0.28

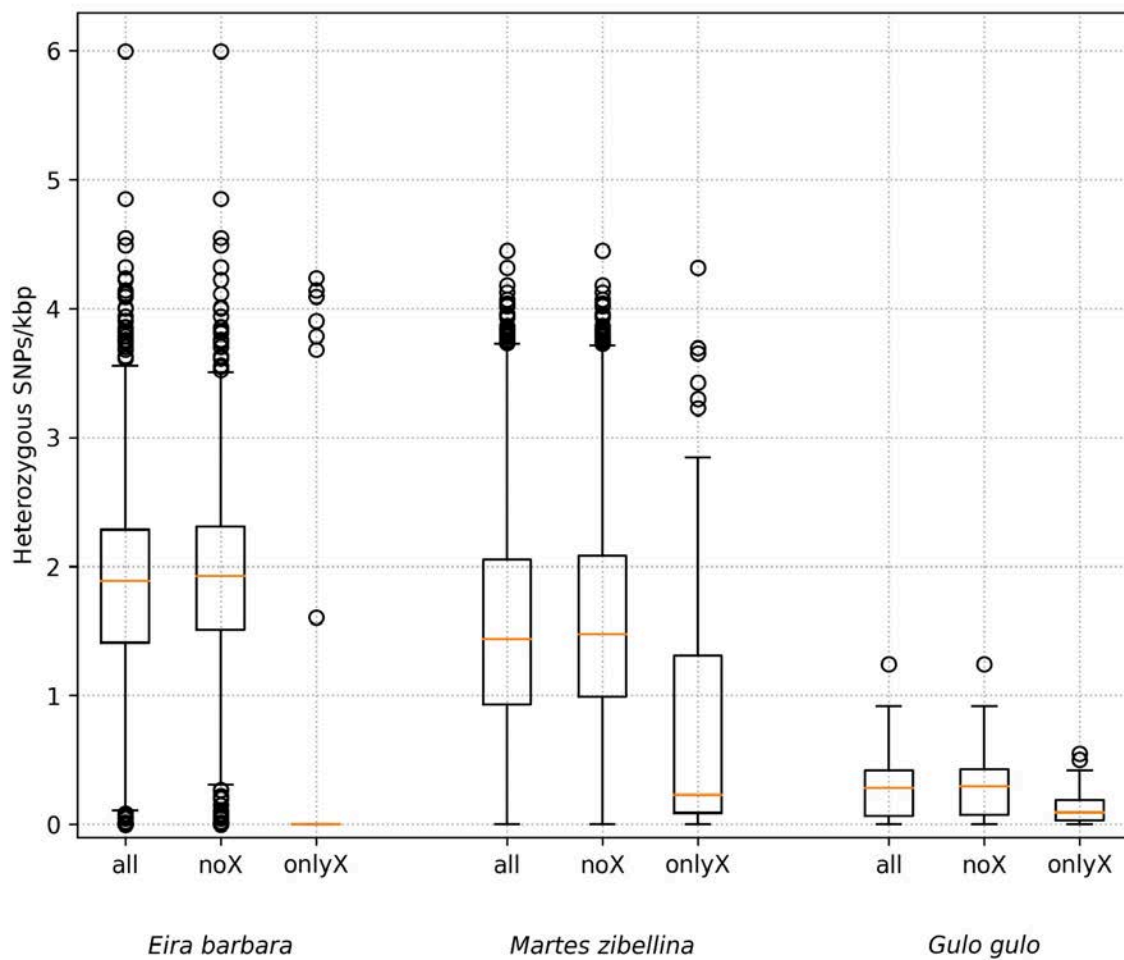


Figure S6. Genome-wide heterozygosity for tayra (*Eira barbara*), sable (*Martes zibellina*) and wolverine (*Gulo gulo*).

SNPs are counted in 1Mbp non-overlapping sliding windows and scaled to heterozygous SNPs per kbp. Corresponding pseudochromosome assemblies based on the domestic ferret assembly were used as reference. ‘all’ includes windows from all scaffolds of at least 1 Mbp, ‘noX’ - without windows from the X pseudochromosome C-scaffold (ps_chrX), ‘onlyX’ - only windows from ps_chrX . Exclusion of ps_chrX slightly affected the borders of boxes and whiskers (all vs noX). Orange lines on boxplots correspond to median values, box edges - to 25th and 75th percentiles (Q1 and Q3), lower whisker - to $\max(0, Q1 - 1.5 \cdot IQR)$, upper whisker - to $Q3 + 1.5 \cdot IQR$, respectively. Interquartile range equals $Q3 - Q1$. Circles are outliers, i.e values outside $[\max(0, Q1 - 1.5 \cdot IQR), Q3 + 1.5 \cdot IQR]$.

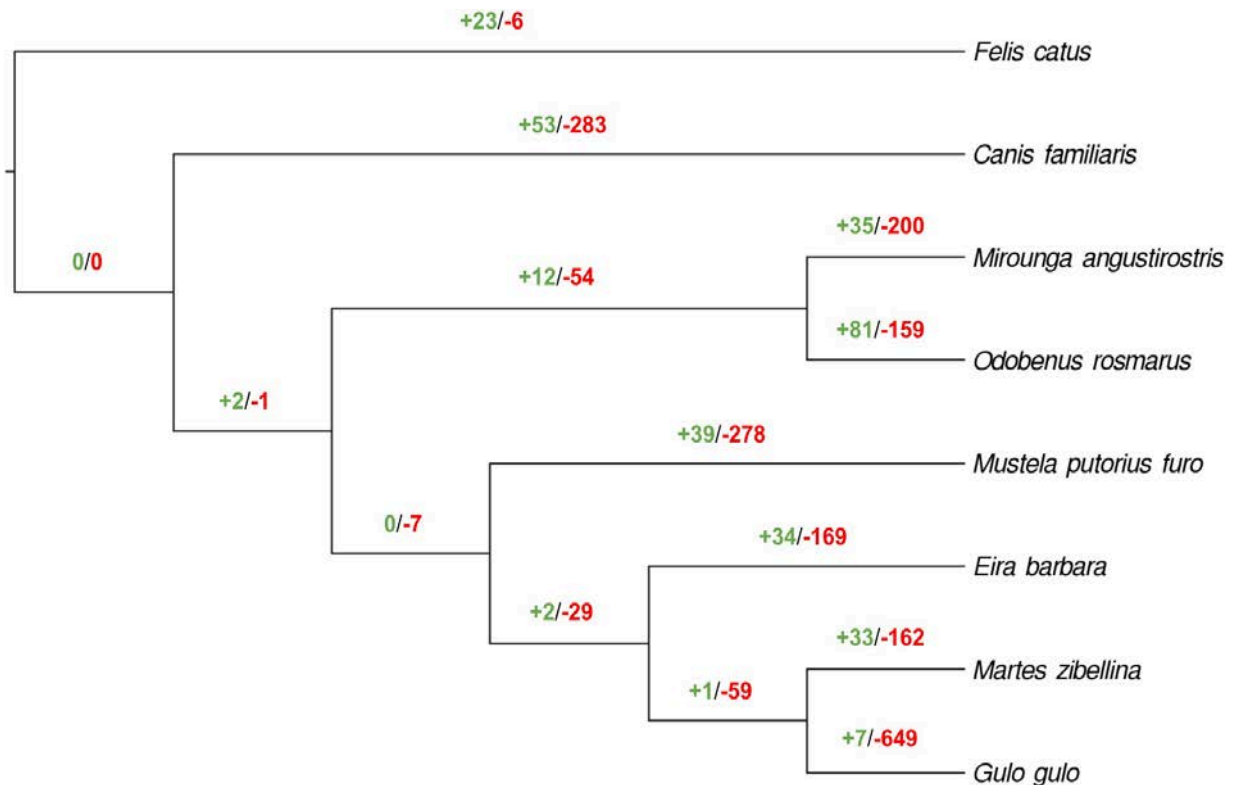


Figure S7. Gene family expansions and contractions for eight carnivoran species.

The species tree was built with FigTree v1.4.4 (<https://github.com/rambaut/figtree>). On each branch, the number of gene family expansions (+, green) and contractions (-, red) are represented. Counts are based on an error-corrected model of gene family evolution (analysed with CAFE).

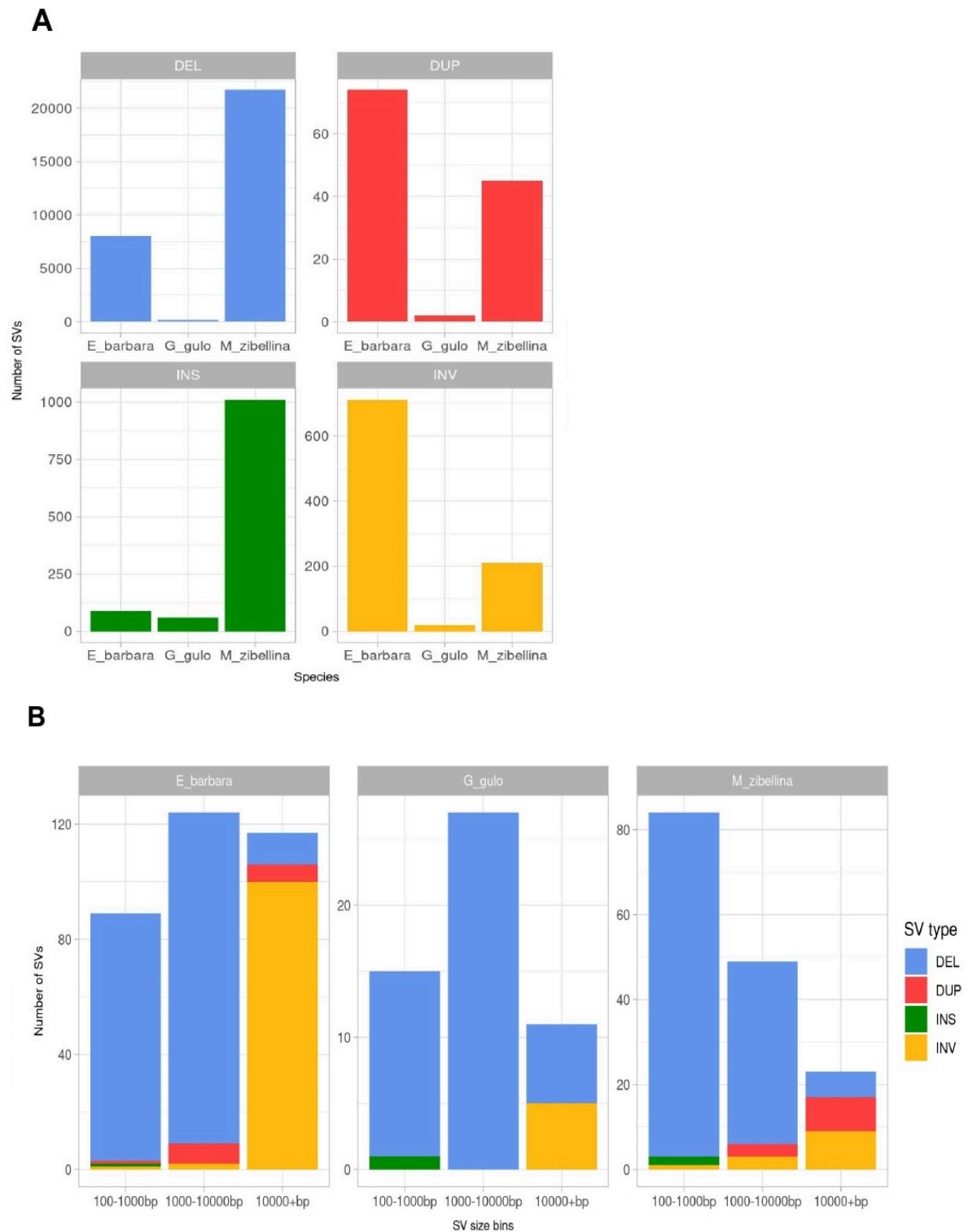


Figure S8. Structural variants detected in gulonine species.

A) Counts of different types of species-specific structural variants detected in tayra (*E_barbara*), wolverine (*G_gulo*), and sable (*M_zibellina*), **B)** Length distribution of species-specific structural variants overlapping genic regions detected in tayra (*E_barbara*), wolverine (*G_gulo*), and sable (*M_zibellina*).

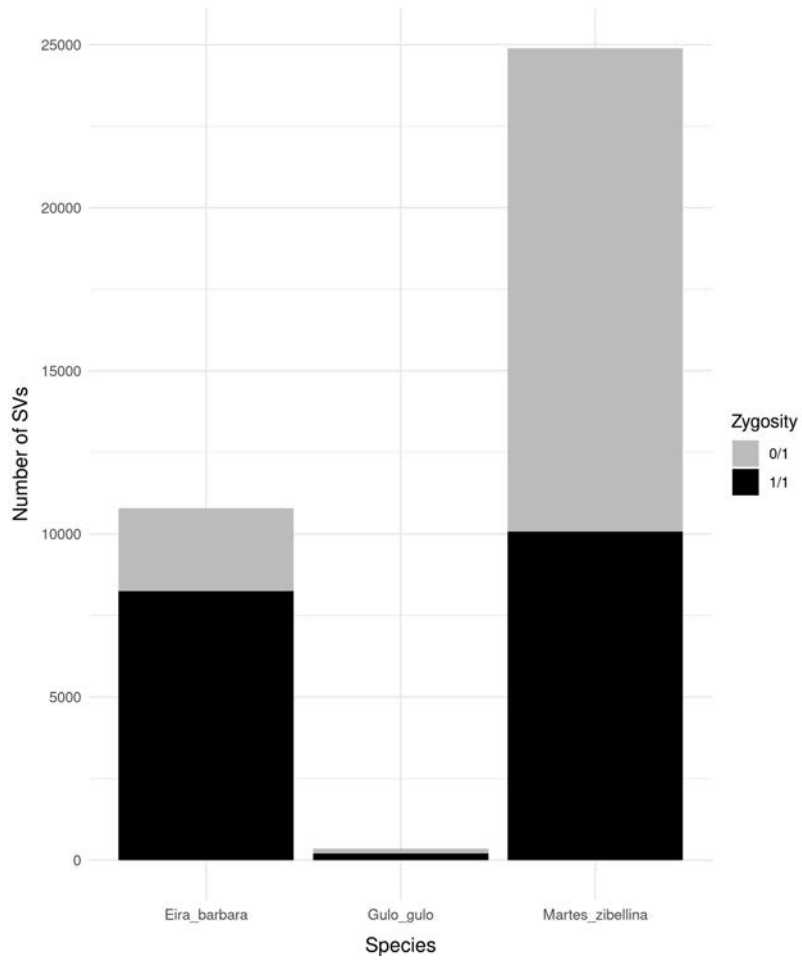


Figure S9. Heterozygous and homozygous SVs detected in each gulonine species. Heterozygous SVs represented with “0/1” (grey), homozygous SVs with “1/1” (black).

SUPPLEMENTAL INFORMATION

Genome assembly and processing of sequencing data

Due to uncertainty regarding the optimum number of reads for genome assembly, we generated three *de novo* genome assemblies with default parameters, each with a different read input: 900 million, 1.3 billion and 1.7 billion paired reads, respectively. Based on Supernova assembly metrics (contig, scaffold, and phase block N50 sizes), we chose the assembly generated using 1.3 billion paired reads for further analysis (SI Table S1).

Prior to mapping, linked-read barcodes were trimmed from tayra sequencing reads with `proc10xG` (`filter_Reads.py`, <https://github.com/ucdavis-bioinformatics/proc10xSC>). For the other two species, we used the reads used for generating the respective genome assemblies of Scandinavian wolverine (*Gulo gulo*, GenBank ID: GCA_900006375.2, PRJEB10674), and sable (*Martes zibellina*, GCA_012583365.1, PRJNA495455).

The sequencing data used to generate wolverine reference genome consist of Illumina PE short reads (library insert size range: ~200 - 500 bp), and mate-pair libraries: 3,000 - 4,500 bp. The muscle sample was obtained from a carcass of a female wolverine, originating from Jämtland County, Western Sweden, which represents the middle point of Scandinavian wolverine distribution (Ekblom et al., 2018). The sable reference genome was generated from Illumina PE short-read libraries (230 bp, 500 bp) and mate-pair libraries with different insert sizes (2 kb, 5 kb, 10 kb, and 15 kb), prepared from the sample of the individual from the Greater Khingan mountains, Heilongjiang province, Northeast China (Liu et al., 2020). This region represents the SE border of sable distribution range (Zhang et al., 2017).

Adapter clipping and quality trimming (Q30, min. length 80 bp) were performed on all samples with `TrimGalore v0.6.4` (Krueger et al., 2021). For the structural variation analysis, reads were mapped to the reference ferret genome (*Mustela putorius furo*) genome (MusPutFur1.0_HiC; (Dudchenko et al., 2017, 2018; Peng et al., 2014) in local mode with `Bowtie2 v2.3.5.1` (Langmead & Salzberg, 2012). The proportion of reads mapping to the ferret genome for all three samples was above 96%. Duplicated reads were removed with `Picard Toolkit v2.23` (MarkDuplicates, Broad Institute 2019). Insert size distributions for each sample library were generated with `Svtyper v0.7.1` (Chiang et al., 2015).

Alignment to pseudochromosome assemblies and sex verification

Trimmed reads were aligned to the pseudochromosome assemblies of corresponding species using `BWA version 0.7.17` (Li & Durbin, 2009). Read duplicates were marked using the `markup` utility from `Samtools version 1.10` (Li, 2011). To verify the sex of the tayra, sable and wolverine genomes, two approaches were applied: marker-based and coverage-based. For the marker approach, we used the Y-chromosome-specific SRY gene (sex determining region Y). The amino acid sequence of the *Martes melampus* SRY (GenBank ID: BAJ05096.1; (Yamada & Masuda, 2010) was downloaded from the NCBI protein database and aligned to the pseudochromosome assemblies using `Exonerate v 2.2.4` (Slater & Birney, 2005) in `protein2genome` mode to identify orthologous sequences. For the second approach, a per-base genome coverage was estimated using `Bedtools v2.29` (Quinlan & Hall, 2010). Mean and median values for both non-overlapping sliding windows of 1Mbp and whole genomes were calculated and visualized using scripts from the `MACE` package (<https://github.com/mahajrod/MACE>).

Phylogenomic data preparation, analysis and dating

We performed sequence alignment of 6020 coding genomic regions of single-copy orthologs shared across eight carnivoran taxa. Our taxon set included domestic cat (*Felis catus*), domestic dog (*Canis familiaris*), northern elephant seal (*Mirounga angustirostris*), and walrus (*Odobenus rosmarus*), in addition to our four mustelid species. Each genomic region was first filtered for highly divergent segments of sequences using the option trimNonHomologousFragments in MACSE v2 (Ranwez et al., 2011). This was followed by coding-region-aware sequence alignment using the alignSequences option. Sequences that were highly divergent and suspected of suffering from problematic alignment or sequencing were also identified using TreeShrink (Mai & Mirarab, 2018), based on a phylogenetic gene tree estimated using IQ-TREE v2 (Minh et al., 2020). This was followed by a second round of alignment excluding the flagged sequences. We also minimized alignment error by excluding codons with >50% of missing data, and with heterozygosity in the translated amino-acid sequences above >50%.

Possible sources of biased phylogenetic inferences due to mis-specification of the substitution model were minimized by assessing model adequacy for each sequence alignment. We performed assessment of model adequacy using methods based on simulations (Duchêne et al., 2018a) and divergence matrices (Naser-Khdour et al., 2019) as implemented in PhyloMAd (Duchêne et al., 2018b) and IQ-TREE v2. Sequence alignments were retained if they passed all tests of substitution model adequacy and retained at least four taxa and 100 nucleotides. Gene trees were estimated from the acceptable gene regions by first selecting the best substitution model from the GTR+F+ Γ +I+R family (Kalyaanamoorthy et al., 2017) and calculating approximate likelihood-ratio test (aLRT) branch supports (Anisimova & Gascuel, 2006), as implemented in IQ-TREE v2. Sequence alignment, cleaning and model adequacy assessment led to a phylogenomic data set with 2457 gene regions comprising over 3.2 million nucleotide sites.

Species tree estimates were performed using two methods. First, we concatenated sequence alignments and performed an analysis assuming that differences between gene trees are caused exclusively by stochastic error arising from having a finite sample of sites. Analysis of the concatenated data set was performed by taking the best model selected per gene region, and a model where branch lengths can vary among gene trees but maintain their relative lengths among branches (Duchêne et al., 2020). Second, we performed species tree inference under the multi-species coalescent using ASTRAL-III (Zhang et al., 2018), assuming that gene tree discordance arises from ancestral population-level processes. Gene trees were used as input for this analysis, collapsing into polytomies all aLRT branch supports below 50 prior to analysis to minimize the impact of stochastic error in gene trees. The estimate of the species tree was accompanied by local posterior probabilities as metrics of branch support.

Concordance factors were calculated using IQ-TREE v2 to explore the decisiveness of the phylogenomic signals found across gene trees (gCF) and alignment sites (sCF) (Minh et al., 2020). In addition to concordance factors, the values of the two discordance factors for each branch provide information about the relative contribution of stochastic error or more complex evolutionary processes to the signal, such as introgression. Highly uneven discordance factors are an indication of introgression, while discordance factors that are very similar to their concordance factors are indicative of substantial phylogenetic error or incomplete lineage sorting (Huson et al., 2005). We estimated gCFs using the gene trees and sCFs using the concatenated alignment, and repeated the analysis in the case that multiple phylogenetic resolutions were identified using IQ-TREE and ASTRAL-III.

The reconstructed species tree from ASTRAL-III was used as input for a Bayesian molecular dating analysis. We used highly efficient Bayesian dating using approximate likelihood computation (Thorne et al., 1998) as implemented in MCMCtree in PAML v4.8 (Yang, 2007). To reduce violation

of the most common tree priors for molecular dating (Angelis & Dos Reis, 2015) and the impact of gene tree discordance on substitution rate estimates (Mendes & Hahn, 2016), we only included genomic regions with gene trees concordant with the species tree, in addition to other forms of filtering. Acceptable loci were partitioned by codon position, each modelled under a GTR+ Γ substitution model. We used an uncorrelated gamma prior on rates across lineages and a birth-death prior for divergence times. The posterior distribution was sampled using Markov chain Monte Carlo (MCMC) every 1×10^3 steps over 1×10^7 steps, after a burn-in phase of 1×10^6 steps. We verified convergence to the stationary distribution by comparing the results from two independent runs, and confirming that the effective sample sizes for all parameters were above 1000 using the R package coda (Plummer et al., 2006).

Absolute times of divergence were estimated using a time-calibration strategy focusing on the deepest nodes of the tree (Duchêne et al., 2014). Four fossil calibrations were included, all based on nodes with consistently strong phylogenetic support in previous studies and using routinely-used fossils with robust taxonomic placement (e.g. Law et al., 2018). The split between Caniformes and Feliformes was calibrated using a conservative range for the age of the oldest known fossil of family Viverravidae (65 – 50 Mya; Benton et al., 2015; Fox et al., 2010; Wang & Tedford, 2008). The split between Canidae and other Caniformes was calibrated using a fossil of the genus *Amphicticeps* (32.8 – 30.4 Mya; Wang et al., 2005). The split between Pinnipedia and Musteloidea was calibrated using fossils of the stem musteloid of the genus *Mustelictis* (32.8 – 23.3 Mya; Rybczynski et al., 2009; Wang et al., 2005). To maximize the quality of our estimates of molecular rates and dates, we also calibrated the timing of the split between the pinnipeds *Odobenus rosmarus* and *Mirounga angustirostris*, using a fossil of the genus *Proneotherium* (20.4 – 13.8 Mya; Deméré et al., 2003). All fossil calibrations were set to follow a uniform distribution with soft maximum bounds.

Nucleotide diversity

For the sable, analysis of the heterozygosity distribution along chromosomes revealed that the ends of many pseudochromosomes had nearly double the heterozygosity than the median (> 2.5 SNPs per kbp vs median of 1.44 SNPs per kbp; Figure 2B, dark red color), while at least eight of the pseudochromosomes had very long stretches of low diversity (Figure 2B, blue color) that might be considered as runs of homozygosity (ROHs). Such distributed ROHs are often interpreted to be a consequence of recent inbreeding (Ceballos et al., 2018). This sable was sampled in the Greater Khingan mountains (Heilongjiang Province, China), very close to the edge of the species' range (Liu et al., 2020). The usual decline in abundance at the periphery, coupled with partial isolation from the main part of the species' range, might have resulted in recent inbreeding, which could explain the observed ROHs. Such a positive relationship between intraspecific SV counts and census (population) size has also been demonstrated for other species (Weissensteiner et al., 2020).

The pseudochromosome X (*ps_chrX*) of sable, had a uniform full (relative) coverage (0.75x - 1.25x relative to median whole genome coverage, SI Figure S5B), similar to the wolverine individual (SI Figure S5C). In contrast, we observed a clearly visible pseudoautosomal region (PAR) in the male tayra individual (SI Figure S5A), which had full (relative) coverage, whereas the rest of *ps_chrX* had half coverage. We detected the pseudoautosomal region (PAR) in the pseudochromosome assembly of tayra using the coverage based method described in Totikov et al., 2021. The detected PAR region encompassed the interval 90000-6660000 of *ps_chrX* (HiC_scaffold_10_RaGOO). Genetic variants of tayra in *ps_chrX* outside of PAR were called in haploid mode. Additional verification using the SRY protein gene sequence as a Y-specific marker confirmed this result. The orthologous full-length CDS of SRY was detected only in the tayra assembly, while only partial hits with low similarity were observed in the two other assemblies. This provides evidence that the sable individual is female, not male, not XXY (Klinefelter syndrome in humans; Wikström & Dunkel, 2011), nor XX with

translocation of the SRY locus to the X chromosome (de la Chapelle syndrome in humans; De la Chapelle et al., 1972).

Gene copies completeness assessment

The orthologs are listed as complete, whether single-copy or duplicated, if the BUSCO matches have scored within the expected range of scores and length alignments to the BUSCO profile. If the BUSCO matches of orthologs have scored within the range of scores but not within the range of length alignments to the BUSCO profile, then they are noted as fragmented and not considered to be duplicated. Presence of premature stop codon in the nucleotide sequence would qualify the ortholog as fragmented. Fragmented results go in a second round of sequence searches and gene predictions with parameters trained on those BUSCOs that were found to be complete, but this can still fail to recover the whole gene if present in the assembly (Simão et al., 2015). Due to this uncertainty, the focus of our study was the analysis of candidate genes sourced from the BUSCO single-copy and multi-copy gene sets qualified as complete.

Nonsynonymous and synonymous substitution rates (denoted as Ka and Ks, respectively) and Ka/Ks ratios were estimated using the software KaKs_Calculator v2.0 (Zhang et al., 2006). Maximum likelihood method of model selection and model averaging was based on a set of 14 candidate substitution models defined by Posada, 2003. Model selection was done with MODELTEST, coupled with PAUP* (in-built in KaKs_Calculator) to find the best-fit model of nucleotide substitution for pairwise sequence alignments, following the Akaike information criterion (AIC) to measure fitness between models and data. After the best-fit model selection, parameters are averaged across the candidate models to include as many features as needed to reflect the "true" model, which is seldom the one of the candidate models in practice (Posada & Buckley, 2004). Fisher's exact test is applied to determine if there was a significant association between the Ka and Ks substitution rates.

Structural variants

Prior to SV calling, tayra reads were downsampled to ~38x using seqtk v1.3 (<https://github.com/lh3/seqtk>) to match the total coverage of wolverine and sable libraries so as to avoid bias in variant calling. Three SV callers, Manta v1.6.0 (Chen et al., 2016), Whamg v1.7.0 (Kronenberg et al., 2015) and Lumpy v0.2.13 (Layer et al., 2014) were selected based on their sensitivity and precision (Cameron et al., 2019), and used to identify putative SV events in the three Guloninae genomes. Manta combines paired-read (PR) and split-read (SR) evidence during SV discovery, along with SV breakend assembly (AS) to base-pair resolution. Whamg implements PR and SR support and Lumpy, a probabilistic CNV discovery tool, uses a combined approach of PR, SR and read-depth (RD) for SV detection. Intersecting results of multiple SV callers utilizing different methods for SV detection has previously been shown to improve accuracy of variant call sets (Kosugi et al., 2019; Pirooznia et al., 2015) although this approach is highly sensitive to the chosen set of callers (Cameron et al., 2019).

High coverage joint Manta variant calls (depth greater than 3x the median chromosome depth near one or both SV breakends) mainly caused by reads mapping in low complexity regions were filtered out as well as reads with MAPQ<30 for SV breakpoint support. We retained variant calls for which all samples passed all sample-level filters (FILTER=PASS), filtered out calls with genotype quality below 30 (GQ<30), and kept calls with paired-read (PR) and split-read (SR) support of PR>=3 and SR>=3. Unlocalized and unplaced scaffolds were removed and only scaffolds assigned to chromosomes were included in further analysis.

Whamg SV calls of size <50bp and >2Mb were filtered out to improve call accuracy, as well as calls with fewer than 10 supporting reads (total support, INFO field “A”). Calls with GQ<30 were filtered out. To reduce the number of false positive calls, we filtered out events flagged as “BND” with high cross-chromosomal mappings (CW>0.2), as Whamg is aware of but does not specifically call translocations. We further removed calls of all SV types with max(CW)<0.2 as these calls are associated with poorly mapped regions. We filtered out calls for which total evidence (paired and/or split reads) supporting the variant (FORMAT/SU field) was below 10 (SU<10).

Both Whamg and Lumpy SV call sets were genotyped with Svtlyper v0.7.1 (Chiang et al., 2015) prior to filtering for genotype quality. Survivor v1.0.7 (Jeffares et al., 2017) was used to merge and compare SV call sets obtained from the three SV caller methods within and among samples. For each species and library insert size, we first merged SV events of the same type, called by at least two SV callers, with start/end positions detected within +/-1000 bp. We then intersected SV calls among the three species to obtain SVs private for each species (species-specific) and shared among all three species. To filter out SV calls overlapping gaps in the reference genome assemblies, we removed all SV calls overlapping stretches of “Ns” longer than 30 bases.

To annotate the final SV set, we first mapped annotations in gff3 format with LiftOff v1.5.1 (Shumate & Salzberg, 2020) between the domestic ferret draft genome assembly (MusPutFur1.0, GCF_000215625.1) and the chromosome-length domestic ferret assembly (DNA Zoo), which were used as a reference. We then reordered parent and children features with gff3sort (Zhu et al., 2017). Ensembl Variant Effect Predictor v101.0 (McLaren et al., 2016) was used to annotate variants affecting protein-coding genes with the maximum SV size set to 200 Mb. Annotated SV sets were further filtered for species-specific variants and SV type. Functional and biological roles of genes affected by SVs were explored using literature sources and online databases, including OrthoDB v10 (Kriventseva et al., 2019), Uniprot (The UniProt Consortium, 2017), and NCBI Entrez Gene (Maglott et al., 2011). Gene ontology analysis was performed with Shiny GO (Ge et al., 2020) with an FDR < 0.05 for each SV (excluding inversions) overlapping multiple protein-coding genes.

REFERENCES

- Angelis, K., & Dos Reis, M. (2015). The impact of ancestral population size and incomplete lineage sorting on Bayesian estimation of species divergence times. *Current Zoology*, 61(5), 874–885.
- Anisimova, M., & Gascuel, O. (2006). Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic Biology*, 55(4), 539–552.
- Benton, M. J., Donoghue, P. C. J., Vinther, J., Asher, R. J., Friedman, M., & Near, T. J. (2015). Constraints on the timescale of animal evolutionary history. *Palaeontologia Electronica*. <https://doi.org/10.26879/424>
- Cameron, D. L., Di Stefano, L., & Papenfuss, A. T. (2019). Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nature Communications*, 10(1), 3240.
- Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M., & Wilson, J. F. (2018). Runs of homozygosity: windows into population history and trait architecture. *Nature Reviews. Genetics*, 19(4), 220–234.
- Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A. J., Kruglyak, S., & Saunders, C. T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, 32(8), 1220–1222.
- Chiang, C., Layer, R. M., Faust, G. G., Lindberg, M. R., Rose, D. B., Garrison, E. P., Marth, G. T., Quinlan, A. R., & Hall, I. M. (2015). SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nature Methods*, 12(10), 966–968.
- De la Chapelle, A., Schröder, J., & Pernu, M. (1972). Isochromosome for the short arm of X, a human

- 46, XXpi syndrome. *Annals of Human Genetics*, 36(1), 79–87.
- Deméré, T. A., Berta, A., & Adam, P. J. (2003). Chapter 3. *Bulletin of the American Museum of Natural History*, 2003(279), 32–76.
- Duchêne, D. A., Duchêne, S., & Ho, S. Y. W. (2018a). Differences in Performance among Test Statistics for Assessing Phylogenomic Model Adequacy. *Genome Biology and Evolution*, 10(6), 1375–1388.
- Duchêne, D. A., Duchêne, S., & Ho, S. Y. W. (2018b). PhyloMAd: efficient assessment of phylogenomic model adequacy. *Bioinformatics*, 34(13), 2300–2301.
- Duchêne, D. A., Tong, K. J., Foster, C. S. P., Duchêne, S., Lanfear, R., & Ho, S. Y. W. (2020). Linking Branch Lengths across Sets of Loci Provides the Highest Statistical Support for Phylogenetic Inference. *Molecular Biology and Evolution*, 37(4), 1202–1210.
- Duchêne, S., Lanfear, R., & Ho, S. Y. W. (2014). The impact of calibration and clock-model choice on molecular estimates of divergence times. *Molecular Phylogenetics and Evolution*, 78, 277–289.
- 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.
- Dudchenko, O., Shamim, M. S., Batra, S. S., Durand, N. C., Musial, N. T., Mostofa, R., Pham, M., St Hilaire, B. G., Yao, W., Stamenova, E., Hoeger, M., Nyquist, S. K., Korchina, V., Pletch, K., Flanagan, J. P., Tomaszewicz, A., McAloose, D., Estrada, C. P., Novak, B. J., ... Aiden, E. L. (2018). *The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under \$1000* (p. 254797). <https://doi.org/10.1101/254797>
- Ekblom, R., Brechlin, B., Persson, J., Smeds, L., Johansson, M., Magnusson, J., Flagstad, Ø., & Ellegren, H. (2018). Genome sequencing and conservation genomics in the Scandinavian wolverine population. *Conservation Biology: The Journal of the Society for Conservation Biology*, 32(6), 1301–1312.
- Fox, R. C., Scott, C. S., & Rankin, B. D. (2010). New early carnivoran specimens from the Puercan (Earliest Paleocene) of Saskatchewan, Canada. *Journal of Paleontology*, 84(6), 1035–1039.
- Ge, S. X., Jung, D., & Yao, R. (2020). ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics*, 36(8), 2628–2629.
- Huson, D. H., Klöpper, T., Lockhart, P. J., & Steel, M. A. (2005). Reconstruction of Reticulate Networks from Gene Trees. *Research in Computational Molecular Biology*, 233–249.
- Jeffares, D. C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., Balloux, F., Dessimoz, C., Bähler, J., & Sedlazeck, F. J. (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature Communications*, 8, 14061.
- Kalyanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermini, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, 14(6), 587–589.
- Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., & Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology*, 20(1), 117.
- Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F. A., & Zdobnov, E. M. (2019). OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research*, 47(D1), D807–D811.
- Kronenberg, Z. N., Osborne, E. J., Cone, K. R., Kennedy, B. J., Domyan, E. T., Shapiro, M. D., Elde, N. C., & Yandell, M. (2015). Wham: Identifying Structural Variants of Biological Consequence. *PLoS Computational Biology*, 11(12), e1004572.
- Krueger, F., James, F., Ewels, P., Afyounian, E., & Schuster-Boeckler, B. (2021). *FelixKrueger/TrimGalore: v0.6.7 - DOI via Zenodo*. <https://doi.org/10.5281/zenodo.5127899>
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), 357–359.
- Law, C. J., Slater, G. J., & Mehta, R. S. (2018). Lineage Diversity and Size Disparity in Musteloidea:

- Testing Patterns of Adaptive Radiation Using Molecular and Fossil-Based Methods. *Systematic Biology*, 67(1), 127–144.
- Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: a probabilistic framework for structural variant discovery. *Genome Biology*, 15(6), R84.
- Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21), 2987–2993.
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), 1754–1760.
- Liu, G., Zhao, C., Xu, D., Zhang, H., Monakhov, V., Shang, S., Gao, X., Sha, W., Ma, J., Zhang, W., Tang, X., Li, B., Hua, Y., Cao, X., Liu, Z., & Zhang, H. (2020). First Draft Genome of the Sable, *Martes zibellina*. *Genome Biology and Evolution*, 12(3), 59–65.
- Maglott, D., Ostell, J., Pruitt, K. D., & Tatusova, T. (2011). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, 39(Database issue), D52–D57.
- Mai, U., & Mirarab, S. (2018). TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genomics*, 19(Suppl 5), 272.
- McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1), 122.
- Mendes, F. K., & Hahn, M. W. (2016). Gene Tree Discordance Causes Apparent Substitution Rate Variation. *Systematic Biology*, 65(4), 711–721.
- Minh, B. Q., Hahn, M. W., & Lanfear, R. (2020). New Methods to Calculate Concordance Factors for Phylogenomic Datasets. *Molecular Biology and Evolution*, 37(9), 2727–2733.
- Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, 37(5), 1530–1534.
- Naser-Khdour, S., Minh, B. Q., Zhang, W., Stone, E. A., & Lanfear, R. (2019). The Prevalence and Impact of Model Violations in Phylogenetic Analysis. *Genome Biology and Evolution*, 11(12), 3341–3352.
- Peng, X., Alföldi, J., Gori, K., Einfeld, A. J., Tyler, S. R., Tisoncik-Go, J., Brawand, D., Law, G. L., Skunca, N., Hatta, M., Gasper, D. J., Kelly, S. M., Chang, J., Thomas, M. J., Johnson, J., Berlin, A. M., Lara, M., Russell, P., Swofford, R., ... Katze, M. G. (2014). The draft genome sequence of the ferret (*Mustela putorius furo*) facilitates study of human respiratory disease. *Nature Biotechnology*, 32(12), 1250–1255.
- Pirooznia, M., Goes, F. S., & Zandi, P. P. (2015). Whole-genome CNV analysis: advances in computational approaches. *Frontiers in Genetics*, 6, 138.
- Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: convergence diagnosis and output analysis for MCMC. *R News*, 6(1), 7–11.
- Posada, D. (2003). Using MODELTEST and PAUP* to select a model of nucleotide substitution. *Current Protocols in Bioinformatics / Editorial Board, Andreas D. Baxevanis ... [et Al.]*, Chapter 6, Unit 6.5.
- Posada, D., & Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic Biology*, 53(5), 793–808.
- Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841–842.
- Ranwez, V., Harispe, S., Delsuc, F., & Douzery, E. J. P. (2011). MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. *PloS One*, 6(9), e22594.
- Rybczynski, N., Dawson, M. R., & Tedford, R. H. (2009). A semi-aquatic Arctic mammalian carnivore from the Miocene epoch and origin of Pinnipedia. *Nature*, 458(7241), 1021–1024.
- Shumate, A., & Salzberg, S. L. (2020). *Liftoff: an accurate gene annotation mapping tool* (p. 2020.06.24.169680). <https://doi.org/10.1101/2020.06.24.169680>
- Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.

- Bioinformatics*, 31(19), 3210–3212.
- Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6, 31.
- The UniProt Consortium. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1), D158–D169.
- Thorne, J. L., Kishino, H., & Painter, I. S. (1998). Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution*, 15(12), 1647–1657.
- Totikov, A., Tomarovsky, A., Prokopov, D., Yakupova, A., Bulyonkova, T., Derezanin, L., Rasskazov, D., Wolfsberger, W. W., Koepfli, K.-P., Oleksyk, T. K., & Kliver, S. (2021). Chromosome-Level Genome Assemblies Expand Capabilities of Genomics for Conservation Biology. *Genes*, 12(9), 1336.
- Wang, X., McKenna, M. C., & Dashzeveg, D. (2005). Amphicticeps and Amphicynodon (Arctoidea, Carnivora) from Hsanda Gol Formation, Central Mongolia and Phylogeny of Basal Arctoids with Comments on Zoogeography. *American Museum Novitates*, 2005(3483), 1–60.
- Wang, X., & Tedford, R. H. (2008). *Dogs: Their Fossil Relatives and Evolutionary History*. Columbia University Press.
- Weissensteiner, M. H., Bunikis, I., Catalán, A., Francoijs, K.-J., Knief, U., Heim, W., Peona, V., Pophaly, S. D., Sedlazeck, F. J., Suh, A., Warmuth, V. M., & Wolf, J. B. W. (2020). Discovery and population genomics of structural variation in a songbird genus. *Nature Communications*, 11(1), 3403.
- Wikström, A. M., & Dunkel, L. (2011). Klinefelter syndrome. *Best Practice & Research. Clinical Endocrinology & Metabolism*, 25(2), 239–250.
- Yamada, C., & Masuda, R. (2010). Molecular Phylogeny and Evolution of Sex-Chromosomal Genes and SINE Sequences in the Family Mustelidae. *Mammal Study*, 35(1), 17–30.
- Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24(8), 1586–1591.
- Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, 19(Suppl 6), 153.
- Zhang, R., Yang, L., Ai, L., Yang, Q., Chen, M., Li, J., Yang, L., & Luan, X. (2017). Geographic characteristics of sable (*Martes zibellina*) distribution over time in Northeast China. *Ecology and Evolution*, 7(11), 4016–4023.
- Zhang, Z., Li, J., Zhao, X.-Q., Wang, J., Wong, G. K.-S., & Yu, J. (2006). KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics, Proteomics & Bioinformatics*, 4(4), 259–263.
- Zhu, T., Liang, C., Meng, Z., Guo, S., & Zhang, R. (2017). GFF3sort: a novel tool to sort GFF3 files for tabix indexing. *BMC Bioinformatics*, 18(1), 482.

Supplemental Tables S4-S6 - References

- Amos, J. S., Huang, L., Thevenon, J., Kariminedjad, A., Beaulieu, C. L., Masurel-Paulet, A., Najmabadi, H., Fattahi, Z., Beheshtian, M., Tonekaboni, S. H., Tang, S., Helbig, K. L., Alcaraz, W., Rivière, J.-B., Faivre, L., Innes, A. M., Lebel, R. R., Boycott, K. M., & Care4Rare Canada Consortium. (2017). Autosomal recessive mutations in THOC6 cause intellectual disability: syndrome delineation requiring forward and reverse phenotyping. *Clinical Genetics*, 91(1), 92–99.
- Avellino, R., Carrella, S., Pirozzi, M., Risolino, M., Salierno, F. G., Franco, P., Stoppelli, P., Verde, P., Banfi, S., & Conte, I. (2013). miR-204 targeting of Ankrd13A controls both mesenchymal neural crest and lens cell migration. *PLoS One*, 8(4), e61099.

- Avnet, S., Lemma, S., Errani, C., Falzetti, L., Panza, E., Columbaro, M., Nanni, C., & Baldini, N. (2020). Benign albeit glycolytic: MCT4 expression and lactate release in giant cell tumour of bone. *Bone*, *134*, 115302.
- Bähler, M., & Rhoads, A. (2002). Calmodulin signaling via the IQ motif. *FEBS Letters*, *513*(1), 107–113.
- Banci, L., Bertini, I., Ciofi-Baffoni, S., Jaiswal, D., Neri, S., Peruzzini, R., & Winkelmann, J. (2012). Structural characterization of CHCHD5 and CHCHD7: two atypical human twin CX9C proteins. *Journal of Structural Biology*, *180*(1), 190–200.
- Bao, J., Zhang, J., Zheng, H., Xu, C., & Yan, W. (2010). UBQLN1 interacts with SPEM1 and participates in spermiogenesis. *Molecular and Cellular Endocrinology*, *327*(1-2), 89–97.
- Bauer, A., De Lucia, M., Jagannathan, V., Mezzalana, G., Casal, M. L., Welle, M. M., & Leeb, T. (2017). A Large Deletion in the NSDHL Gene in Labrador Retrievers with a Congenital Cornification Disorder. *G3*, *7*(9), 3115–3121.
- Baumann, P., Schriever, S. C., Kullmann, S., Zimprich, A., Feuchtinger, A., Amarie, O., Peter, A., Walch, A., Gailus-Durner, V., Fuchs, H., Hrabě de Angelis, M., Wurst, W., Tschöp, M. H., Heni, M., Höfner, S. M., & Pfluger, P. T. (2019). Dusp8 affects hippocampal size and behavior in mice and humans. *Scientific Reports*, *9*(1), 19483.
- Belizaire, R., Komanduri, C., Wooten, K., Chen, M., Thaller, C., & Janz, R. (2004). Characterization of synaptogyrin 3 as a new synaptic vesicle protein. *The Journal of Comparative Neurology*, *470*(3), 266–281.
- Blume, M., Inoguchi, F., Sugiyama, T., Owada, Y., Osumi, N., Aimi, Y., Taki, K., & Katsuyama, Y. (2017). Dab1 contributes differently to the morphogenesis of the hippocampal subdivisions. *Development, Growth & Differentiation*, *59*(8), 657–673.
- Browning, A. C., Figueiredo, G. S., Baylis, O., Montgomery, E., Beesley, C., Molinari, E., Figueiredo, F. C., & Sayer, J. A. (2019). A case of ocular cystinosis associated with two potentially severe CTNS mutations. *Ophthalmic Genetics*, *40*(2), 157–160.
- Brown, J. A., Eberhardt, D. M., Schrick, F. N., Roberts, M. P., & Godkin, J. D. (2003). Expression of retinol-binding protein and cellular retinol-binding protein in the bovine ovary. *Molecular Reproduction and Development*, *64*(3), 261–269.
- Cabral, W. A., Ishikawa, M., Garten, M., Makareeva, E. N., Sargent, B. M., Weis, M., Barnes, A. M., Webb, E. A., Shaw, N. J., Ala-Kokko, L., Lachawan, F. L., Högler, W., Leikin, S., Blank, P. S., Zimmerberg, J., Eyre, D. R., Yamada, Y., & Marini, J. C. (2016). Absence of the ER Cation Channel TMEM38B/TRIC-B Disrupts Intracellular Calcium Homeostasis and Dysregulates Collagen Synthesis in Recessive Osteogenesis Imperfecta. *PLoS Genetics*, *12*(7), e1006156.
- Chen, D., Teng, J. M., North, P. E., Lapinski, P. E., & King, P. D. (2019). RASA1-dependent cellular export of collagen IV controls blood and lymphatic vascular development. *The Journal of Clinical Investigation*, *129*(9), 3545–3561.
- Chen, Q., Denard, B., Lee, C.-E., Han, S., Ye, J. S., & Ye, J. (2016). Inverting the Topology of a Transmembrane Protein by Regulating the Translocation of the First Transmembrane Helix. *Molecular Cell*, *63*(4), 567–578.
- Cheong, A., Lingutla, R., & Mager, J. (2020). Expression analysis of mammalian mitochondrial ribosomal protein genes. *Gene Expression Patterns: GEP*, *38*, 119147.
- Choi, Y. J., Kim, S., Choi, Y., Nielsen, T. B., Yan, J., Lu, A., Ruan, J., Lee, H.-R., Wu, H., Spellberg, B., & Jung, J. U. (2019). SERPINB1-mediated checkpoint of inflammatory caspase activation. *Nature Immunology*, *20*(3), 276–287.
- Coan, P. M., Vaughan, O. R., Sekita, Y., Finn, S. L., Burton, G. J., Constancia, M., & Fowden, A. L. (2010). Adaptations in placental phenotype support fetal growth during undernutrition of pregnant mice. *The Journal of Physiology*, *588*(Pt 3), 527–538.
- Crocco, P., Saiardi, A., Wilson, M. S., Maletta, R., Bruni, A. C., Passarino, G., & Rose, G. (2016). Contribution of polymorphic variation of inositol hexakisphosphate kinase 3 (IP6K3) gene promoter to the susceptibility to late onset Alzheimer’s disease. *Biochimica et Biophysica Acta*, *1862*(9), 1766–1773.
- D’Ambrosio, D. N., Clugston, R. D., & Blaner, W. S. (2011). Vitamin A metabolism: an update.

- Nutrients*, 3(1), 63–103.
- Davydova, E., Ho, A. Y. Y., Malecki, J., Moen, A., Enserink, J. M., Jakobsson, M. E., Loenarz, C., & Falnes, P. Ø. (2014). Identification and characterization of a novel evolutionarily conserved lysine-specific methyltransferase targeting eukaryotic translation elongation factor 2 (eEF2). *The Journal of Biological Chemistry*, 289(44), 30499–30510.
- Del Dotto, V., Fogazza, M., Musiani, F., Maresca, A., Aleo, S. J., Caporali, L., La Morgia, C., Nolli, C., Lodi, T., Goffrini, P., Chan, D., Carelli, V., Rugolo, M., Baruffini, E., & Zanna, C. (2018). Deciphering OPA1 mutations pathogenicity by combined analysis of human, mouse and yeast cell models. *Biochimica et Biophysica Acta, Molecular Basis of Disease*, 1864(10), 3496–3514.
- Dong, F., Shinohara, K., Botilde, Y., Nabeshima, R., Asai, Y., Fukumoto, A., Hasegawa, T., Matsuo, M., Takeda, H., Shiratori, H., Nakamura, T., & Hamada, H. (2014). Pih1d3 is required for cytoplasmic preassembly of axonemal dynein in mouse sperm. *The Journal of Cell Biology*, 204(2), 203–213.
- Dreiza, C. M., Komalavilas, P., Furnish, E. J., Flynn, C. R., Sheller, M. R., Smoke, C. C., Lopes, L. B., & Brophy, C. M. (2010). The small heat shock protein, HSPB6, in muscle function and disease. *Cell Stress & Chaperones*, 15(1), 1–11.
- Feng, W., & Zhang, M. (2009). Organization and dynamics of PDZ-domain-related supramodules in the postsynaptic density. *Nature Reviews. Neuroscience*, 10(2), 87–99.
- Feng, Y., Xu, X., Zhang, Y., Ding, J., Wang, Y., Zhang, X., Wu, Z., Kang, L., Liang, Y., Zhou, L., Song, S., Zhao, K., & Ye, Q. (2015). HPIP is upregulated in colorectal cancer and regulates colorectal cancer cell proliferation, apoptosis and invasion. *Scientific Reports*, 5, 9429.
- Geisinger, A., Alsheimer, M., Baier, A., Benavente, R., & Wettstein, R. (2005). The mammalian gene pecanex 1 is differentially expressed during spermatogenesis. *Biochimica et Biophysica Acta*, 1728(1-2), 34–43.
- Gerits, N., Van Belle, W., & Moens, U. (2007). Transgenic mice expressing constitutive active MAPKAPK5 display gender-dependent differences in exploration and activity. *Behavioral and Brain Functions: BBF*, 3, 58.
- Gòdia, M., Casellas, J., Ruiz-Herrera, A., Rodríguez-Gil, J. E., Castelló, A., Sánchez, A., & Clop, A. (2020). Whole genome sequencing identifies allelic ratio distortion in sperm involving genes related to spermatogenesis in a swine model. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, 27(5).
- Goldfarb, K. C., & Cech, T. R. (2017). Targeted CRISPR disruption reveals a role for RNase MRP RNA in human preribosomal RNA processing. *Genes & Development*, 31(1), 59–71.
- Grayson, P., & Civetta, A. (2012). Positive Selection and the Evolution of izumo Genes in Mammals. *International Journal of Evolutionary Biology*, 2012, 958164.
- Griffiths, M. R., Botto, M., Morgan, B. P., Neal, J. W., & Gasque, P. (2018). CD93 regulates central nervous system inflammation in two mouse models of autoimmune encephalomyelitis. *Immunology*, 155(3), 346–355.
- Guo, H. X., Cun, W., Liu, L. D., Dong, S. Z., Wang, L. C., Dong, C. H., & Li, Q. H. (2006). Protein encoded by HSV-1 stimulation-related gene 1 (HSRG1) interacts with and inhibits SV40 large T antigen. *Cell Proliferation*, 39(6), 507–518.
- He, Y.-W., Li, H., Zhang, J., Hsu, C.-L., Lin, E., Zhang, N., Guo, J., Forbush, K. A., & Bevan, M. J. (2004). The extracellular matrix protein mindin is a pattern-recognition molecule for microbial pathogens. *Nature Immunology*, 5(1), 88–97.
- Hill, M., Pařízek, A., Cibula, D., Kancheva, R., Jirásek, J. E., Jirkovská, M., Velíková, M., Kubátová, J., Klímková, M., Pašková, A., Zizka, Z., Kancheva, L., Kazihnitková, H., Zamrazilová, L., & Stárka, L. (2010). Steroid metabolome in fetal and maternal body fluids in human late pregnancy. *The Journal of Steroid Biochemistry and Molecular Biology*, 122(4), 114–132.
- Hnia, K., Tronchère, H., Tomczak, K. K., Amoasii, L., Schultz, P., Beggs, A. H., Payrastre, B., Mandel, J. L., & Laporte, J. (2011). Myotubularin controls desmin intermediate filament architecture and mitochondrial dynamics in human and mouse skeletal muscle. *The Journal of Clinical Investigation*, 121(1), 70–85.
- Horibata, Y., Elpeleg, O., Eran, A., Hirabayashi, Y., Savitzki, D., Tal, G., Mandel, H., & Sugimoto, H.

- (2018). EPT1 (selenoprotein I) is critical for the neural development and maintenance of plasmalogen in humans. *Journal of Lipid Research*, 59(6), 1015–1026.
- Im, S.-H., Kim, S.-H., Azam, T., Venkatesh, N., Dinarello, C. A., Fuchs, S., & Souroujon, M. C. (2002). Rat interleukin-18 binding protein: cloning, expression, and characterization. *Journal of Interferon & Cytokine Research: The Official Journal of the International Society for Interferon and Cytokine Research*, 22(3), 321–328.
- Inoue, K., Maeda, N., Mori, T., Sekimoto, R., Tsushima, Y., Matsuda, K., Yamaoka, M., Suganami, T., Nishizawa, H., Ogawa, Y., Funahashi, T., & Shimomura, I. (2014). Possible involvement of Opa-interacting protein 5 in adipose proliferation and obesity. *PLoS One*, 9(2), e87661.
- Ishizawa, T., Nozaki, Y., Ueda, T., & Takeuchi, N. (2008). The human mitochondrial translation release factor HMRF1L is methylated in the GGQ motif by the methyltransferase HMPPrmC. *Biochemical and Biophysical Research Communications*, 373(1), 99–103.
- Jakobsson, T., Venteclef, N., Toresson, G., Damdimopoulos, A. E., Ehlund, A., Lou, X., Sanyal, S., Steffensen, K. R., Gustafsson, J.-A., & Treuter, E. (2009). GPS2 is required for cholesterol efflux by triggering histone demethylation, LXR recruitment, and coregulator assembly at the ABCG1 locus. *Molecular Cell*, 34(4), 510–518.
- Johnson, R. J., Stenvinkel, P., Andrews, P., Sánchez-Lozada, L. G., Nakagawa, T., Gaucher, E., Andres-Hernando, A., Rodriguez-Iturbe, B., Jimenez, C. R., Garcia, G., Kang, D.-H., Tolan, D. R., & Lanaspa, M. A. (2020). Fructose metabolism as a common evolutionary pathway of survival associated with climate change, food shortage and droughts. *Journal of Internal Medicine*, 287(3), 252–262.
- Kang, C. B., Hong, Y., Dhe-Paganon, S., & Yoon, H. S. (2008). FKBP family proteins: immunophilins with versatile biological functions. *Neuro-Signals*, 16(4), 318–325.
- Katzen, J., Wagner, B. D., Venosa, A., Kopp, M., Tomer, Y., Russo, S. J., Headen, A. C., Basil, M. C., Stark, J. M., Mulugeta, S., Deterding, R. R., & Beers, M. F. (2019). An SFTPC BRICHOS mutant links epithelial ER stress and spontaneous lung fibrosis. *JCI Insight*, 4(6). <https://doi.org/10.1172/jci.insight.126125>
- Kazmierczak, M., Harris, S. L., Kazmierczak, P., Shah, P., Starovoytov, V., Ohlemiller, K. K., & Schwander, M. (2015). Progressive Hearing Loss in Mice Carrying a Mutation in Usp53. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 35(47), 15582–15598.
- Kennell, J. A., Richards, N. W., Schaner, P. E., & Gumucio, D. L. (2001). cDNA cloning, chromosomal localization and evolutionary analysis of mouse vacuolar ATPase subunit D, Atp6m. *Cytogenetics and Cell Genetics*, 92(3-4), 337–341.
- Kielkowsky, P., Buchsbaum, I. Y., Kirsch, V. C., Bach, N. C., Drukker, M., Cappello, S., & Sieber, S. A. (2020). FICD activity and AMPylation remodelling modulate human neurogenesis. *Nature Communications*, 11(1), 517.
- Kim, J. H., Gurumurthy, C. B., Band, H., & Band, V. (2010). Biochemical characterization of human Ecdysoneless reveals a role in transcriptional regulation. *Biological Chemistry*, 391(1), 9–19.
- Kim, J., Ishiguro, K.-I., Nambu, A., Akiyoshi, B., Yokobayashi, S., Kagami, A., Ishiguro, T., Pendas, A. M., Takeda, N., Sakakibara, Y., Kitajima, T. S., Tanno, Y., Sakuno, T., & Watanabe, Y. (2015). Meikin is a conserved regulator of meiosis-I-specific kinetochore function. *Nature*, 517(7535), 466–471.
- Kim, M. Y., Lee, H. K., Park, J. S., Park, S. H., Kwon, H. B., & Soh, J. (1999). Identification of a zeta-crystallin (quinone reductase)-like 1 gene (CRYZL1) mapped to human chromosome 21q22.1. *Genomics*, 57(1), 156–159.
- Kitchener, A. C., Meloro, C., & Williams, T. M. (2018). Form and function of the musteloids. In *Biology and Conservation of Musteloids*. Oxford University Press.
- Koc, E. C., Burkhart, W., Blackburn, K., Moseley, A., Koc, H., & Spremulli, L. L. (2000). A Proteomics Approach to the Identification of Mammalian Mitochondrial Small Subunit Ribosomal Proteins *. *The Journal of Biological Chemistry*, 275(42), 32585–32591.
- Kohroki, J., Nishiyama, T., Nakamura, T., & Masuho, Y. (2005). ASB proteins interact with Cullin5 and Rbx2 to form E3 ubiquitin ligase complexes. *FEBS Letters*, 579(30), 6796–6802.

- Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F. A., & Zdobnov, E. M. (2019). OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research*, *47*(D1), D807–D811.
- Kuo, P.-L., Chiang, H.-S., Wang, Y.-Y., Kuo, Y.-C., Chen, M.-F., Yu, I.-S., Teng, Y.-N., Lin, S.-W., & Lin, Y.-H. (2013). SEPT12-microtubule complexes are required for sperm head and tail formation. *International Journal of Molecular Sciences*, *14*(11), 22102–22116.
- Kuriakose, T., & Kanneganti, T.-D. (2018). ZBP1: Innate Sensor Regulating Cell Death and Inflammation. *Trends in Immunology*, *39*(2), 123–134.
- Lal-Nag, M., & Morin, P. J. (2009). The claudins. *Genome Biology*, *10*(8), 235.
- Larsen, K., Momeni, J., Farajzadeh, L., & Callesen, H. (2017). Splice variants of porcine PPHLN1 encoding periphilin-1. *Gene Reports*, *7*, 176–183.
- Lei, Y., Wen, H., Yu, Y., Taxman, D. J., Zhang, L., Widman, D. G., Swanson, K. V., Wen, K.-W., Damania, B., Moore, C. B., Giguère, P. M., Siderovski, D. P., Hiscott, J., Razani, B., Semenkovich, C. F., Chen, X., & Ting, J. P.-Y. (2012). The mitochondrial proteins NLRX1 and TUFM form a complex that regulates type I interferon and autophagy. *Immunity*, *36*(6), 933–946.
- Li, Q., Korzan, W. J., Ferrero, D. M., Chang, R. B., Roy, D. S., Buchi, M., Lemon, J. K., Kaur, A. W., Stowers, L., Fendt, M., & Liberles, S. D. (2013). Synchronous evolution of an odor biosynthesis pathway and behavioral response. *Current Biology: CB*, *23*(1), 11–20.
- Liu, S., Gong, X., Yan, X., Peng, T., Baker, J. C., Li, L., Robben, P. M., Ravindran, S., Andersson, L. A., Cole, A. B., & Roche, T. E. (2001). Reaction mechanism for mammalian pyruvate dehydrogenase using natural lipoyl domain substrates. *Archives of Biochemistry and Biophysics*, *386*(2), 123–135.
- Luczkowska, K., Stekelenburg, C., Sloan-Béna, F., Ranza, E., Gastaldi, G., Schwitzgebel, V., & Maechler, P. (2020). Hyperinsulinism associated with GLUD1 mutation: allosteric regulation and functional characterization of p.G446V glutamate dehydrogenase. *Human Genomics*, *14*(1), 9.
- Lupo, G., Nisi, P. S., Esteve, P., Paul, Y.-L., Novo, C. L., Sidders, B., Khan, M. A., Biagioni, S., Liu, H.-K., Bovolenta, P., Cacci, E., & Rugg-Gunn, P. J. (2018). Molecular profiling of aged neural progenitors identifies Dbx2 as a candidate regulator of age-associated neurogenic decline. *Aging Cell*, *17*(3), e12745.
- Maglott, D., Ostell, J., Pruitt, K. D., & Tatusova, T. (2011). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, *39*(Database issue), D52–D57.
- Makarova, O. V., Makarov, E. M., & Lührmann, R. (2001). The 65 and 110 kDa SR-related proteins of the U4/U6.U5 tri-snRNP are essential for the assembly of mature spliceosomes. *The EMBO Journal*, *20*(10), 2553–2563.
- Martinez-Duncker, I., Dupré, T., Piller, V., Piller, F., Candelier, J.-J., Trichet, C., Tchernia, G., Oriol, R., & Mollicone, R. (2005). Genetic complementation reveals a novel human congenital disorder of glycosylation of type II, due to inactivation of the Golgi CMP-sialic acid transporter. *Blood*, *105*(7), 2671–2676.
- Masure, S., Cik, M., Hoefnagel, E., Nosrat, C. A., Van der Linden, I., Scott, R., Van Gompel, P., Lesage, A. S., Verhasselt, P., Ibáñez, C. F., & Gordon, R. D. (2000). Mammalian GFRalpha -4, a divergent member of the GFRalpha family of coreceptors for glial cell line-derived neurotrophic factor family ligands, is a receptor for the neurotrophic factor persephin. *The Journal of Biological Chemistry*, *275*(50), 39427–39434.
- Matsubara, T., Kokabu, S., Nakatomi, C., Kinbara, M., Maeda, T., Yoshizawa, M., Yasuda, H., Takano-Yamamoto, T., Baron, R., & Jimi, E. (2018). The Actin-Binding Protein PPP1r18 Regulates Maturation, Actin Organization, and Bone Resorption Activity of Osteoclasts. *Molecular and Cellular Biology*, *38*(4).
- McDaniel, P., & Wu, X. (2009). Identification of oocyte-selective NLRP genes in rhesus macaque monkeys (*Macaca mulatta*). *Molecular Reproduction and Development*, *76*(2), 151–159.
- Metzger, J., Karwath, M., Tonda, R., Beltran, S., Águeda, L., Gut, M., Gut, I. G., & Distl, O. (2015). Runs of homozygosity reveal signatures of positive selection for reproduction traits in breed and non-breed horses. *BMC Genomics*, *16*, 764.

- Mironova, E., & Millette, C. F. (2008). Expression of the diaphanous-related formin proteins mDia1 and mDia2 in the rat testis. *Developmental Dynamics: An Official Publication of the American Association of Anatomists*, 237(8), 2170–2176.
- Mizuarai, S., Miki, S., Araki, H., Takahashi, K., & Kotani, H. (2005). Identification of dicarboxylate carrier Slc25a10 as malate transporter in de novo fatty acid synthesis. *The Journal of Biological Chemistry*, 280(37), 32434–32441.
- Montanez, E., Wickström, S. A., Altstätter, J., Chu, H., & Fässler, R. (2009). Alpha-parvin controls vascular mural cell recruitment to vessel wall by regulating RhoA/ROCK signalling. *The EMBO Journal*, 28(20), 3132–3144.
- Nakamura, S., Kahyo, T., Tao, H., Shibata, K., Kurabe, N., Yamada, H., Shinmura, K., Ohnishi, K., & Sugimura, H. (2015). Novel roles for LIX1L in promoting cancer cell proliferation through ROS1-mediated LIX1L phosphorylation. *Scientific Reports*, 5, 13474.
- Nelson, L., Anderson, S., Archibald, A. L., Rhind, S., Lu, Z. H., Condie, A., McIntyre, N., Thompson, J., Nenutil, R., Vojtesek, B., Whitelaw, C. B. A., Little, T. J., & Hupp, T. (2008). An animal model to evaluate the function and regulation of the adaptively evolving stress protein SEP53 in oesophageal bile damage responses. *Cell Stress & Chaperones*, 13(3), 375–385.
- Nomura, Y., Roston, D., Montemayor, E. J., Cui, Q., & Butcher, S. E. (2018). Structural and mechanistic basis for preferential deadenylation of U6 snRNA by Usb1. *Nucleic Acids Research*, 46(21), 11488–11501.
- Papadopoulos, C., Kirchner, P., Bug, M., Grum, D., Koerver, L., Schulze, N., Poehler, R., Dressler, A., Fengler, S., Arhzaouy, K., Lux, V., Ehrmann, M., Wehl, C. C., & Meyer, H. (2017). VCP/p97 cooperates with YOD1, UBXD1 and PLAA to drive clearance of ruptured lysosomes by autophagy. *The EMBO Journal*, 36(2), 135–150.
- Park, K. M., Kang, E., Jeon, Y.-J., Kim, N., Kim, N.-S., Yoo, H.-S., Yeom, Y. I., & Kim, S. J. (2007). Identification of novel regulators of apoptosis using a high-throughput cell-based screen. *Molecules and Cells*, 23(2), 170–174.
- Paschen, S. A., Rothbauer, U., Káldi, K., Bauer, M. F., Neupert, W., & Brunner, M. (2000). The role of the TIM8-13 complex in the import of Tim23 into mitochondria. *The EMBO Journal*, 19(23), 6392–6400.
- Pei, J., & Grishin, N. V. (2012). Unexpected diversity in Shisa-like proteins suggests the importance of their roles as transmembrane adaptors. *Cellular Signalling*, 24(3), 758–769.
- Perumal, K., Sinha, K., Henning, D., & Reddy, R. (2001). Purification, characterization, and cloning of the cDNA of human signal recognition particle RNA 3'-adenylating enzyme. *The Journal of Biological Chemistry*, 276(24), 21791–21796.
- Piccolo, A., & Pusch, M. (2005). Chloride/proton antiporter activity of mammalian CLC proteins CLC-4 and CLC-5. *Nature*, 436(7049), 420–423.
- Pierre, K., Parent, A., Jayet, P.-Y., Halestrap, A. P., Scherrer, U., & Pellerin, L. (2007). Enhanced expression of three monocarboxylate transporter isoforms in the brain of obese mice. *The Journal of Physiology*, 583(Pt 2), 469–486.
- Premzl, M. (2016). Comparative genomic analysis of eutherian tumor necrosis factor ligand genes. *Immunogenetics*, 68(2), 125–132.
- Pu, J., Schindler, C., Jia, R., Jarnik, M., Backlund, P., & Bonifacino, J. S. (2015). BORC, a multisubunit complex that regulates lysosome positioning. *Developmental Cell*, 33(2), 176–188.
- Qureshi, T., Bjørkmo, M., Nordengen, K., Gundersen, V., Utheim, T. P., Watne, L. O., Storm-Mathisen, J., Hassel, B., & Chaudhry, F. A. (2020). Slc38a1 Conveys Astroglia-Derived Glutamine into GABAergic Interneurons for Neurotransmitter GABA Synthesis. *Cells*, 9(7). <https://doi.org/10.3390/cells9071686>
- Raffaello, A., De Stefani, D., Sabbadin, D., Teardo, E., Merli, G., Picard, A., Checchetto, V., Moro, S., Szabò, I., & Rizzuto, R. (2013). The mitochondrial calcium uniporter is a multimer that can include a dominant-negative pore-forming subunit. *The EMBO Journal*, 32(17), 2362–2376.
- Rauniyar, K., Jha, S. K., & Jeltsch, M. (2018). Biology of Vascular Endothelial Growth Factor C in the Morphogenesis of Lymphatic Vessels. *Frontiers in Bioengineering and Biotechnology*, 6, 7.
- Reynolds, A., Qiao, H., Yang, Y., Chen, J. K., Jackson, N., Biswas, K., Holloway, J. K., Baudat, F., de

- Massy, B., Wang, J., Höög, C., Cohen, P. E., & Hunter, N. (2013). RNF212 is a dosage-sensitive regulator of crossing-over during mammalian meiosis. *Nature Genetics*, *45*(3), 269–278.
- Rismanchi, N., Soderblom, C., Stadler, J., Zhu, P.-P., & Blackstone, C. (2008). Atlantin GTPases are required for Golgi apparatus and ER morphogenesis. *Human Molecular Genetics*, *17*(11), 1591–1604.
- Rivas, M. A., Graham, D., Sulem, P., Stevens, C., Desch, A. N., Goyette, P., Gudbjartsson, D., Jonsdottir, I., Thorsteinsdottir, U., Degenhardt, F., Mucha, S., Kurki, M. I., Li, D., D’Amato, M., Annese, V., Vermeire, S., Weersma, R. K., Halfvarson, J., Paavola-Sakki, P., ... Wang, M. H. (2016). A protein-truncating R179X variant in RNF186 confers protection against ulcerative colitis. *Nature Communications*, *7*, 12342.
- Salleron, L., Magistrelli, G., Mary, C., Fischer, N., Bairoch, A., & Lane, L. (2014). DERA is the human deoxyribose phosphate aldolase and is involved in stress response. *Biochimica et Biophysica Acta*, *1843*(12), 2913–2925.
- Schlager, M. A., Kapitein, L. C., Grigoriev, I., Burzynski, G. M., Wulf, P. S., Keijzer, N., de Graaff, E., Fukuda, M., Shepherd, I. T., Akhmanova, A., & Hoogenraad, C. C. (2010). Pericentrosomal targeting of Rab6 secretory vesicles by Bicaudal-D-related protein 1 (BICDR-1) regulates neuritogenesis. *The EMBO Journal*, *29*(10), 1637–1651.
- Shaheen, R., Jiang, N., Alzahrani, F., Ewida, N., Al-Sheddi, T., Alobeid, E., Musaev, D., Stanley, V., Hashem, M., Ibrahim, N., Abdulwahab, F., Alshenqiti, A., Sonmez, F. M., Saqati, N., Alzaidan, H., Al-Qattan, M. M., Al-Mohanna, F., Gleeson, J. G., & Alkuraya, F. S. (2019). Bi-allelic Mutations in FAM149B1 Cause Abnormal Primary Cilium and a Range of Ciliopathy Phenotypes in Humans. *American Journal of Human Genetics*, *104*(4), 731–737.
- Shang, J., Xia, T., Han, Q.-Q., Zhao, X., Hu, M.-M., Shu, H.-B., & Guo, L. (2018). Quantitative Proteomics Identified TTC4 as a TBK1 Interactor and a Positive Regulator of SeV-Induced Innate Immunity. *Proteomics*, *18*(2).
- Shiio, Y., Rose, D. W., Aur, R., Donohoe, S., Aebersold, R., & Eisenman, R. N. (2006). Identification and characterization of SAP25, a novel component of the mSin3 corepressor complex. *Molecular and Cellular Biology*, *26*(4), 1386–1397.
- Shi, Y.-Q., Li, Y.-C., Hu, X.-Q., Liu, T., Liao, S.-Y., Guo, J., Huang, L., Hu, Z.-Y., Tang, A. Y. B., Lee, K.-F., Yeung, W. S. B., Han, C.-S., & Liu, Y.-X. (2009). Male germ cell-specific protein Trs4 binds to multiple proteins. *Biochemical and Biophysical Research Communications*, *388*(3), 583–588.
- Shyu, M.-K., Lin, M.-C., Shih, J.-C., Lee, C.-N., Huang, J., Liao, C.-H., Huang, I.-F., Chen, H.-Y., Huang, M.-C., & Hsieh, F.-J. (2007). Mucin 15 is expressed in human placenta and suppresses invasion of trophoblast-like cells in vitro. *Human Reproduction*, *22*(10), 2723–2732.
- Siepkka, S. M., Yoo, S.-H., Park, J., Song, W., Kumar, V., Hu, Y., Lee, C., & Takahashi, J. S. (2007). Circadian mutant Overtime reveals F-box protein FBXL3 regulation of cryptochrome and period gene expression. *Cell*, *129*(5), 1011–1023.
- Sillibourne, J. E., Delaval, B., Redick, S., Sinha, M., & Doxsey, S. J. (2007). Chromatin remodeling proteins interact with pericentrin to regulate centrosome integrity. *Molecular Biology of the Cell*, *18*(9), 3667–3680.
- Singh, B. N., Gong, W., Das, S., Theisen, J. W. M., Sierra-Pagan, J. E., Yannopoulos, D., Skie, E., Shah, P., Garry, M. G., & Garry, D. J. (2019). Etv2 transcriptionally regulates Yes1 and promotes cell proliferation during embryogenesis. *Scientific Reports*, *9*(1), 9736.
- Singh, P., Patel, R. K., Palmer, N., Grenier, J. K., Paduch, D., Kaldis, P., Grimson, A., & Schimenti, J. C. (2019). CDK2 kinase activity is a regulator of male germ cell fate. *Development*, *146*(21).
- Sonna, L. A., Fujita, J., Gaffin, S. L., & Lilly, C. M. (2002). Invited review: Effects of heat and cold stress on mammalian gene expression. *Journal of Applied Physiology*, *92*(4), 1725–1742.
- Sun-Wada, G. H., Murakami, H., Nakai, H., Wada, Y., & Futai, M. (2001). Mouse Atp6f, the gene encoding the 23-kDa proteolipid of vacuolar proton translocating ATPase. *Gene*, *274*(1-2), 93–99.
- Tashita, C., Hoshi, M., Hirata, A., Nakamoto, K., Ando, T., Hattori, T., Yamamoto, Y., Tezuka, H., Tomita, H., Hara, A., & Saito, K. (2020). Kynurenine plays an immunosuppressive role in 2,4,6-

- trinitrobenzene sulfate-induced colitis in mice. *World Journal of Gastroenterology: WJG*, 26(9), 918–932.
- Tejada-Jiménez, M., Galván, A., & Fernández, E. (2011). Algae and humans share a molybdate transporter. *Proceedings of the National Academy of Sciences of the United States of America*, 108(16), 6420–6425.
- Tezuka, Y., Okada, M., Tada, Y., Yamauchi, J., Nishigori, H., & Sanbe, A. (2013). Regulation of neurite growth by inorganic pyrophosphatase 1 via JNK dephosphorylation. *PLoS One*, 8(4), e61649.
- The UniProt Consortium. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1), D158–D169.
- Touyama, K., Khan, M., Aoki, K., Matsuda, M., Hiura, F., Takakura, N., Matsubara, T., Harada, Y., Hirohashi, Y., Tamura, Y., Gao, J., Mori, K., Kokabu, S., Yasuda, H., Fujita, Y., Watanabe, K., Takahashi, Y., Maki, K., & Jimi, E. (2019). Bif-1/Endophilin B1/SH3GLB1 regulates bone homeostasis. *Journal of Cellular Biochemistry*, 120(11), 18793–18804.
- Tsantoulas, C., Denk, F., Signore, M., Nassar, M. A., Futai, K., & McMahon, S. B. (2018). Mice lacking *Kcns1* in peripheral neurons show increased basal and neuropathic pain sensitivity. *Pain*, 159(8), 1641–1651.
- Valente, P., Romei, A., Fadda, M., Sterlini, B., Lonardoni, D., Forte, N., Fruscione, F., Castroflorio, E., Michetti, C., Giansante, G., Valtorta, F., Tsai, J.-W., Zara, F., Nieus, T., Corradi, A., Fassio, A., Baldelli, P., & Benfenati, F. (2019). Constitutive Inactivation of the PRRT2 Gene Alters Short-Term Synaptic Plasticity and Promotes Network Hyperexcitability in Hippocampal Neurons. *Cerebral Cortex*, 29(5), 2010–2033.
- Wehner, K. A., & Baserga, S. J. (2002). The sigma(70)-like motif: a eukaryotic RNA binding domain unique to a superfamily of proteins required for ribosome biogenesis. *Molecular Cell*, 9(2), 329–339.
- Weinstat-Saslow, D. L., Germino, G. G., Somlo, S., & Reeders, S. T. (1993). A transducin-like gene maps to the autosomal dominant polycystic kidney disease gene region. *Genomics*, 18(3), 709–711.
- Wild, T., Budzowska, M., Hellmuth, S., Eibes, S., Karemore, G., Barisic, M., Stemmann, O., & Choudhary, C. (2018). Deletion of APC7 or APC16 Allows Proliferation of Human Cells without the Spindle Assembly Checkpoint. *Cell Reports*, 25(9), 2317–2328.e5.
- Wiley, S. R., Cassiano, L., Lofton, T., Davis-Smith, T., Winkles, J. A., Lindner, V., Liu, H., Daniel, T. O., Smith, C. A., & Fanslow, W. C. (2001). A novel TNF receptor family member binds TWEAK and is implicated in angiogenesis. *Immunity*, 15(5), 837–846.
- Wu, S. Y., Thomas, M. C., Hou, S. Y., Likhite, V., & Chiang, C. M. (1999). Isolation of mouse TFIID and functional characterization of TBP and TFIID in mediating estrogen receptor and chromatin transcription. *The Journal of Biological Chemistry*, 274(33), 23480–23490.
- Wu, X., Quondamatteo, F., Lefever, T., Czuchra, A., Meyer, H., Chrostek, A., Paus, R., Langbein, L., & Brakebusch, C. (2006). Cdc42 controls progenitor cell differentiation and beta-catenin turnover in skin. *Genes & Development*, 20(5), 571–585.
- Xiao, Q., Wu, X.-L., Michal, J. J., Reeves, J. J., Busboom, J. R., Thorgaard, G. H., & Jiang, Z. (2006). A novel nuclear-encoded mitochondrial poly(A) polymerase PAPD1 is a potential candidate gene for the extreme obesity related phenotypes in mammals. *International Journal of Biological Sciences*, 2(4), 171–178.
- Xie, L., Qin, W.-X., He, X.-H., Shu, H.-Q., Yao, G.-F., Wan, D.-F., & Gu, J.-R. (2004). Differential gene expression in human hepatocellular carcinoma Hep3B cells induced by apoptosis-related gene BNIPL-2. *World Journal of Gastroenterology: WJG*, 10(9), 1286–1291.
- Xie, X.-K., Xu, Z.-K., Xu, K., & Xiao, Y.-X. (2020). DUSP19 mediates spinal cord injury-induced apoptosis and inflammation in mouse primary microglia cells via the NF- κ B signaling pathway. *Neurological Research*, 42(1), 31–38.
- Xu, Z., Yang, L., Xu, S., Zhang, Z., & Cao, Y. (2015). The receptor proteins: pivotal roles in selective autophagy. *Acta Biochimica et Biophysica Sinica*, 47(8), 571–580.
- Yamakoshi, T., Makino, T., Ur Rehman, M., Yoshihisa, Y., Sugimori, M., & Shimizu, T. (2013).

- Trichohyalin-like 1 protein, a member of fused S100 proteins, is expressed in normal and pathologic human skin. *Biochemical and Biophysical Research Communications*, 432(1), 66–72.
- Yang, X., Matsuda, K., Bialek, P., Jacquot, S., Masuoka, H. C., Schinke, T., Li, L., Brancorsini, S., Sassone-Corsi, P., Townes, T. M., Hanauer, A., & Karsenty, G. (2004). ATF4 is a substrate of RSK2 and an essential regulator of osteoblast biology; implication for Coffin-Lowry Syndrome. *Cell*, 117(3), 387–398.
- Ye, W., Zhou, Y., Xu, B., Zhu, D., Rui, X., Xu, M., Shi, L., Zhang, D., & Jiang, J. (2019). CD247 expression is associated with differentiation and classification in ovarian cancer. *Medicine*, 98(51), e18407.
- Ye, Z., & Ting, J. P.-Y. (2008). NLR, the nucleotide-binding domain leucine-rich repeat containing gene family. *Current Opinion in Immunology*, 20(1), 3–9.
- Yockey, L. J., & Iwasaki, A. (2018). Interferons and Proinflammatory Cytokines in Pregnancy and Fetal Development. *Immunity*, 49(3), 397–412.
- Yu, Y.-H., Chang, Y.-C., Su, T.-H., Nong, J.-Y., Li, C.-C., & Chuang, L.-M. (2013). Prostaglandin reductase-3 negatively modulates adipogenesis through regulation of PPAR γ activity. *Journal of Lipid Research*, 54(9), 2391–2399.
- Zhang, C., & Liang, Y. (2018). Latexin and hematopoiesis. *Current Opinion in Hematology*, 25(4), 266–272.
- Zhang, H., Ge, Y., He, P., Chen, X., Carina, A., Qiu, Y., Aga, D. S., & Ren, X. (2015). Interactive Effects of N6AMT1 and As3MT in Arsenic Biomethylation. *Toxicological Sciences: An Official Journal of the Society of Toxicology*, 146(2), 354–362.
- Zhang, W. J., & Wu, J. Y. (1998). Sip1, a novel RS domain-containing protein essential for pre-mRNA splicing. *Molecular and Cellular Biology*, 18(2), 676–684.
- Zhang, X., Azhar, G., Zhong, Y., & Wei, J. Y. (2006). Zipzap/p200 is a novel zinc finger protein contributing to cardiac gene regulation. *Biochemical and Biophysical Research Communications*, 346(3), 794–801.
- Zhao, H., Zhu, L., Zhu, Y., Cao, J., Li, S., Huang, Q., Xu, T., Huang, X., Yan, X., & Zhu, X. (2013). The Cep63 paralogue Deup1 enables massive de novo centriole biogenesis for vertebrate multiciliogenesis. *Nature Cell Biology*, 15(12), 1434–1444.

Appendix B.

Supporting information for Chapter III, Palma-Vera et al., 2021
Genomic characterization of the world's longest selection experiment in mouse reveals the complexity of polygenic traits

Supporting tables can be accessed at:

<https://bmcbiol.biomedcentral.com/articles/10.1186/s12915-022-01248-9#Sec24>

Supplementary Methods

1. Establishment of the Dummerstorf mouse lines

During the years 1969 and 1970, four outbred strains (NMRI orig., Han:NMRI, CFW, CF1) and four inbred strains (CBA/Bln, AB/Bln, C57BL/Bln, XVII/Bln) were systematically crossed to establish the line **FZTDU** (*Forschungszentrum für Tierproduktion Dummerstorf*) [5,6]. Full-sib mating was avoided by random mating. This line has been kept unselected for almost 200 generations. The line was originally maintained with 200 breeding pairs per generation until animals were moved from a conventional semi-barrier housing into a specific pathogen-free (SPF) environment on generation ~160. This transition could only be accomplished through a limited number of embryo transfers and as a result, the number of breeding pairs dropped to 55. Thereafter, the number of breeding pairs has been kept at 125 (the current number of breeding pairs) avoiding consanguinity by random mating. All the Dummerstorf selection lines were derived from FZTDU starting at different time-points.

In general, all trait-selected lines were developed through among-family selection [108], whereby litters were ranked according to each trait of interest (Table 2) and then parents were chosen at random from the highest ranked litters. The proportion of litters selected varied generation to generation (Additional file 2: Figure S2). The trait-selected lines were maintained with 60-100 breeding pairs. However, when animals had to be relocated to the new SPF animal housing building in 2011 after 120-165 generations, the number of breeding pairs drastically dropped to ~20 for DUK, DUC, DU6P and DUhLB, and as low as 7 for DU6 (Table 1).

The process of selection to establish the **fertility lines DUK and DUC** began shortly after the establishment of FZTDU. In 1971, each fertility line was started with 60 breeding pairs. These animals were drawn from FZTDU by phenotypic maternal selection for number of offspring and litter weight at birth in the first litter (for more details see [22]). However, for an interim period of 10 generations families were ranked only by litter weight. Selection for fertility continues to this day spanning more than 190 generations. The lowest number of breeding pairs was 19 (DUK) and 24 (DUC) at generation ~164 when animals were transferred to a new facility by embryo transfer. Nowadays, the number of breeding pairs is 60 for both lines. Before relocation, the number of breeding pairs per generation and the selection intensity ranged between 60% and 100% and between 25% and 45%, respectively. A few generations after relocation, family information and individual information was combined in a pedigree-based BLUP (Best Linear Unbiased Prediction) [109] estimation of breeding values and selection was conducted accordingly.

The **body mass line DU6** was started in 1975 by phenotypic selection of FZTDU by ranking litters according to total weight of two randomly sampled males from each litter at 42 days of age.

Whenever possible these males were not chosen as sires. DU6 was founded by mating 80 pairs at around 9 weeks of age. The number of breeding pairs until generation ~154 was kept at 60-100 pairs. On generation ~154, animals were transferred to the new facility and the number of breeding pairs decreased to only 7 because of embryo transfer yield. Thereafter the number of breeding pairs was increased to 60 and phenotyping was massively extended by taking body weights at day 42 from all progeny, including females. The selection strategy was later changed on generation 161 to selection based on estimated breeding values. The breeding values for body mass at day 42 were calculated using with BLUP [109]. As of generation 173 the proportion of female to male breeders has been kept at 2:1 (120 females and 60 males approx.) to mitigate the decreased pregnancy rate observed in DU6 females. This line continues to be selected (selection intensity of 45-90%).

Also in 1975, the **DU6P** line was established by selection for weight and protein content of the carcass of a single male from each litter at 42 days of age. Occasionally protein mass could not be determined (e.g. because of technical issues or limited lab capacities), in which case litters were ranked by the combined weight of two males, as described for line DU6. The number of breeding pairs at the start of this line was 80 and was kept at 60-80 breeding pairs per generation. Then, on generation ~154, animals were relocated with 19 breeding pairs as founders, which were then increased to 60 pairs. Selection in DU6P stopped at generation 152 and the line is currently preserved without selection pressure by allowing random mating.

Finally, the high **endurance line DUhLB** was started in 1982 based on selection for high treadmill performance. It is thus the youngest of the Dummerstorf selection lines, as well as the shortest selected one (selection stopped at generation 141). Male running performance was evaluated based on distance (meters) covered on a treadmill before exhaustion (submaximal test). Trials were conducted after mating at 11 weeks of age. Subjects had no previous access to any kind of equipment that would influence their performance (untrained). Offspring of the highest scoring subjects were chosen for breeding. The line was founded with 100 breeding pairs and a selection intensity of 40% for the first 25 generations. Thereafter, the line was maintained with 60-80 breeding pairs at a selection intensity of 45-100%. On generation ~120, 44 breeding pairs were used as founders after transferring to the new facility. Together with DU6P, DUhLB is currently preserved without selection.

2. Structural Variant Calling

Mapped and deduplicated short PE reads were used in detection of structural variants. As depth of coverage of reads mapped to the reference mouse genome sequence varied between 5 and 20x, we

have split samples for each line into a high (10 samples) and low coverage (15 samples) set, and conducted structural variation analysis separately on these sets.

Three SV callers, Manta v.1.6.0 [110], Whamg v.1.7.0 [111] and Lumpy v.0.2.13 [112] were selected based on their sensitivity and precision [113]. Manta integrates paired-read (PR), split-read (SR) evidence and SV breakend assembly (AS) during SV discovery. Whamg implements PR and SR support while Lumpy relies on a probabilistic copy number variation discovery combining PR, SR and read-depth (RD). Intersecting results of multiple SV callers applying different SV detection approaches has previously been shown to improve accuracy of variant call sets [113,114].

Manta SV calls in genomic regions with depth greater than 3x the median chromosome depth near one or both SV breakends mainly caused by reads mapping in low complexity regions, as well as reads with MAPQ<30 were filtered out. Furthermore, variant calls for which samples did not pass Manta caller quality filters, and have genotype quality below 20 (GQ<20) have been filtered out. Calls with paired-read (PR) and split-read (SR) support of PR≥3 and SR≥3 were retained.

Whamg SV calls of size <50bp and >2Mb were filtered out, along with calls with fewer than 5 supporting reads and GQ<20. Calls associated with poorly mapped regions and BND-type calls with high cross-chromosomal mapping scores (CW>0.2) were removed, as Whamg does not specifically call translocations. Lumpy calls for which evidence supporting the variant were below 5 (SU<5) and calls with GQ<20 were filtered out. Both Whamg and Lumpy SV call sets were genotyped with Svtiper v0.7.1 [101].

Unlocalized and unplaced scaffolds have been removed from all SV sets and only scaffolds assigned to chromosomes have been included in further analysis. Survivor v.1.0.7 [102] was used to merge SV call sets within and among samples. For each mice line sample, we first merged SV events of the same type, called by at least two SV callers, with start/ end positions detected within +/-1000 bp, identified in at least 60% of samples in a low coverage set (10 samples out of 15) and 100% of samples in a high coverage set (all 10 samples) for each mice line.

The union of SVs detected in two separate sample sets for each line were further used. We then intersected SV calls among all mice lines to obtain SVs private for each mice line (line-specific) and shared among lines (Additional file 2: Figure S11). To further reduce the FDR, SV calls overlapping gaps and high coverage regions (> 80x) in the reference genome assembly were filtered out. High coverage regions were determined for each mouse line based on the intersection among samples of 1-kb windows containing reads mapping with depth of coverage > 80x. Variants specified as “BND”

(translocations) were removed and deletions, and duplications and inversions were further investigated (Additional file 2: Figure S12-S15).

We annotated the final SV set with the Ensembl VEP v. 101.0 [103] focusing on variants overlapping protein-coding genes (maximum SV size = 200 Mb). Functional classification of genes was based on literature and database search (OrthoDB v10 [104]; Uniprot [107]; NCBI Entrez gene [105]), and Gene Ontology enrichment analysis (Shiny GO, FDR < 0.05 [106]).

Supplementary figures

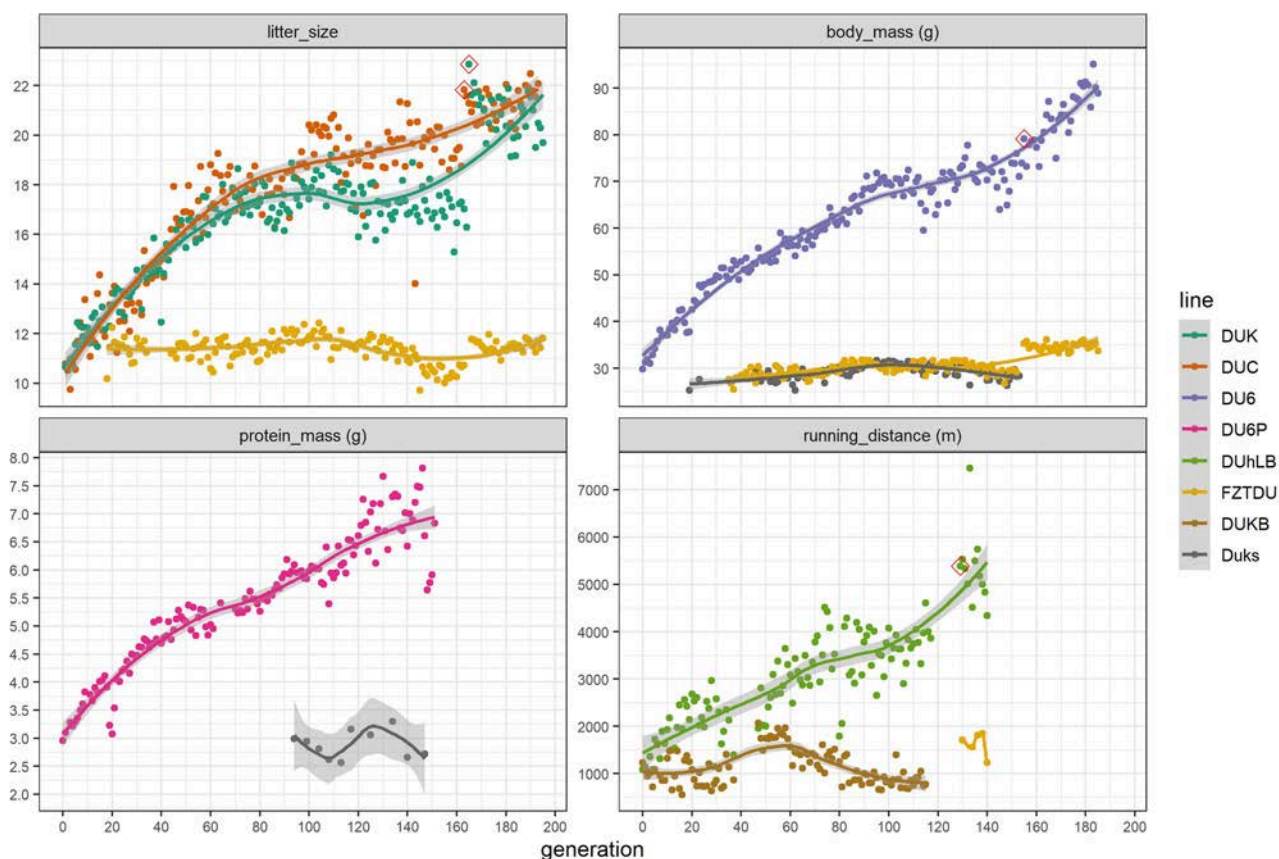


Figure S1. Response to selection throughout the selection experiment. Data points indicate the trait-value at each generation for the trait-selected lines and their controls. The first measure taken after relocation to a new mouse house is indicated by a red diamond shape surrounding the data point. Trend lines and confidence intervals are based on local regression (locally estimated scatterplot smoothing). For the case of the protein-mass line DU6P, measurements were not collected after relocation. Though FZTDU can be considered the control line, as it has been evolving neutrally over the span of the breeding experiment, other lines have been used as controls in the past and are indicated in the figure as “Duks” (specific control line for the body (DU6) and protein (DU6P) mass lines) and “DUKB” (specific control line for the treadmill performance line DUhLB). None of these line-specific control lines exist anymore and for the case of “Duks”, data is incomplete.

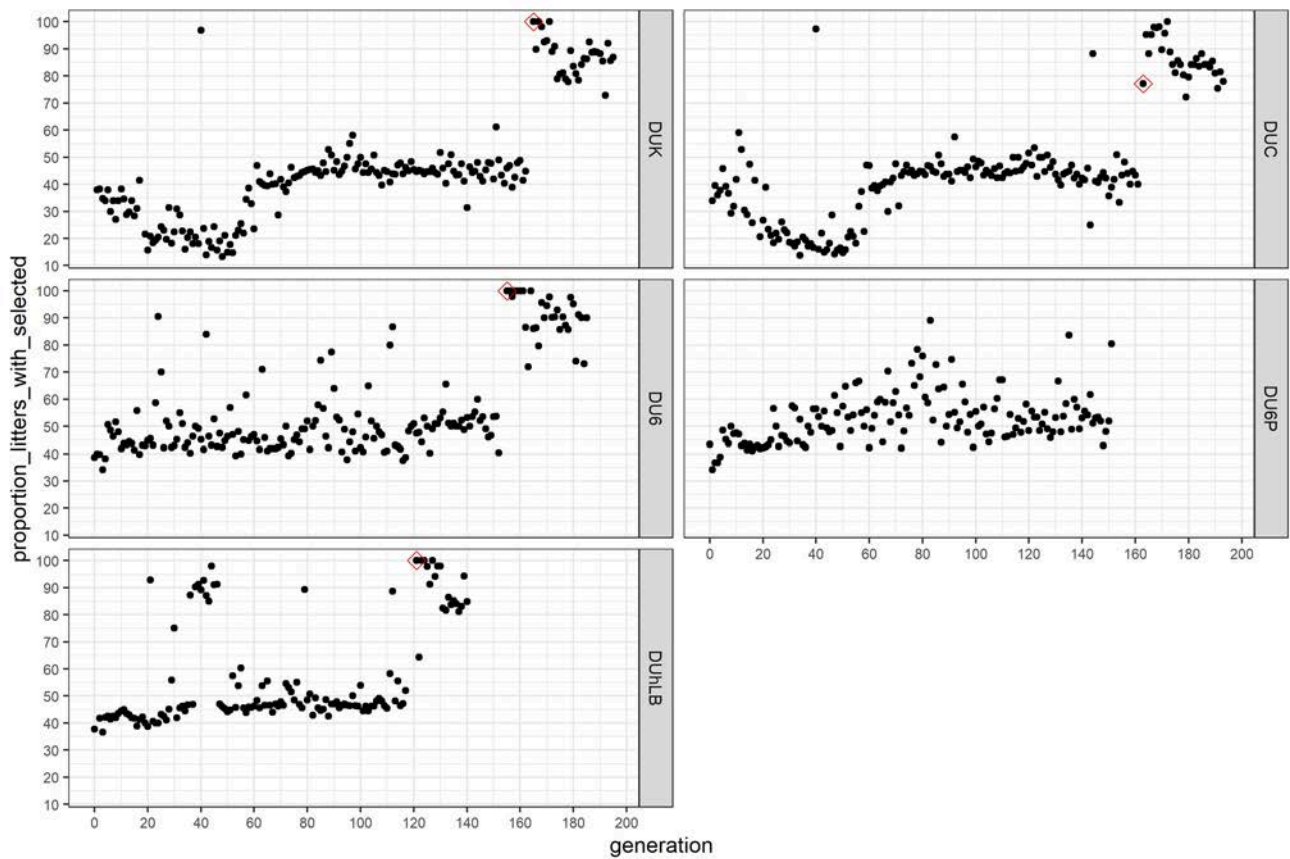


Figure S2. Proportion of litters supplying parents for the next generation. Proportions varied according to the selection intensities required by the breeding program over the last ~50 years. Indicated by a diamond shape are the data points corresponding to the first generation after relocation to a new mouse house. For the case of the protein-mass line DU6P, there was no data after relocation.

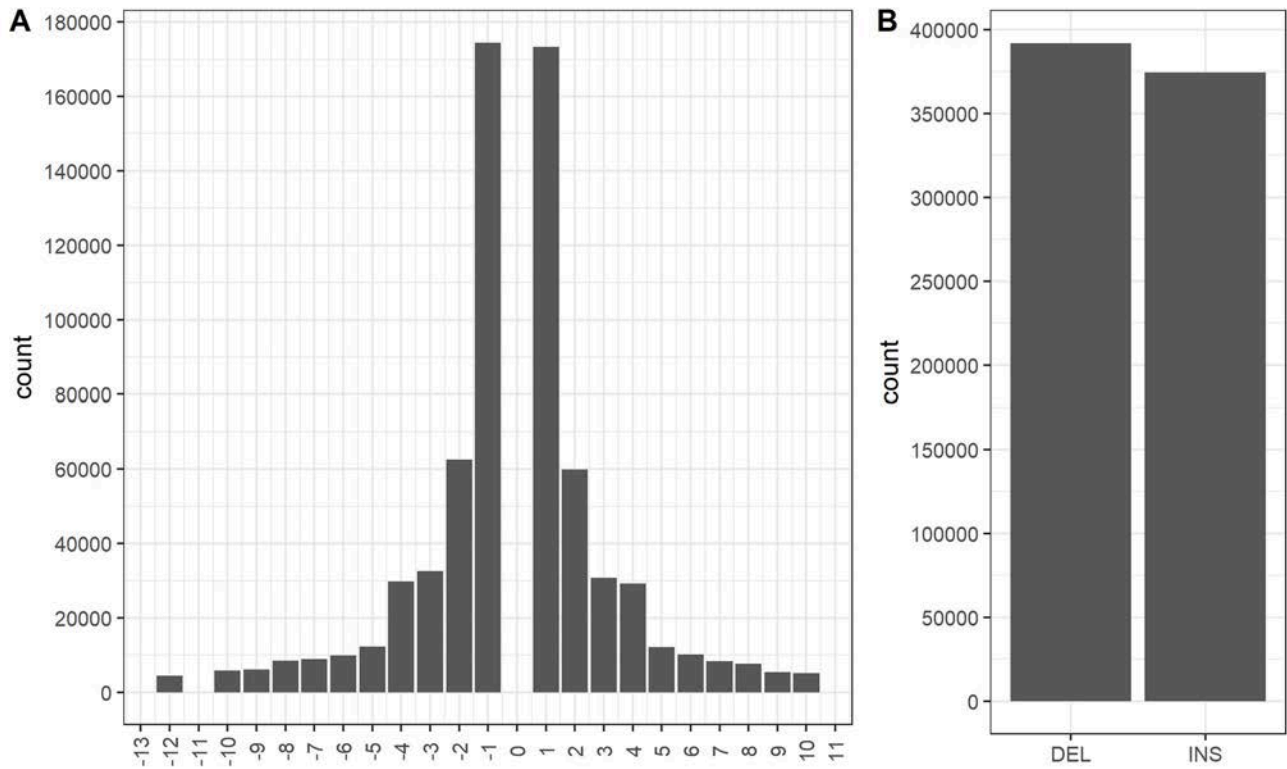


Figure S3. Distribution of INDEL lengths. (A) INDEL length distribution comprising ~90% of INDELS. (B) Total number of insertions and mutations (392,051 deletions and 374,604 insertions). Only INDELS outside microsatellites were considered.

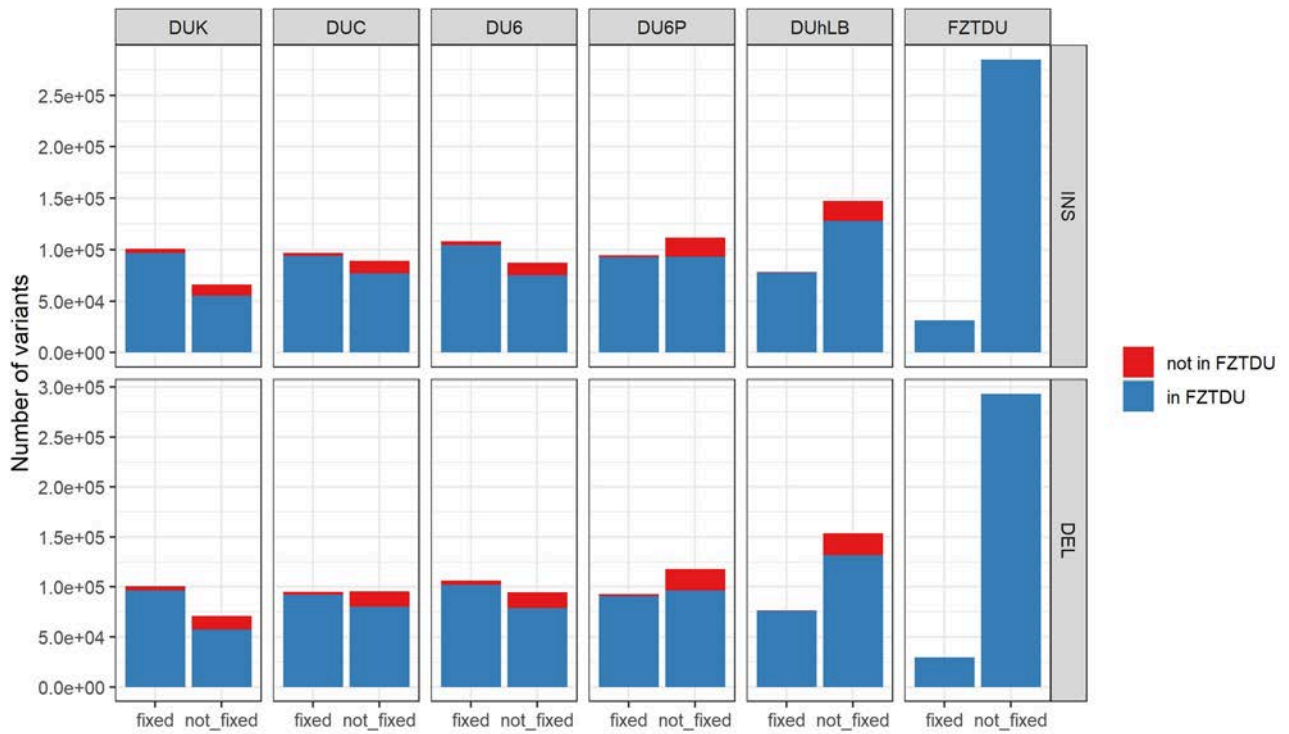


Figure S4. Classification of INDELs by type, fixation and presence in control line. INDEL sites were classified as fixed or not-fixed if their allele frequencies were 1 or <1, respectively. At each line, the fraction of INDELs shared (in FZTDU, blue) and not shared (not in FZTDU, red) with FZTDU is also shown. INDELs overlapping microsatellite regions were removed from the analysis.

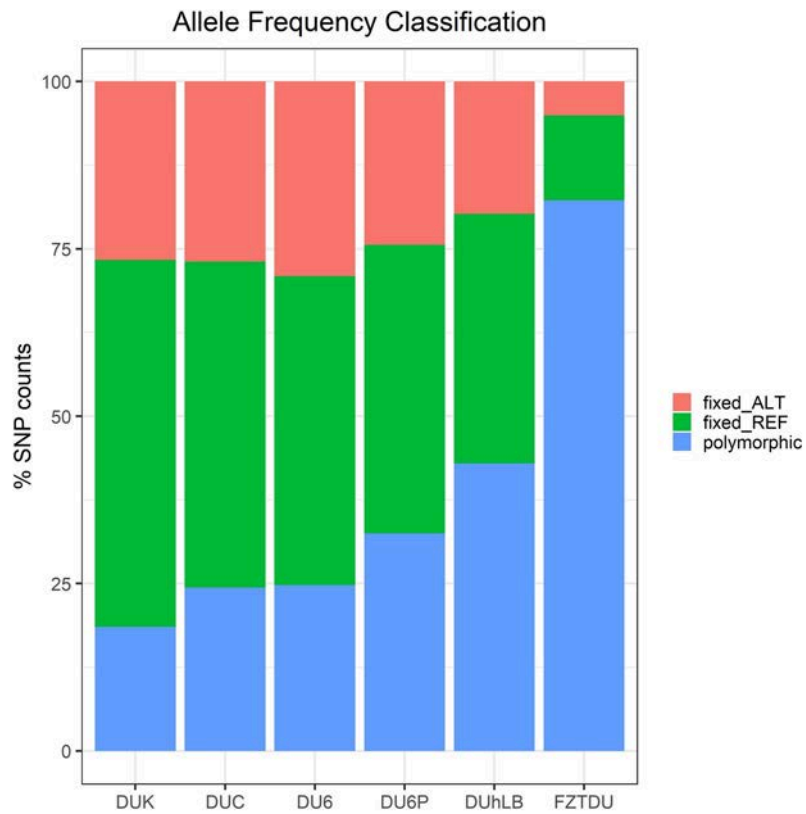


Figure S5. SNP allele frequency state classification. Sites in which SNPs were observed across the whole set of samples are classified for each mouse line as fixed-alternative (alternative allele homozygous in all subjects, fixed_ALT), fixed-reference (reference allele homozygous in all subjects, fixed_REF) and polymorphic (not fixed).

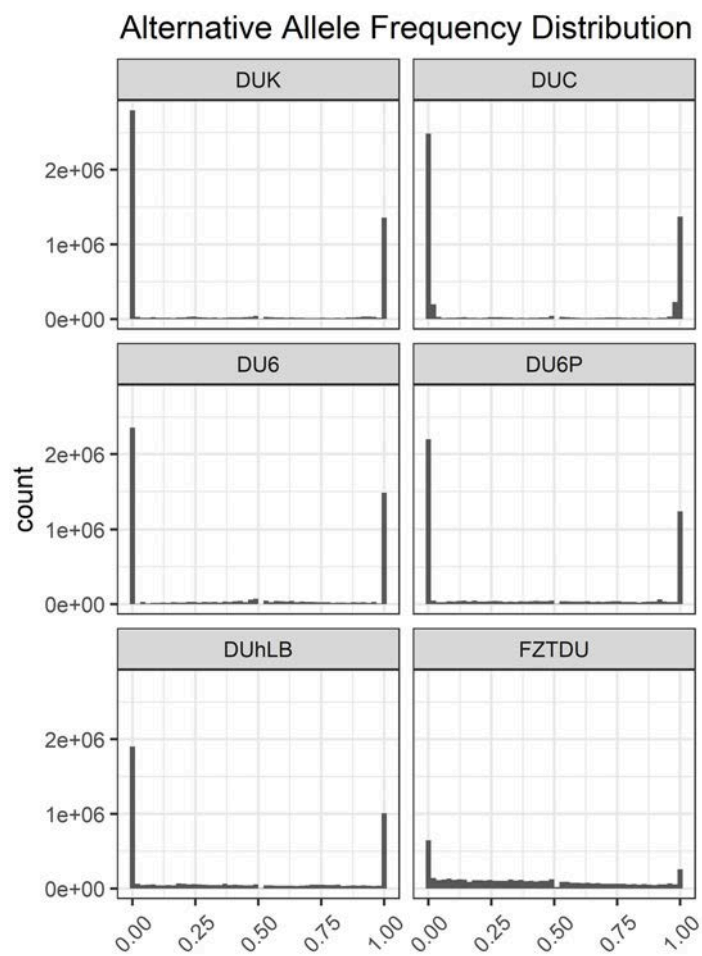


Figure S6. Alternative allele frequency distribution. Counts of SNPs along the allele frequency spectrum for each mouse line.

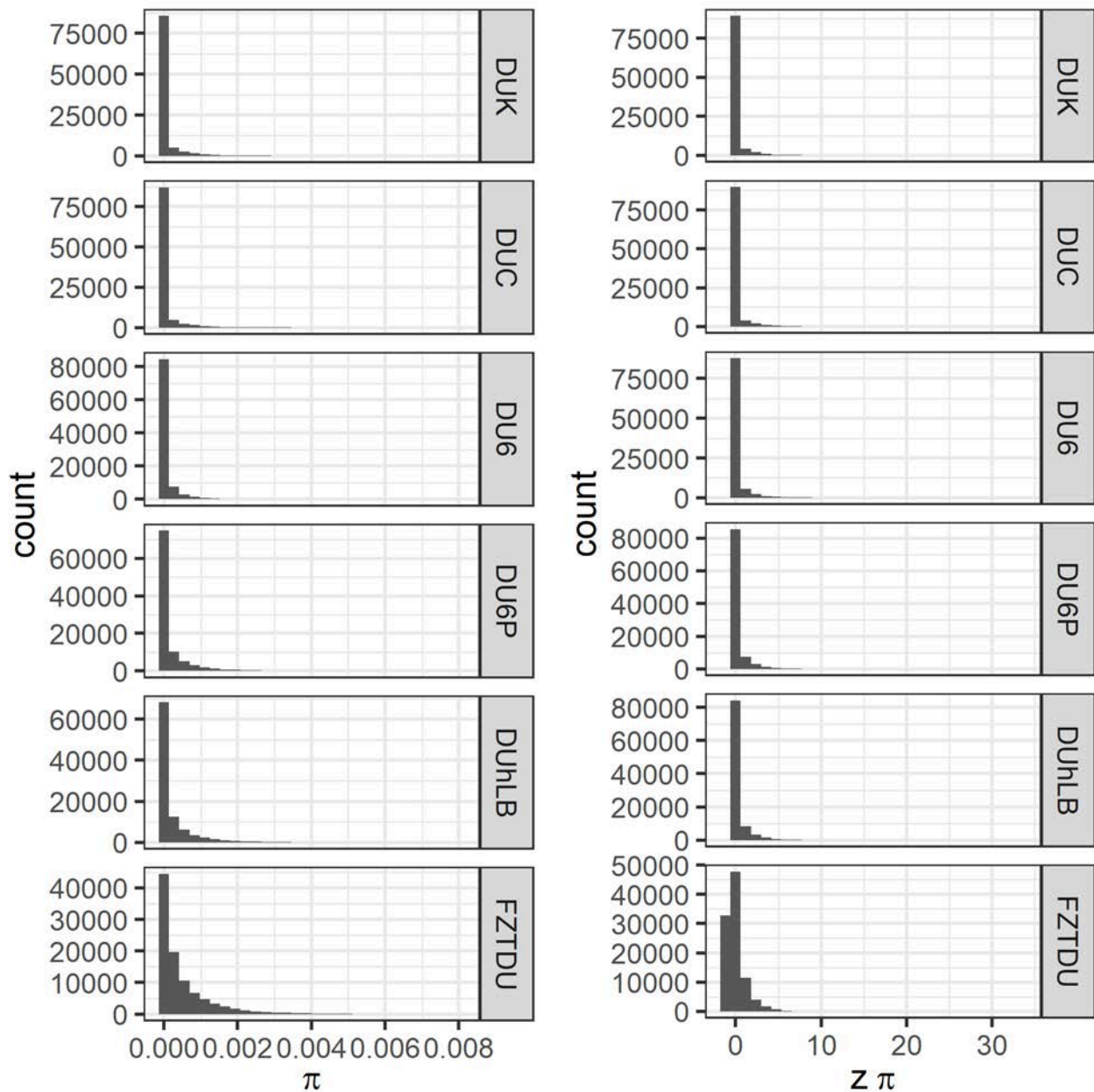


Figure S7. Nucleotide diversity (π) distribution in the Dummerstorf mouse lines. Nucleotide diversity (π) was calculated in sliding window mode (size=50Kb, step=25Kb, ≥ 10 SNPs). Scores were transformed to z-scores in order to represent the data in terms of standard deviations from the genomic mean. The distributions illustrate the low levels of genetic diversity within lines, with the most scores accumulate at the lower end of the distribution (left), thus highly diverse regions are rare events.

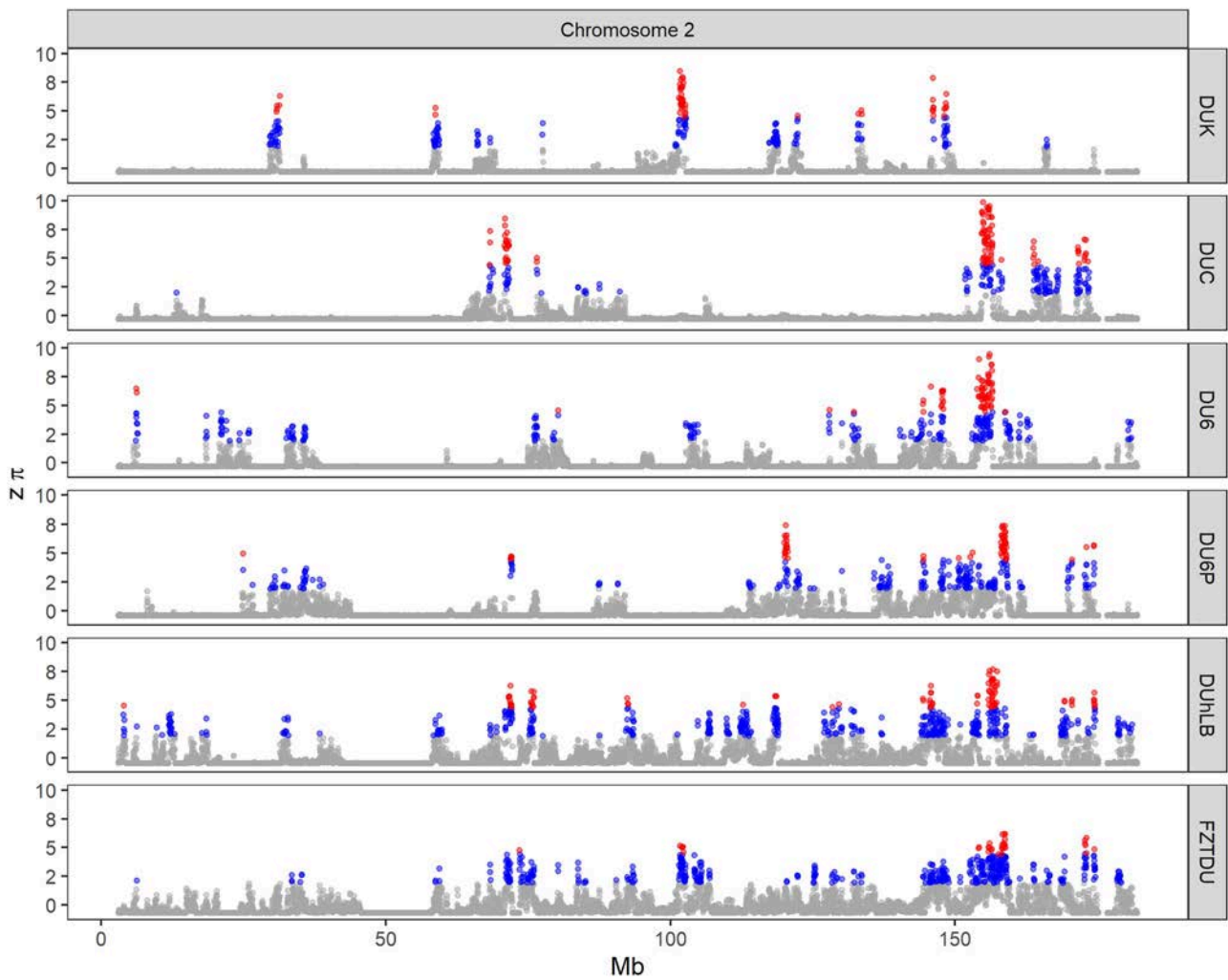


Figure S8. Example of one chromosome representative of the level of genetic diversity observed in the Dummerstorf mouse lines. The stretches of low diversity are longer and more abundant than in FZTDU. Regions of extreme genetic diversity are shown in blue (top 5% most diverse windows) and red (top 1% most diverse windows).

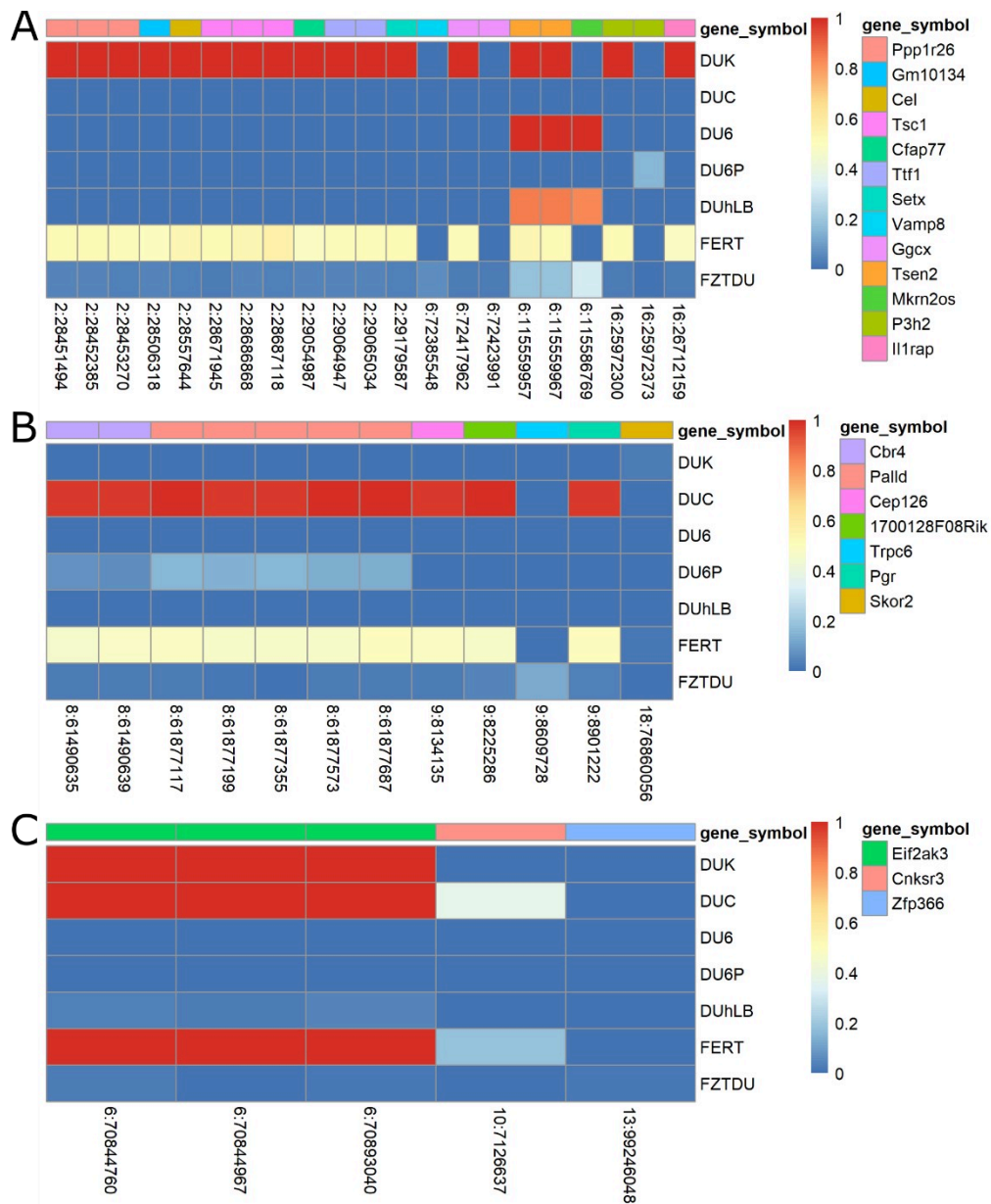


Figure S9. Allele frequency heatmap of non-synonymous mutations in RDD genes. Allele frequencies of non-synonymous SNPs in genes overlapping regions of distinct genetic differentiation for DUK (A), DUC (B) and the joint fertility population FERT (C). The gradient scale represents the allele frequency from low (blue) to high (red).

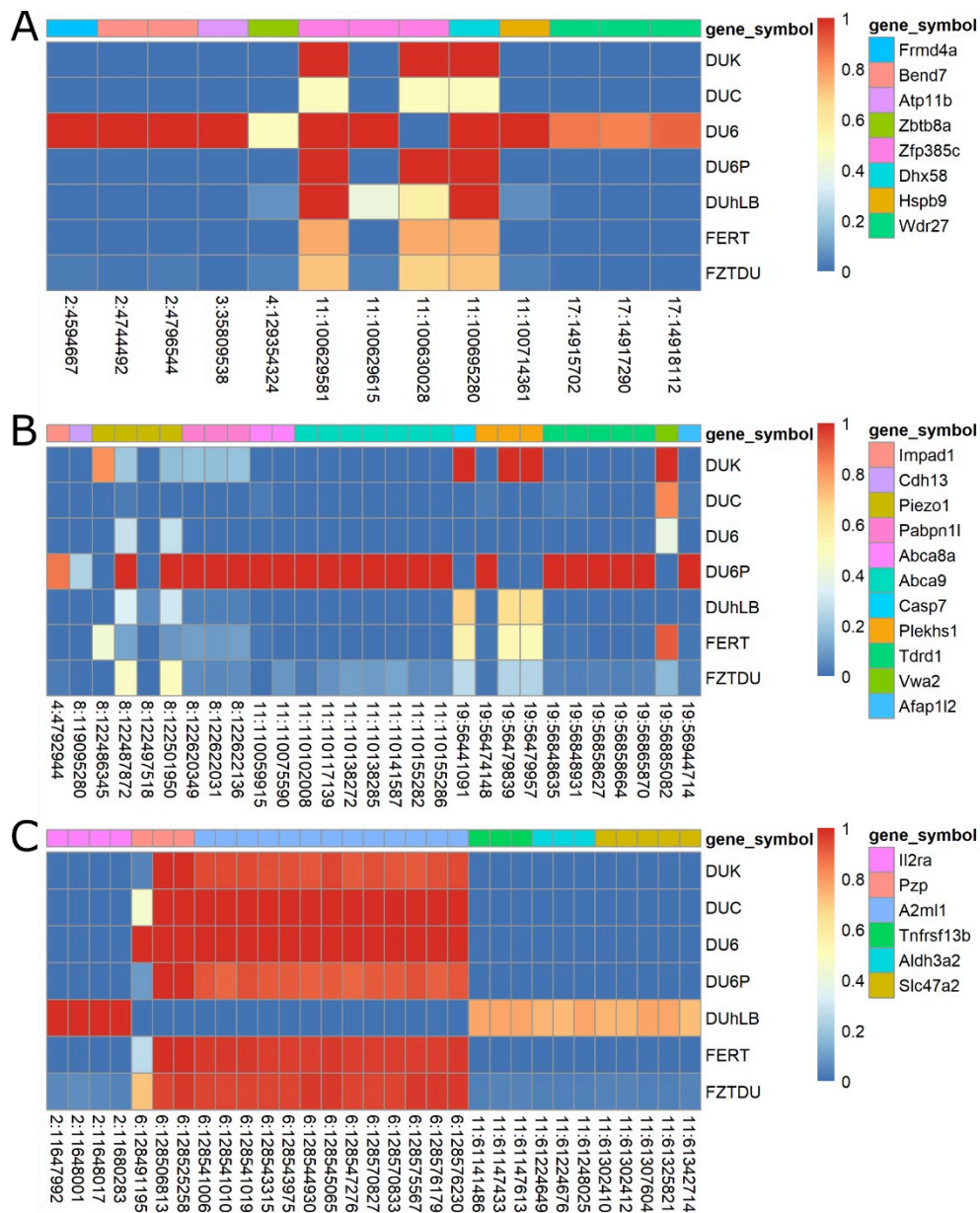


Figure S10. Allele frequency heatmap of non-synonymous mutations in RDD genes. Allele frequencies of non-synonymous SNPs in genes overlapping regions of distinct genetic differentiation for DU6 (**A**), DU6P (**B**) and DUhLB (**C**). The gradient scale represents the allele frequency from low (blue) to high (red).

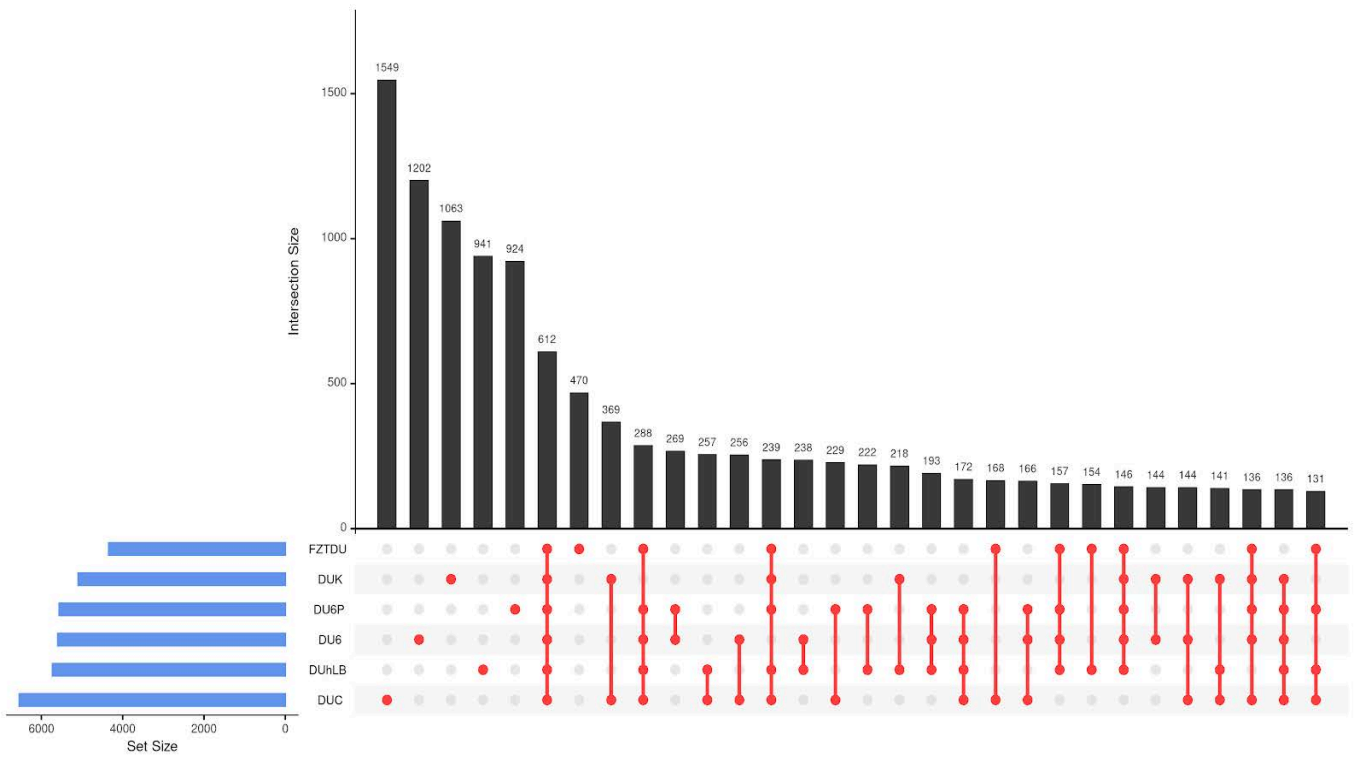


Figure S11. Shared and line-specific structural variants. SVs detected in a union of high and low coverage sample sets for each mice line.

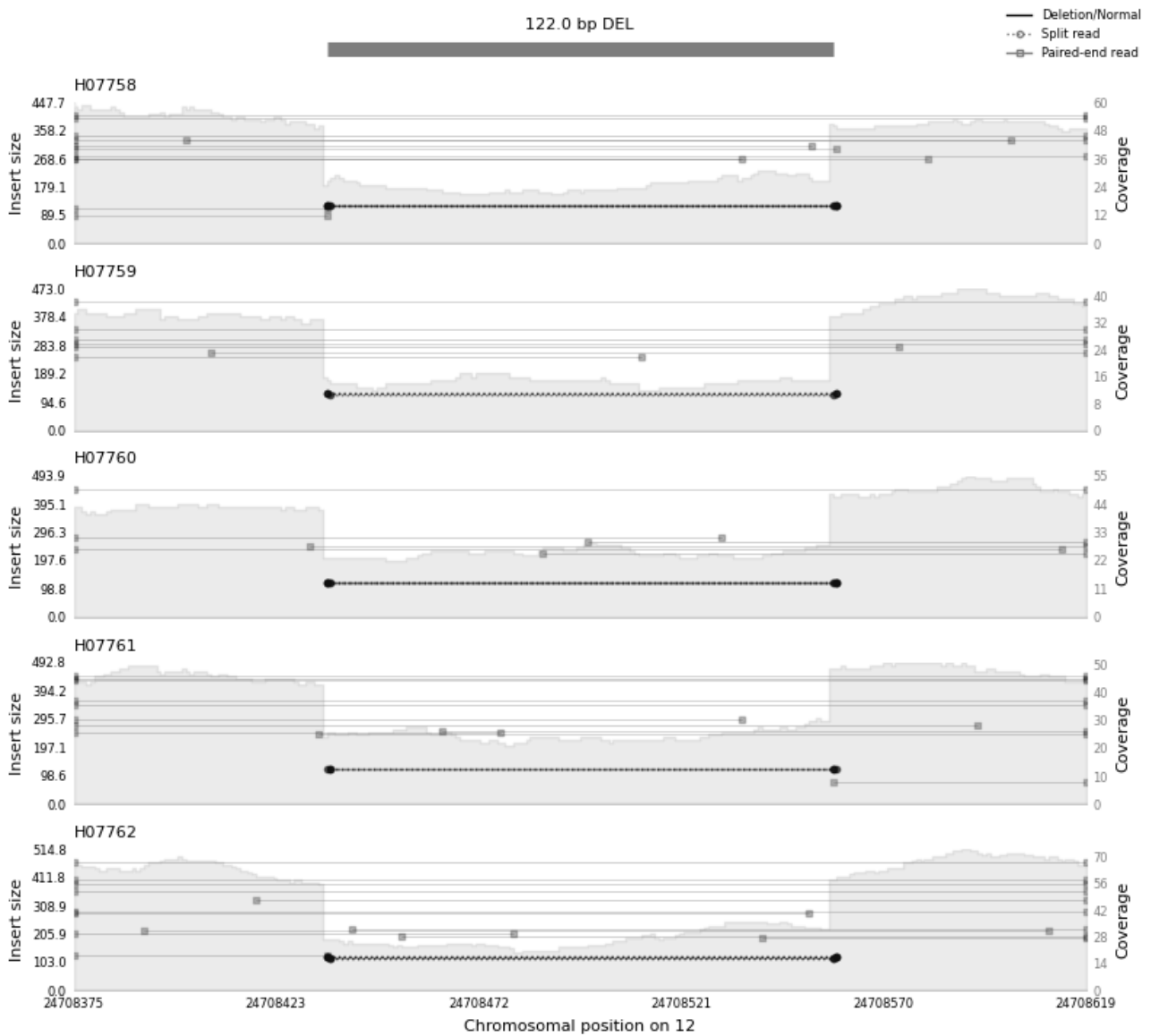


Figure S12. Example of a polymorphic deletion. SV is shown in 5 out of 10 samples of the high coverage set, indicating the genomic location (x-axis), the insert size (y-axis, left) and coverage (y-axis, right).

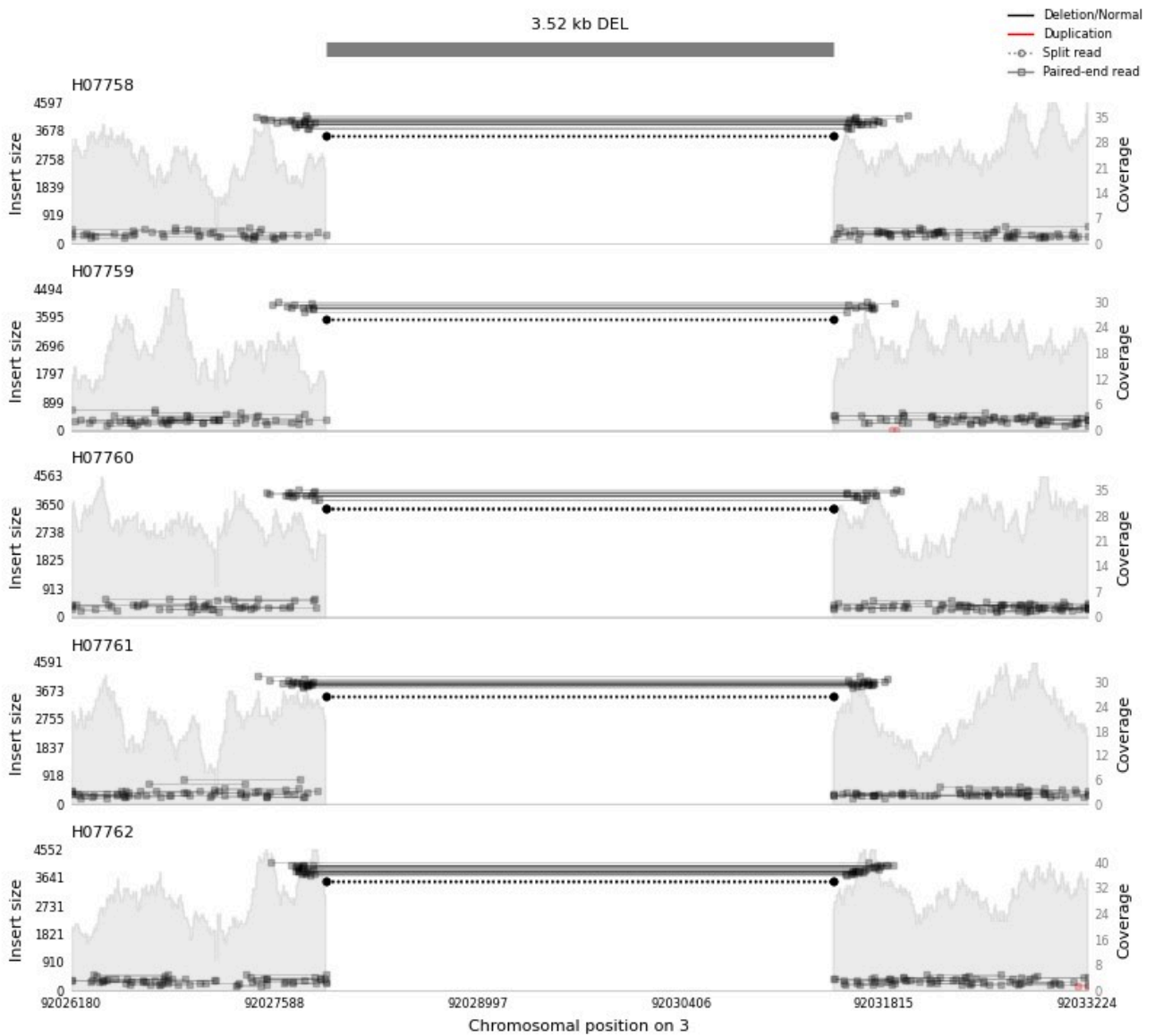


Figure S13. Example of a fixed deletion. SV is shown in 5 out of 10 samples of the high coverage set, indicating the genomic location (x-axis), the insert size (y-axis, left) and coverage (y-axis, right).

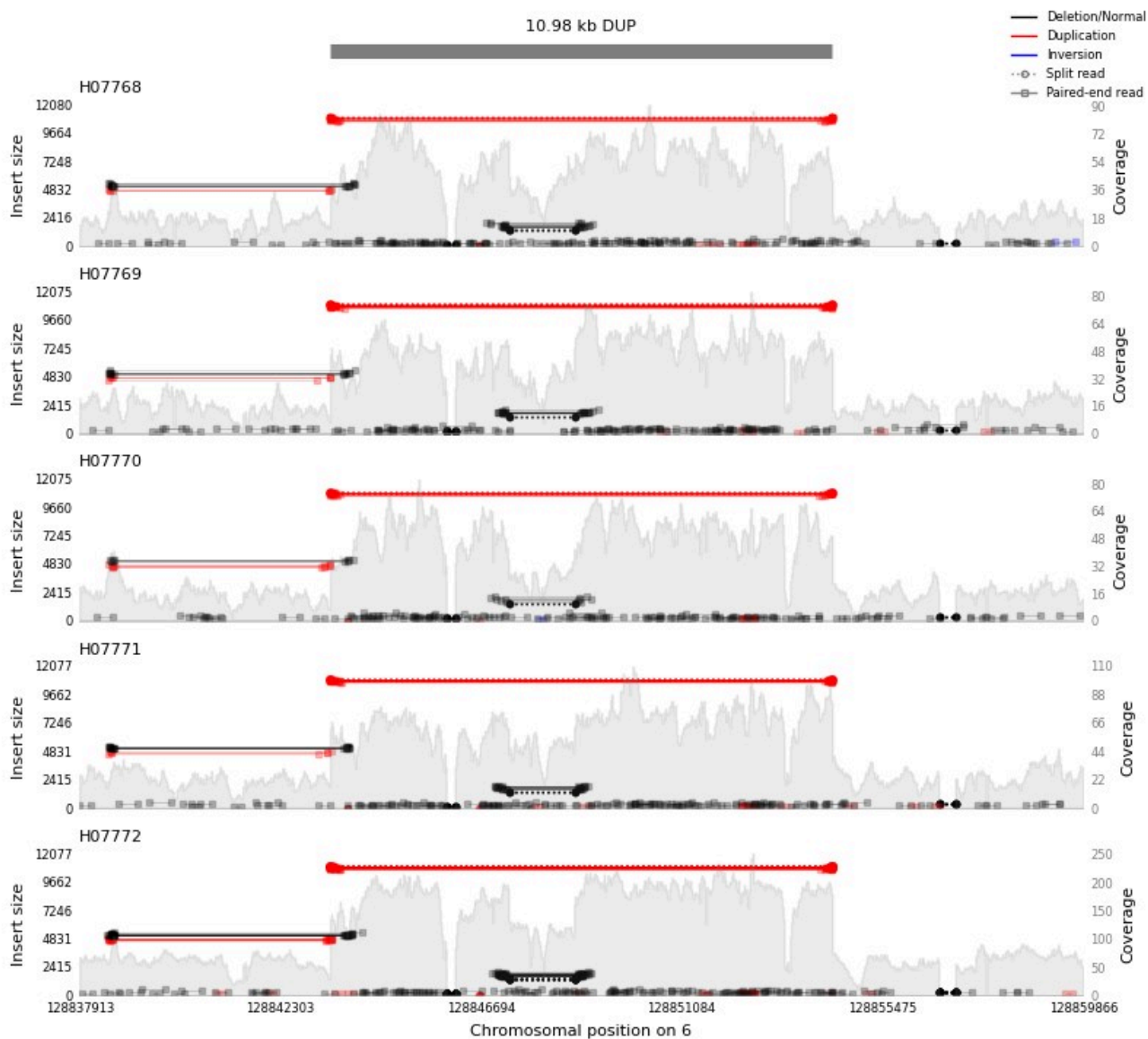


Figure S14. Example of a fixed duplication. SV is shown in 5 out of 10 samples of the high coverage set, indicating the genomic location (x-axis), the insert size (y-axis, left) and coverage (y-axis, right).

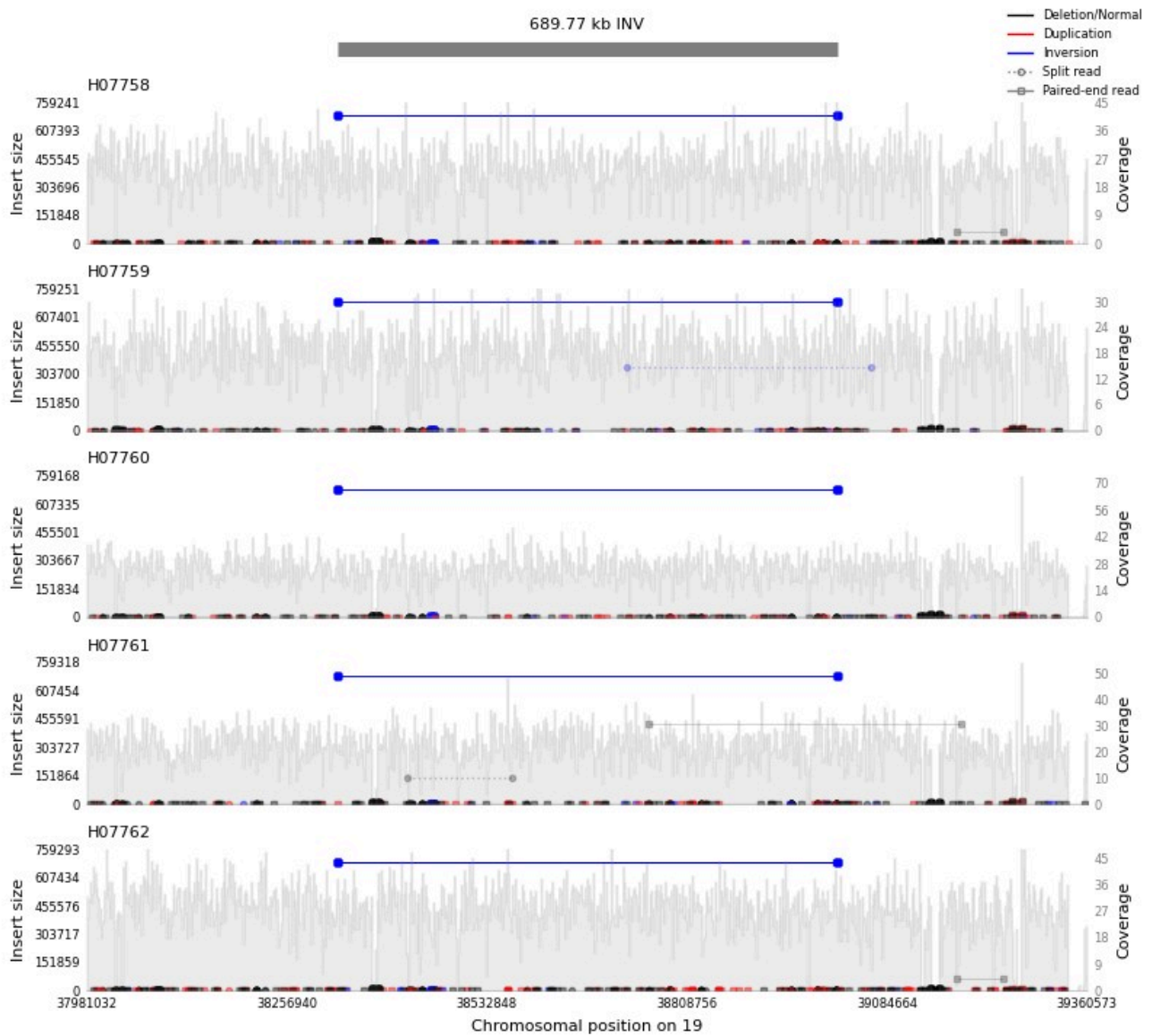


Figure S15. Example of a fixed inversion. SV is shown in 5 out of 10 samples of the high coverage set, indicating the genomic location (x-axis), the insert size (y-axis, left) and coverage (y-axis, right).

Supplementary tables

Table S1. Alternative names of Dummerstorf mouse lines and reference list of scientific articles based on these mouse lines. Status: November 2021

Line	Alternative Name and Reference
DUK	DU-K [37], FL1 [20,22-24,38-41]
DUC	DU-C [37], FL2 [20,22,38,40]
DU6	BW [25,26], Titan [21]
DU6P	PA [25]
DUhLB	DU-hTP [10,27-32]
FZTDU	Fzt: DU [6,8,25,26,32,37], DUK [34-36], Ctrl [22,23]

Table S2. Number of SNP and INDEL sites discovered in each line

Line	SNPs	% in FZTDU	INDELs (Insertions + Deletions)	% in FZTDU
DUK	2,305,349	92.76	338,380 (166,838 + 171,542)	90.41
DUC	2,615,584	92.76	376,453 (185,791 + 190,662)	91.20
DU6	2,744,788	93.18	396,022 (195,415 + 200,607)	90.97
DU6P	2,899,902	91.04	417,027 (206,164 + 210,863)	89.52
DUhLB	3,196,655	92.26	455,789 (225,415 + 230,374)	90.65
FZTDU	4,453,865	--	638,500 (315,851 + 322,649)	--

Table S3. Number of private variants with predicted high/moderate effects according to SnpEff.

	SNPs	INDELs	Genes
DUK	996	92	517
DUC	640	101	465
DU6	752	127	546
DU6P	783	109	534
DUhLB	1970	176	1027

Private variants for FZTDU not included as this line was unselected.

Table S4. Counts per length up to the 90% most frequent INDELS sorted in decreasing order of frequency

LENGTH	COUNT	PRCT	CLASS	CUMSUM_ PRCT
-1	174467	22.76	deletion	22.76
1	173290	22.6	insertion	45.36
-2	62514	8.15	deletion	53.51
2	59929	7.82	insertion	61.33
-3	32609	4.25	deletion	65.58
3	30779	4.01	insertion	69.6
-4	29781	3.88	deletion	73.48
4	29271	3.82	insertion	77.3
-5	12315	1.61	deletion	78.91
5	12167	1.59	insertion	80.5
6	10170	1.33	insertion	81.82
-6	9875	1.29	deletion	83.11
-7	8919	1.16	deletion	84.27
-8	8512	1.11	deletion	85.38
7	8360	1.09	insertion	86.47
8	7651	1	insertion	87.47
-9	6098	0.8	deletion	88.27
-10	5838	0.76	deletion	89.03
9	5403	0.7	insertion	89.73
10	5147	0.67	insertion	90.41
-12	4400	0.57	deletion	90.98

Table S5. Significantly enriched terms based on RDD gene lists

Line	Term/Pathway	Genes	FDR
DUK	mmu04072: Phospholipase D signaling pathway	Raf1 Adcy6 Grm8 Tsc1 Ralgds	0.0082875
DUC	GO:0009755: hormone-mediated signaling pathway	Pias2 Pgr Rxfp1 Yap1	0.030545
DUC	GO:0048814: regulation of dendrite morphogenesis	Pias2 Trpc6 Skor2	0.074441
DUC	GO:0030518: intracellular steroid hormone receptor signaling pathway	Pias2 Pgr Yap1	0.074441
DUhLB	mmu00340: Histidine metabolism	Aldh3a1 Aldh3a2	0.099122
DUhLB	mmu00410: beta-Alanine metabolism	Aldh3a1Aldh3a2	0.099122

Significantly enriched GO terms and pathways at FDR < 0.1

Table S6. Proportion of line-specific fixed and polymorphic structural variants in genic regions

	Fixed		Polymorphic	
	Number	Length (mean)	Number	Length (mean)
DUK	5	13.58 (2.7) kbp	15	204.62 (13.64) Mb
DUC	2	2.19 kbp (1.1 kbp)	34	186.92 (5.5) Mb
DU6	4	7.75 (1.94) kbp	14	302 (21.57) kb
DU6P	7	29.11 (4.65) kbp	17	129.96 (7.6) Mb
DUhLB	6	11.16 (1.9) kbp	14	3.72 (0.265) Mb
FZTDU	1	1.14 kbp	8	14.09 (1.8) Mb

Table S7. Types and lengths of line-specific fixed structural variants in genic regions

	DEL	DEL length	DUP	DUP length	INV	INV length
DUK	5	13.6 Kb	--	--	--	--
DUC	1	1.3 Kb	--	--	1	0.923 Kb
DU6	4	7.7 Kb	--	--	--	--
DU6P	3	8.5 Kb	1	11 Kb	3	9.6 Kb
DUhLB	4	3.3 Kb	--	--	2	7.9 Kb
FZTDU	1	1.1 Kb	--	--	--	--

Table S8. Number of genes affected by line-specific fixed and polymorphic structural variants

	Fixed			Polymorphic		
	DEL	DUP	INV	DEL	DUP	INV
DUK	5	--	--	4	28	1694
DUC	1	--	1	9	38	1363
DU6	4	--	--	6	--	7
DU6P	3	1	3	11	--	1130
DUhLB	4	--	2	11	3	7
FZTDU	1	--	--	3	--	266

Table S9. Number of genes in functional groups affected by line-specific structural variants

	DUK	DUC	DU6	DU6P	DUhLB	FZTDU
Reproduction	8	13	1	1	3	1
Metabolism/Energy conversion	--	27	3	16	8	--
Immune system	--	3	1	4	1	13
Nervous system	5	13	1	3	4	1
Cardiovascular system	2	1	1	1	2	15
Endocrine system	2	2	2	1	2	--
Sensory perception	297	36	2	--	3	--
Other (cell cycle, transcription)	2	88	3	15	1	2

Table S10. Summary of structural variants detected in low and high coverage variant calling sets for each mice line.

	Low-coverage set				High-coverage set			
	DEL	DUP	INV	Total	DEL	DUP	INV	Total
DUK	1965	11	81	2057	3548	31	515	4094
DUC	1693	6	36	1735	4897	51	1241	6189
DU6	63	3	5	71	5014	28	554	5596
DU6P	2735	10	104	2849	2716	21	2063	4800
DUhLB	1992	6	87	2085	3755	21	1481	5257
FZTDU	2408	6	107	2521	2454	12	540	3006

Table S11. Number of SNP sites per window analyzed with F_{ST}

Contrast	Number of windows	Mean sites/window	SD sites/window	Min sites/window	Max sites/window
DUK_FZTDU	70621	125.03	115.02	10	1274
DUC_FZTDU	71172	124.42	114.54	10	1274
DU6_FZTDU	71036	124.16	115.1	10	1274
DU6P_FZTDU	71633	125.71	115.51	10	1274
DUhLB_FZTDU	71222	125.38	115.45	10	1274
FZTDU_FZTDU	72702	125.99	115.88	10	1274

“There is no knowledge that is not power.”

Mortal Kombat III