SINGLE-HAPLOTYPE GENOME ASSEMBLIES REVEAL REMARKABLE RATES OF X-LINKED SATELLITE EVOLUTION AND A ROLE IN REPRODUCTIVE ISOLATION

A Dissertation

by

KEVIN ROSS BREDEMEYER

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	William J. Murphy
Committee Members,	Heath Blackmon
	Scott V. Dindot
	Terje Raudsepp
	Paul B. Samollow
Interdisciplinary Faculty Chair,	David W. Threadgill

August 2021

Major Subject: Genetics

ABSTRACT

Complex and highly repetitive regions are absent from nearly all genome assemblies and are often referred to as genomic "dark matter". Many of these regions also harbor large copy number repeats and/or other structural differences within and between species, further confounding their assembly from a diploid genome, and eluding our understanding of their biological properties. We identified and characterized one such region, the macrosatellite *DXZ4*, as a candidate hybrid-sterility locus in a cat interspecific hybrid cross. Previous investigations of *DXZ4* have been limited to female somatic cells, where it plays a role in structural organization of the inactive X chromosome. We demonstrate that *DXZ4* is transcriptionally active in germ cells undergoing meiotic sex chromosome inactivation and reveal that divergence across this macrosatellite results in perturbed methylation and ncRNA expression in testes of male interspecific hybrids. Due to its structurally complex nature, *DXZ4* was incomplete in the human reference genome and nearly all other mammalian genomes, limiting insights into its structure and functional evolution. Trio binning was developed to sort and independently assemble the divergent parental haplotypes of an F1 hybrid using a combination of short-read and long-read sequence data from the parents and hybrid. Here we applied this method to three feline F1 hybrids to generate six ultra-continuous single-haplotype genome assemblies from five species: domestic cat, Asian leopard cat, Geoffroy's cat, tiger, and lion. For the first time, we were able to fully resolve X-linked macrosatellites in other mammal species and thereby discovered the feline *DXZ4* satellite is composed of a compound tandem repeat with two distinct and highly divergent repeat arrays (A and B). Interestingly, the presence and organization of the two felid *DXZ4*

repeat arrays have diverged rapidly across mammalian orders, revealing a far greater scope of complexity and divergence than previously appreciated, especially for a locus involved in a conserved developmental process like X-chromosome inactivation (XCI). As ultracontinuous genomes become available for a wider variety of organisms, “dark matter” regions previously missing from genomes might hold the key to answering pervasive questions in disease biology, genome organization, gene regulation, and speciation.

ACKNOWLEDGEMENTS

I would like to begin by acknowledging Bill Murphy, who in his lab has established a perfect environment for higher learning. Despite being a top tier scientist at the cutting edge of his field, he shows nothing but respect and compassion for everyone he mentors and is the gold-standard for how an advisor should operate. He has been a true pleasure to work with and has contributed massively to both my growth of genetic knowledge, as well as wisdom in matters of life and family, and for that I am truly thankful.

I would also like to thank the entire Murphy lab. While I could go into MUCH detail regarding the special qualities of every member in the interest of keeping it brief I simply say, thanks for always being supportive and willing to discuss matters of science and life over a few beers.

I would also like to thank my parents, Malcolm and Sharen. They both worked hard to provide me with every opportunity to succeed and were always nothing but supportive when it came to choosing a path in life. They are the best role models for how to be a parent and a friend and I am eternally grateful for all they have done for me. Finally, I would like to thank Hannah Bredemeyer, my beautiful wife. I can only guess that taking a pay cut and starting grad-school wasn't on her list of priorities right after getting married, but she took it all in stride and was there to support me every step of the way. While I could go on forever singing her praises, I'll just say I love you, and am proud of the woman and mother you have become, and the man you inspire me to be.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This dissertation was completed under the supervision and guidance of the following committee members: William J. Murphy (Advisor), Department of Veterinary Integrative Biosciences, Faculty of Interdisciplinary Program in Genetics; Heath Blackmon, Department of Biology, Faculty of Interdisciplinary Program in Genetics; Scott V. Dindot, Department of Veterinary Pathobiology, Faculty of Interdisciplinary Program in Genetics; Terje Raudsepp, Department of Veterinary Integrative Biosciences, Faculty of Interdisciplinary Program in Genetics; Paul B. Samollow, Department of Veterinary Integrative Biosciences, Faculty of Interdisciplinary Program in Genetics. Section pertaining to "Assembly Quality Control" and Figure 2 of Chapter II was generated by Andrew Harris. Chapter III RNA-Seq libraries were generated in conjunction with Jinhong Wang. The hybrid sterility genome-wide association study in Chapter III was performed by Christopher Seabury. Fine mapping of the *DXZ4* macrosatellite in Chapter III was performed by Gang Li and William Murphy. Germ cell sorting in Chapter III was performed by John McCarrey. Analysis of RRBS data in the Chapter III was conducted by Bridgett vonHoldt. The F1 liger cell lines used to generate high molecular weight DNA for assembly in Chapter IV were provided by Brian Davis. The cytogenetic analysis of F1 Safari Cat cell lines in Chapter IV was performed by Terje Raudsepp. All other work conducted for this dissertation was independently completed by Kevin R Bredemeyer, with significant input, guidance, and

consideration from William J. Murphy.

Funding Sources

This work was funded by grants from the U.S. National Science Foundation, the Morris Animal Foundation, and the Texas A&M University College of Veterinary Medicine.

NOMENCLATURE

BSC	Biological species concept
BDM	Bateson, Dobzhansky, Muller
DE	Differential expression
FACS	Fluorescence-activated cell sorting
FISH	Fluorescence *in situ* hybridization
GWAS	Genome-wide association study
MSCI	Meiotic sex-chromosome inactivation
MSUC	Meiotic silencing of unsynapsed chromatin
MYA	Million years ago
PAR	Pseudoautosomal region
PMSC	Post-meiotic sex chromatin
PSCR	Post-meiotic sex chromatin repression
SHA	Single haplotype assembly
QTL	Quantitative trait locus
T2T	Telomere-to-telomere
TAD	Topologically associated domain
VNTR	Variable-number tandem repeat
XCI	X-chromosome inactivation
Xa	Active X chromosome
Xi	Inactive X chromosome

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I INTRODUCTION

1.1 Speciation and Hybrid Sterility

Evolutionary biologists have long pondered how and why ancestral groups of genetically similar organisms radiated into the multitudes of species that constitute today's global biodiversity. The study of speciation focuses on how evolutionary processes drive divergence across the genome that result in distinct populations of organisms. The Biological Species Concept (BSC) proposed that speciation occurs as a consequence of restricted gene flow resulting from reproductive isolation (Dobzhansky, 1937; Mayr, 1942). Mechanisms of reproductive isolation are divided into pre and post-zygotic reproductive barriers with the latter frequently manifesting in the hybrid offspring produced from successful fertilization by allospecific parental gametes.

Hybrids are often infertile or inviable as a result of genetic incompatibilities between parental haplotypes (Presgraves, 2010). These defects are not equally represented in both sexes, as explained by Haldane's Rule, the long-standing observation of preferential sterility or inviability of the heterogametic sex resulting from an interspecific mating (Haldane, 1922). Hybrids utilizing the X-Y sex determining system also exhibit what is termed the large X-effect, i.e., the observation that the X chromosome is enriched for hybrid sterility factors and plays an increased role in post-zygotic isolation relative to the autosomes (Figure 1A) (Coyne, 1992; Presgraves, 2018). Together these patterns constitute the "two rules of speciation" and form a conserved framework for features of hybrid sterility across a diverse array of animal lineages (Masly & Presgraves, 2007; Coyne, 2018). How interspecific divergence manifests as hybrid sterility is best described by the Bateson-Dobzhansky-Muller

(BDM) model, which suggests that independently derived epistatic relationships acquired in isolated populations, can manifest as interspecific developmental incompatibilities and defective phenotypes (sterility or inviability) when present in a single zygote, as in a hybrid individual (Dobzhansky, 1937; Muller, 1942; Orr, 1996).

While traditionally described in *Drosophila*, more recent *s*upport for the large X-effect and BDM incompatibilities in mammals have been observed through studies using genome-wide introgression analysis (Masly & Presgraves, 2007; Good et al., 2008). This method uses an initial interspecific mating and subsequent backcrossing to generate lines with foreign X chromosomal regions substituted into an otherwise native genomic background. The phenotypic outcome of individuals carrying an interval of the foreign X then allows correlation of specific regions of the genome with hybrid sterility (Figure 1B). Recent work employing these introgression lines and modern genetic mapping studies of sterility phenotypes to the X chromosome using genome-wide association studies (GWAS), QTL and eQTL has allowed fine mapping of regions and loci along the X responsible for hybrid sterility (Good et al., 2008; Bhattacharyya et al., 2014; Turner & Harr, 2014; Turner et al., 2014; Morán & Fontdevila, 2014; Davis et al., 2015; Balcova et al., 2016; Schwahn et al., 2018; Lustyk et al., 2019). Interestingly, many of these candidate regions fail to explain the sweeping pattern of Haldane's Rule and the large X-effect observed across sexually reproducing species.

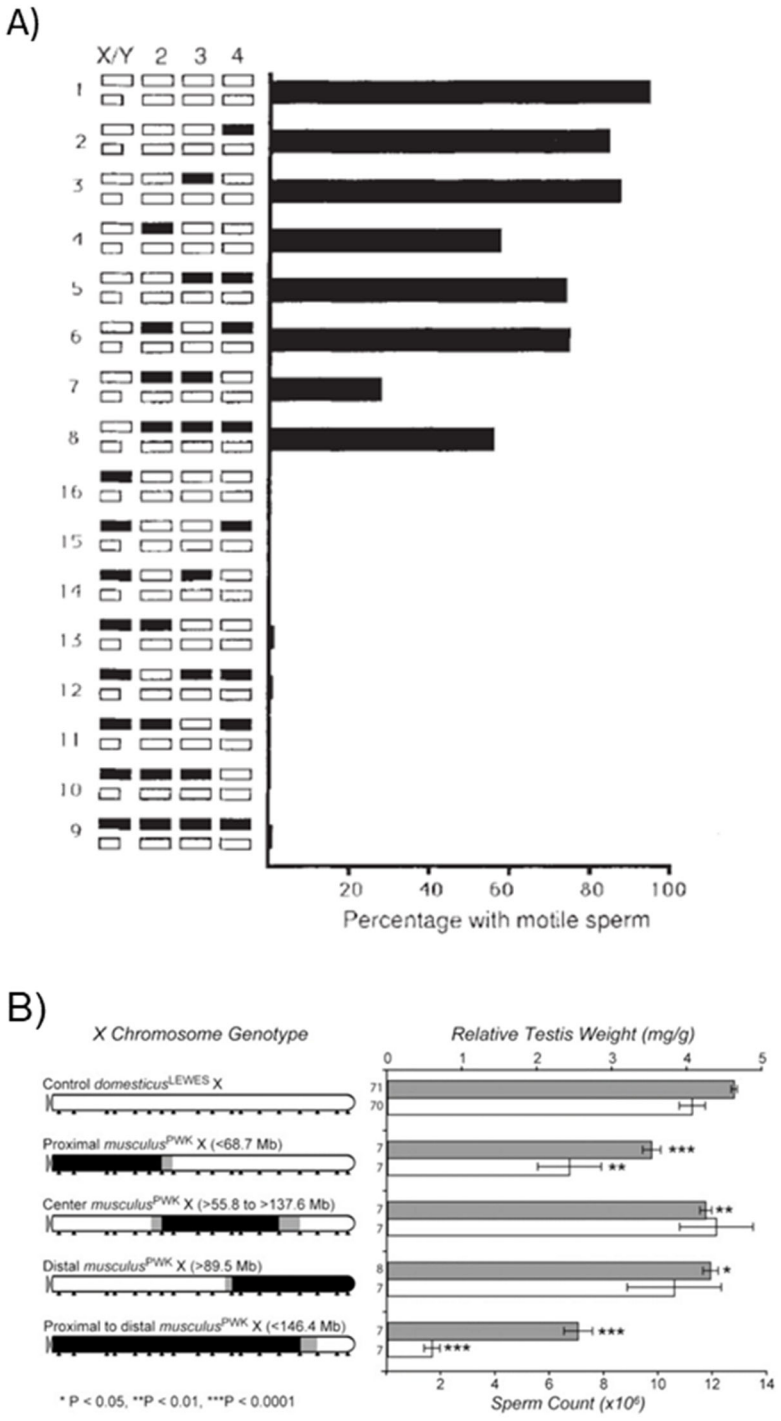


Figure 1. Evidence for the large-X effect

A) Interspecies substitution of the X chromosome yields the largest effect on sterility phenotypes in *Drosophila* (Adapted from Coyne, 1992). B) X chromosome introgression for fine mapping of genetic features effecting sterility in mice (Adapted from Good et al., 2008).

1.2 Felids as a Model System for Comparative Genomics

While mouse has been a useful model for probing mechanisms of speciation and hybrid sterility, it represents only a single mammalian lineage and possess a highly rearranged X chromosome relative to most other mammals, which have maintained conserved linkage across highly divergent clades (Rodriguez Delgado et al., 2009; Proskuryakova et al., 2017). To determine whether established patterns of hybrid dysfunction are conserved across other mammalian lineages and broaden our understanding of hybrid sterility and speciation, our lab utilizes hybrid cat breeds as a model system. The cat lineage is a large, widely dispersed, mammalian radiation with a historical precedent for genome analysis in the domestic cat (Figure 2A). Domestic cat is one of the most popular companion animals globally, which translates into veterinary medical interest that contributes to its use as a biomedical model (O'Brien et al., 1982; 2002). Considerable effort has been expended to generate a high-quality domestic cat assembly, which has proved integral to conservation biology and evolutionary studies in wild cats (e.g., Johnson et al. 2006; O'Brien et al. 2006; Luo et al. 2008; O'Brien et al. 2017; Abascal et al. 2016; Zhang et al. 2019). Felids in general possess several genomic characteristics that make them ideal for genetic analysis including high recombination rates and strong collinearity between divergent genomes (Wurster-Hill & Centerwall, 1982; Modi & O'Brien, 1988; Davis et al., 2009). The latter characteristic facilitates widespread hybridization within and between lineages (Li et al., 2016; 2019).

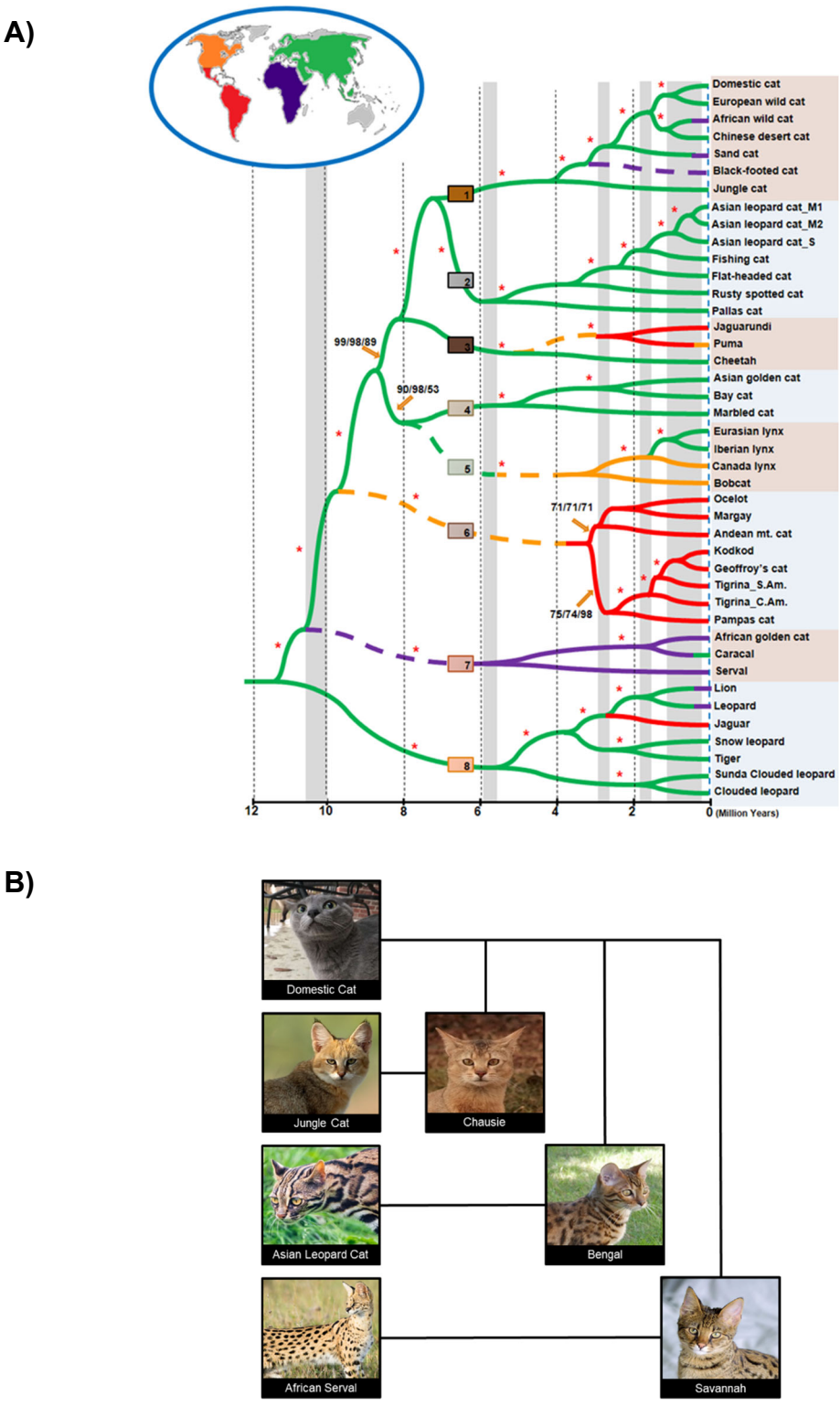


Figure 2. Hybrid cat breeds as a model system for hybrid sterility

A) Phylogenetic and geographic associations between felid species (Adapted from Li et al., 2016). B) Visual representation of parental cat species and their hybrid progeny.

Cat hybrids occur naturally between wild populations and through circumstances of human intervention such as mutual captivity and controlled breeding (Gray, 1972; Schwartz et al., 2004; Homyack et al., 2008; Trigo et al., 2008, 2013; Davis et al., 2015; Li et al. 2016b; Figueiro et al. 2017; Li et al. 2019). Historically, interspecific hybrids have proven invaluable for investigation of divergence and its effect on interaction within and between genomes. The most common hybrid felid breeds include the Chausie, Bengal, and Savannah, which are crosses between a domestic cat (Felis catus) and different wild cat species: Jungle cat (*Felis chaus*), Asian leopard cat (*Prionailurus bengalensis*), and African serval (*Leptailurus serval*), respectively (Figure 2B). Each of these wildcat parent species differ in divergence time from the domestic cat and produce hybrid offspring obeying Haldane's Rule. These attributes, in conjunction with the high recombination rate of felids, and global popularity of these breeds make hybrid cats an attractive model for investigating mechanisms underlying hybrid male sterility and speciation across mammals.

1.3 The Mammalian X Chromosome

Apart from involvement in hybrid sterility, the mammalian X chromosome is distinct from autosomes in many respects. First is the disparate inheritance between males and females. Second, male-specific hemizygosity requires silencing of the X chromosome under two vastly different circumstances, one female-specific and one male-specific.

1.3.1 X Chromosome Inheritance and Evolution

In mammals, females inherit a homologous pair of X chromosomes, while males inherit a single X chromosome and the sex-determining Y chromosome. Once homologs, the X and Y chromosome became morphologically dissimilar through Y-chromosome degradation after acquisition of a sex-determining locus and subsequent suppression of

recombination, driving recurrent loss of Y chromosome sequence (Lahn & Page, 1999; Graves, 2006; Liu, 2019). In males, this disequilibrium results in hemizygosity across much of the X chromosome, with implications in evolution, gene regulation and reproduction. The evolutionary impact of X chromosome hemizygosity, which leaves most genes on the X unprotected by the masking effects of heterozygosity, manifests as increased divergence of the sex chromosomes between populations driven by direct exposure of uncovered genes to selection in males. This phenomenon is detailed in the "Dominance Theory", "Faster X" and "Faster Male" hypotheses and is believed to at least partially explain the patterns described by the "two rules of speciation" mentioned above (Turelli & Orr, 1995; Presgraves, 2008; Meisel & Connallon, 2013; Delph & Demuth, 2016; Charlesworth et al., 2018).

1.3.2 X-Chromosome Inactivation

The XX, XY sex-determining pattern results in a disparity of gene dosage between the sexes, which is resolved in therian mammals (marsupials and eutherians) by a mechanism known as X-Chromosome Inactivation (XCI) **(**Figure 3A**)** (Lyon, 1961; Ohno, 1967; Mahadevaiah et al., 2009; Brockdorff & Turner, 2015). XCI takes place in female somatic cells at different developmental stages depending on species (Okamoto, et al., 2011). The process of X inactivation requires the presence of the long non-coding RNA XIST, as well as a host of long-range chromatin interacting loci and epigenetic modifiers (Jégu et al., 2017). These loci work together to alter the structural, epigenetic, and transcriptional landscape of the inactivated X chromosome (Xi), ultimately resulting in the compact, largely heterochromatic, Barr body (Barr & Bertram, 1949; Chadwick & Willard, 2003). Upon inactivation, the X chromosome acquires a unique structural and positional arrangement, forming two large super-domains (Rao et al., 2014; Deng et al., 2015) and colocalizing with

the nucleolus and nuclear periphery, which is thought to govern the Xi epigenetic state (Bourgeouis et al., 1985; Dyer et al., 1989; Zhang et al., 2007). This new conformation is correlated with a depletion of chromatin modifying proteins and species-dependent reductions in chromatin compartmentalization compared to the transcriptionally active X chromosome (Xa) and autosomes (Minajigi et al., 2015; Giorgetti et al., 2016; Darrow et al., 2016). It has been suggested that this unique silencing mechanism has contributed to the remarkably strong conservation of gene content of X chromosomes across therian mammals (Ohno, 1967).

1.3.3 Meiotic Sex Chromosome Inactivation

While female XCI has been scrutinized for decades, less is known about the male-specific instance of X chromosome silencing resulting from meiotic sex chromosome inactivation (MSCI) (Figure 3B) (Lifshytz & Lindsley, 1972; Turner, 2007). As the name implies, MSCI occurs during meiosis I and is an important feature of spermatogenesis (Turner et al., 2006). Like XCI, MSCI is conserved in therian mammals and hypothesized to have evolved as a mechanism for avoiding arrest at meiotic checkpoints (Burgoyne, 1982; Daish et al., 2015). Throughout meiosis, checkpoint proteins monitor proper synapsis of homologous chromosomes. When chromosomes become heteromorphic, they lose homology along much of their length and no longer synapse, normally resulting in meiotic arrest and germ cell apoptosis (Burgoyne et al., 2009).

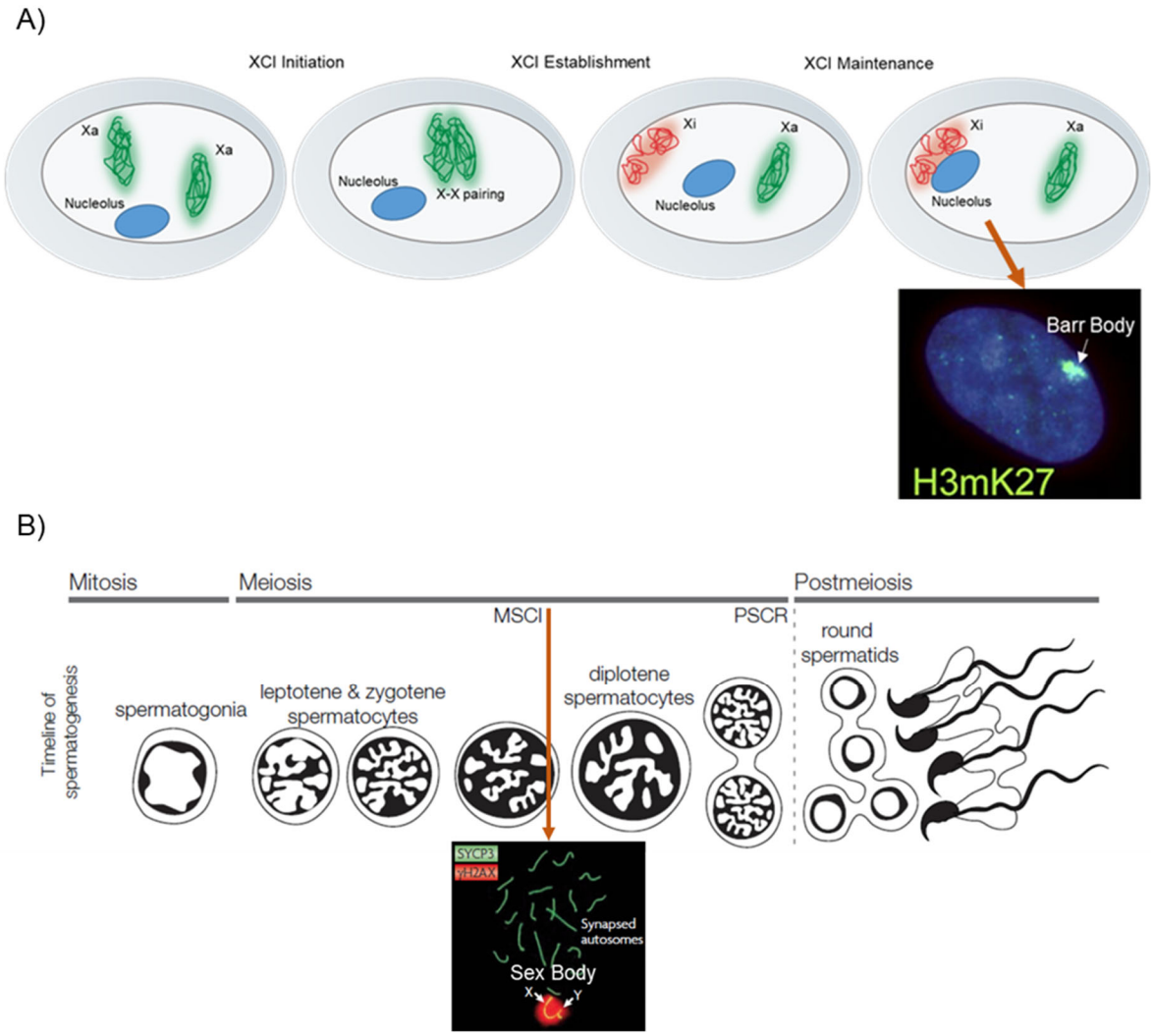


Figure 3. Overview of two sex-specific instances of X chromosome inactivation

A) The process of X chromosome inactivation (XCI) takes place during early development in female somatic cells as a form of dosage compensation. B) Meiotic sex chromosome inactivation (MSCI) takes place during pachynema in male germ cells in response to asynaptic sex chromosomes (Adapted from Larson et al., 2016).

MSCI is a specific version of a more general mechanism termed meiotic suppression of unsynapsed chromatin (MSUC), which silences any regions of asynapsis between otherwise homologous molecules (Schimenti, 2005; Turner et al., 2005). While both mechanisms occur during pachynema of meiosis I, MSUC refers to silencing of any high sequence divergence or structural variation between otherwise homologous chromosomes (Wang & Höög, 2006), while MSCI is specific to the heteromorphic nature of sex chromosomes (McKee & Handel, 1993). In pachytene spermatocytes, the sex chromosomes are subject to MSCI and partitioned into a heterochromatic XY body (Solari, 1974; Handel, 2004). Similar to the Barr body in XCI, the XY body is isolated from the autosomes through association with the nucleolus and maintained throughout spermatogenesis as post-meiotic sex chromatin (PMSC) (Kierszenbaum & Tres, 1974; Knibiehler et al., 1981; Namekawa et al., 2006; Turner, 2006).

Despite the analogous transcriptional and spatial characteristics of the X chromosome in both female somatic and male meiotic instances of inactivation, few studies have attempted to investigate structural and functional links between the two processes. In early studies of MSCI, expression of *XIST*, the key regulator of female X inactivation, was detected in pachytene spermatocytes (McCarrey & Dilworth, 1992). However, a more recent study involving truncation of *XIST* RNA during meiosis, suggests different *XIST* functional domains may be uniquely required for the two processes (Turner, 2002). An additional study comparing histone modifications between the two inactivated states has described exclusive histone methylation patterns between the two cases, indicative of different upstream mechanisms for chromosome silencing (Armstrong et al., 1997). Despite the limited number of studies, several aspects regarding the structural modification of the X chromosome during

XCI and its importance in chromosomal silencing are just beginning to emerge (Finestra & Gribnau, 2017). We hypothesize that undescribed connections between XCI and MSCI exist, and that structural conformation and epigenetic regulation of specific loci associated with chromatin modification may have evolved in the therian ancestor, during the nascent stages of dosage compensation evolution.

1.4 MSCI in Hybrid Sterility

In mammalian hybrids, the testes of sterile males often exhibit two conserved phenotypes or "biomarkers" of sterility. The first is meiotic arrest of spermatogenesis at pachynema, which manifests histologically as vacuolization, accumulation of pachytene spermatocytes and depletion of post-pachynema germ cells. This has been observed in the testes of felid hybrids, as well as other mammalian species (Moore et al., 1999; Thomsen et al., 2011; Bhattacharyya et al., 2013; Davis et al., 2015; Ishishita et al., 2015). The second phenotype is widespread upregulation of X-linked genes in hybrid testes, which contrasts with near-global gene repression via MSCI that is characteristic of non-hybrid testes. First seen in murine testes, X upregulation was later observed across domestic cat hybrids suggesting that this phenomenon might be a general one in sterile mammalian hybrids from divergent orders and could point to a common mechanism of hybrid fertility loss across mammalian lineages. One proposed mechanism linking these two observations involves failure of proper sex chromosome conformational change during gametogenesis (Lifschytz & Lindsley, 1972; Jablonka & Lamb, 1991). Previous studies of subspecific mouse crosses have suggested that this "conformation failure" results in disruption of MSCI, which manifests as meiotic arrest and misregulation of the sex chromosomes in the testes of sterile hybrids. (Good 2010, Larson 2016). In mouse, misexpression was targeted to specific

meiotic stages using fluorescence-activated cell sorting (FACS) to obtain enriched germ-cell populations. These results revealed that X upregulation occurred in all investigated stages of spermatogenesis but were most extreme in cell stages normally exhibiting sex chromosome silencing due to MSCI (Campbell et al., 2013; Larson et al., 2016). While failure of MSCI and post meiotic sex chromatin repression (PSCR) has been linked to pachytene arrest and male sterility in multiple mammalian lineages (Burgoyne et al., 2009; Royo et al., 2010), a mechanism describing failure of MSCI and X upregulation in the context of hybrid incompatibility is currently lacking.

1.5 Hybrid Sterility Candidate *DXZ4*

Chapter III of this dissertation begins with exploration of hybrid male sterility in an interspecific hybrid cat breed, the Chausie, where we discovered a novel mammalian X-linked candidate hybrid sterility locus, *DXZ4*, using a combination of GWAS and ancestry-based fine mapping. *DXZ4* is a variable number tandem repeat (VNTR) DNA sequence that transcribes long non-coding RNAs (lncRNAs) and smallRNAs of unknown function (Tremblay et al., 2011; Pohlers et al., 2014). Until the release of the human X chromosome telomere-to-telomere assembly (Miga et al., 2020), mouse was the only mammal with *DXZ4* fully resolved and accurately represented. This is likely due to the smaller and less complex structure of murine *DXZ4*, which is composed of 7 repeat monomers of varying length (3.8 and 5.7 kb), while human *DXZ4* contains between 12 and 120 repeat monomers of very similar length (3.0 kb) (Tremblay et al., 2011; Horakova et al., 2012a, Schaap et al., 2013). In humans the *DXZ4* region was described as unstable and highly polymorphic due to multiple alleles for three distinct microsatellites within individual monomers and a high degree of meiotic instability (Tremblay et al., 2011). Despite this variability, highly

conserved CTCF binding sites were observed in monomers of both mouse and human. CTCF is a transcription factor protein and the main insulator responsible for partitioning of chromatin domains in vertebrate genomes (Ong & Corces, 2014). This partitioning results from the mechanisms of chromatin loop extrusion, where bound CTCF proteins act as 'stoppers' for chromatin extruded through cohesin rings and underlies a majority of small-scale topologically associated domains (TADs) (Rao et al., 2014; Rao et al., 2017). Further, clustering of CTCF binding motifs, as observed in the *DXZ4* locus, has been shown to increase stability of higher-order chromatin structure relative to isolated sites (Kentepozidou et al., 2020). Repeat copy number of *DXZ4* is highly polymorphic in human populations and is predicted to differ between felid species. *DXZ4* is one of many large tandem repeat loci defined as macrosatellites (Giacalone et al, 1992). Several macrosatellites have been described in humans and share similar features, such as high GC content, large repeat monomers, high variability in copy number, and unique presence at one or two chromosomal locations (Giacalone et al., 1992; Hewitt et al., 1994; Gondo et al., 1998; Dumbovic et al., 2017). An interesting pattern documented in *DXZ4* and another human macrosatellite, *D4Z4*, is the methylation profile of their abundant CpG islands. Both macrosatellites possess a hypermethylated state relative to that of surrounding chromatin (Chadwick, 2009). In *D4Z4*, this state is reversed when the total number of repeat units fall below a certain threshold, leading to aberrant upregulation ultimately leading to facioscapulohumeral muscular dystrophy (Hewitt et al., 1994, van Overveld et al., 2003). The transition between methylation states illustrates the concept that variation in copy number of repeat arrays can cause changes in the surrounding chromatin with downstream epigenetic effects (Chadwick, 2009). While important because of their association with disease processes and speciation,

macrosatellites are incompletely resolved or altogether absent from a majority of genome assemblies due to their complex nature.

1.6 Modern Genome Assembly

Repetitive DNA makes up an estimated 70% of the human genome and poses challenges for genome assembly utilizing both short- and long-read sequences. These challenges stem from the inability of many reads to span long repetitive regions in their entirety, making gapless assembly impossible. While the advent of long-read sequencing by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) helped to remedy this problem and closed many of the gaps associated with less expansive repeats, base-pair level accuracy was poor relative to Illumina short-read data (Pollard et al., 2018). To compensate, alignment programs were developed to "polish" inaccurate long-read assemblies using more accurate Illumina reads. Repetitive regions that also exhibit large or complex allelic differences in the parental haplotypes of diploid genomes present another problem that cannot be overcome by longer reads alone because modern consensus assemblers would still be confounded by divergent versions of a homologous region. As a result, complex and highly repetitive regions are absent from nearly all genome assemblies and are often referred to as genomic "dark matter". Such regions are ignored at our peril, because the same features that confound assembly are increasingly understood to have important roles in disease biology, genome organization, gene regulation, and speciation.

The only way to avoid this problem at present is to use reads representing a single haplotype as input for *de novo* assembly. This was initially accomplished in humans using the essentially haploid hydatidiform mole cell-line CHM13. This haploid DNA source combined with long read data was used to great success, resulting in the first telomere-to-

telomere (T2T) assembly of the human X chromosome followed by human chromosome 8 (Miga et al., 2020; Logsdon et al., 2020). While generation of the human T2T assembly was proof of concept for the contiguity achievable using a haploid data source, similar biological resources are not readily available for most other species of interest.

A new bioinformatic methodology referred to as trio binning was developed to sort and independently assemble divergent parental haplotypes from F1 hybrids (Koren et al., 2018) (Figure 4A). This method utilized a combination of Illumina short- and long-read sequence data from both the parents and hybrid, and output single haploid assemblies for bovid F1 hybrids representing both inter- and subspecific trios (Koren et al., 2018; Rice et al., 2020; Low et al., 2020). Assemblies generated using this approach not only dramatically improved the existing reference genome, but also yielded novel *de novo* assemblies of closely related species. Highly continuous assemblies like these allow interspecific comparison between large and complex regions of interspecific divergence often missing from both long- and short-read diploid assemblies but increasingly understood to play important roles in disease biology, genome organization, gene regulation, and speciation (Hsieh et al., 2019; Vollger et al., 2019; Miga et al., 2020). As described in this dissertation, we applied the trio-binning method to three F1 hybrid cats, the Bengal, Safari, and Liger and generated six highly continuous single haplotype parental assemblies that include two domestic cats, Asian leopard cat, Geoffroy's cat, Tiger and Lion (Figure 4B). While not completely gapless, these assemblies are superior to previous iterations and on-par with their human counterparts.

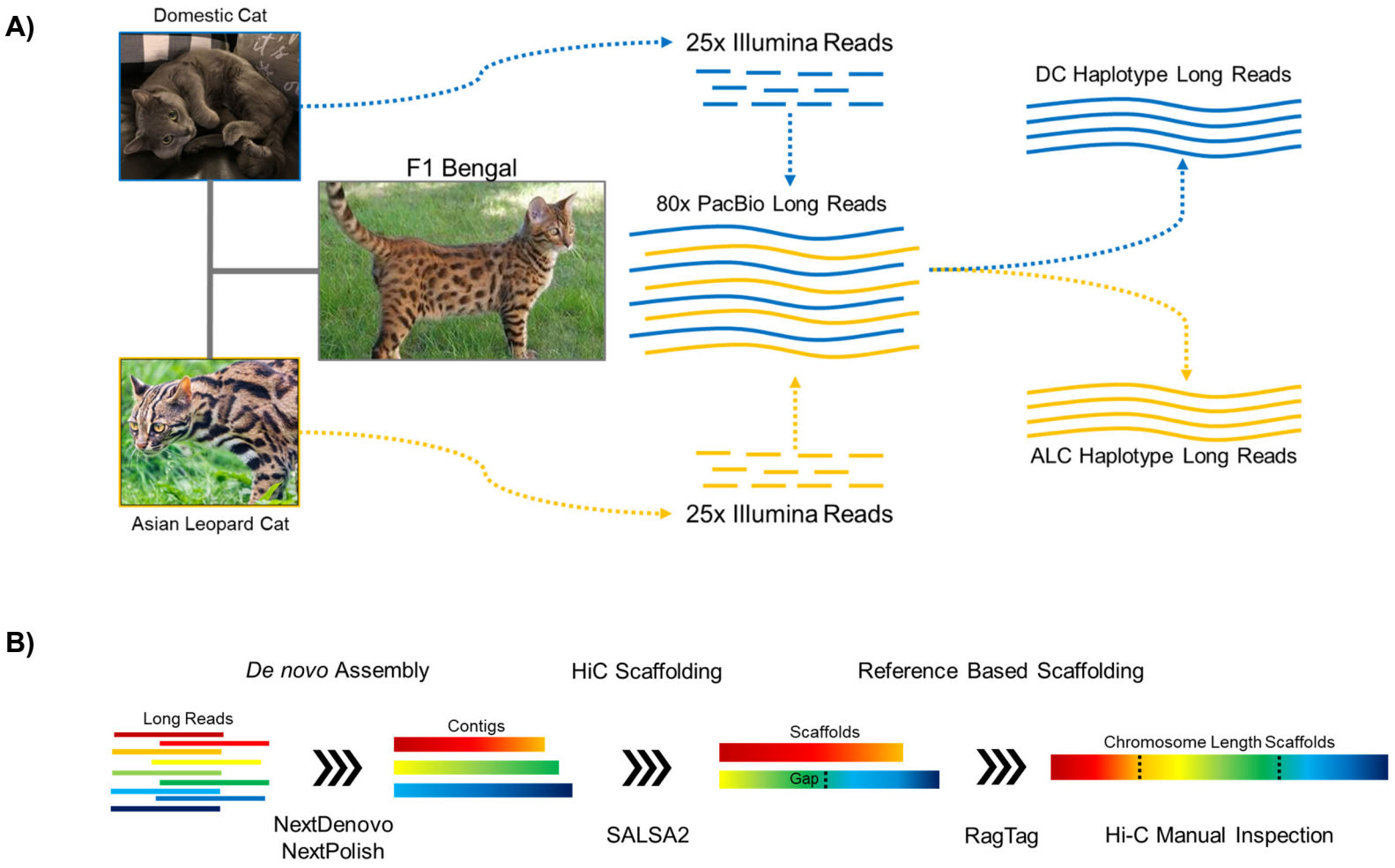


Figure 4. Assembly techniques applied to felid F1 hybrids to obtain ultracontinuous single haplotype assemblies

A) A cartoon outlining the trio-binning method used to identify and partition long reads from an F1 hybrid into parental haplotypes. B) Overview of assembly pipeline used to generate the six chromosome-length single haplotype felid assemblies.

Many regions previously missing from the felid assembly were fully resolved including *DXZ4*, which differed significantly from both human and mouse orthologs.

1.7 *DXZ4* Function

A key characteristic of *DXZ4* that makes it an interesting candidate for structural misregulation of the X chromosome during MSCI is its association with XCI in female somatic cells. *DXZ4* was first connected to human XCI when Giacalone (1992) noticed a large hypomethylated region on the otherwise hypermethylated and heterochromatic Xi. Subsequent investigations of this macrosatellite's epigenetic profile revealed that methylation of the region was also regulating regional binding of the CTCF protein, which acted as a unidirectional insulator on the Xi copy of the locus (Chadwick, 2008). In mouse this pattern is reversed with *DXZ4* hypomethylated on the Xa only and CTCF bound biallelically (Horakova et al., 2012a, Minajigi et al., 2015). In humans, *DXZ4* transcription was detected from both the Xa and Xi alleles, with subsequent analysis revealing two distinct lncRNA transcripts (Tremblay et al., 2011; Figueroa et al., 2015). The genes were labeled *DXZ4*-associated noncoding transcript proximal and distal (*DANT1* and *DANT2*, respectively), and produced alternate isoforms corresponding to *DXZ4* chromatin state. While array traversing transcripts (ATT) from both genes were associated with the euchromatic *DXZ4* state on the Xi, the *DANT1-ATT* isoform alone exhibited Xi specific expression. In addition to lncRNA, several classes of smallRNAs including siRNAs, miRNAs and piRNAs are transcribed from individual repeat monomers and hypothesized to mediate *DXZ4* methylation through an Argonaute dependent pathway (Pohlers et al., 2014). In 2014, while investigating the basis for three-dimensional chromatin architecture in humans, Rao et al. inadvertently observed that the Xi was organized into a unique "bipartite"

chromatin structure relative to the Xa. Composed of two large super-domains with *DXZ4* acting as the hinge region, this structure also appears in cat and mouse despite dramatic divergence of the murine *DXZ4* and X chromosome (Figure 5) (Deng et al., 2015; Brashear et al., 2021). In addition to the formation of super-domains, Rao (2014) and Darrow (2016) have both described the formation of Xi-specific super-loops anchored at *DXZ4* and other macrosatellite sequences along the chromosome. These loci include other X-linked loci, *FIRRE* and *ICCE*, both of which contain occupied CTCF motifs on the inactive X. These loci exhibit conservation of both linkage and spacing between mammalian lineages with structurally divergent X chromosomes, suggesting interaction between them is physically constrained and sensitive to perturbation of orientation and relative positioning (Brashear et al., 2021). Despite these observations, the primary function of *DXZ4* transcripts and *cis* genomic insulator effects of this region remain unknown. Subsequent knockout experiments of *DXZ4* and surrounding regions were concordant in showing that while deletion of *DXZ4* resulted in disruption of the Xi bipartite structure, the silenced state of the X chromosome remained unchanged, thus yielding no insight into *DXZ4* functionality beyond its known influence on Xi structural conformation (Giorgetti et al., 2016; Darrow et al., 2016; Bonora et al., 2018).

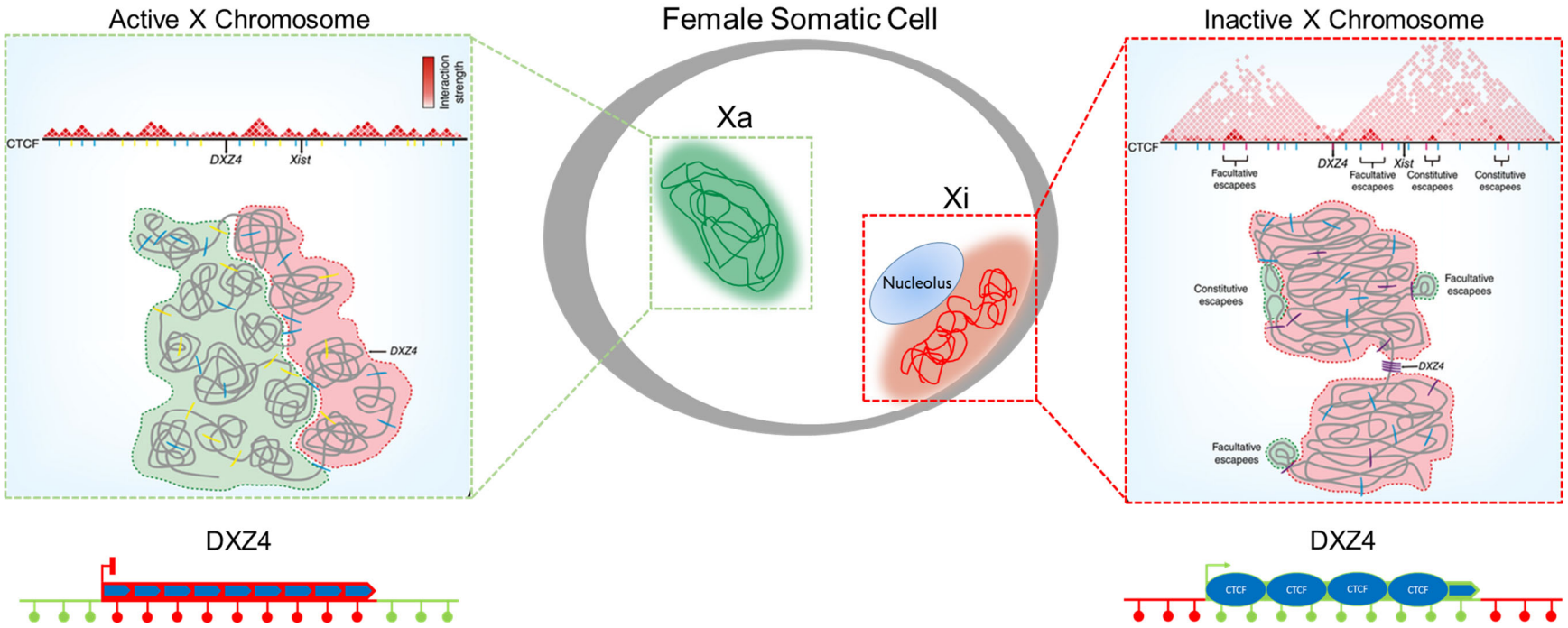


Figure 5. *DXZ4* in female X chromosome inactivation

DXZ4 acts as a central hub for structural organization of the inactive X chromosome by binding CTCF proteins and forming the hinge between distinct chromatin domains, forming the bipartite structure. (Adapted from Rocha & Heard, 2017, Nature Structural and Molecular Biology)

1.8 *DXZ4* in MSCI

While the link between *DXZ4* and architecture of the inactive X chromosome has been extensively studied in females, there have been no published studies describing what role, if any, *DXZ4* may play during spermatogenesis. Investigations into the structural conformations of X chromosomes at various stages of male meiosis, particularly those exhibiting dramatic chromatin modification and reorganization, have also been very limited. In mouse, TADs were depleted across all chromosomes as spermatogenesis progressed and DNA was packaged into heterochromatin (Alavattam et al., 2019; Patel & Kang, 2019). Despite loss of small-scale organization, large-scale A/B compartments were maintained across spermatogenesis in all but the X chromosome, which showed clear changes between meiotic stages, similar to differences observed between the Xa and Xi in females. In this dissertation we make similar observations for felids, supporting the idea of conservation of meiotic chromatin reorganization across divergent species. Additionally, modifications of A/B compartmentalization between the meiotically silenced X and pre-silenced X during spermatogenesis and between X states of female somatic cells suggest conservation of broader features of large-scale nuclear organization likely exist between MSCI and XCI. In Chapter III we generate transcriptional and epigenetic profiles of the locus during normal spermatogenesis and compare epigenetic profiles of fertile and sterile hybrid cat testes.

1.9 *DXZ4* Evolution

After observation of a *DXZ4* structure unique to felids in Chapter III, we sought to compare the *DXZ4* region across multiple divergent mammalian species and determine the ancestral state of the locus. Previously, the only interspecific comparisons of this region were those performed by Horakova et al. in 2012, mostly with Sanger-based and short-read

assemblies, where the macrosatellite was highly fragmented or wholly absent. In chapter IV we reassess DXZ4 structure and repeat units for 18 species representing four super-ordinal clades using updated, long read assemblies. While *DXZ4* remained unresolved across many taxa, making confident assumptions difficult, we felt this was an important step for profiling features that lend insight into the biological significance of this generally conserved, yet highly polymorphic, macrosatellite.

1.10 Motivation

At the onset of this project our primary motivation was to investigate a general mechanism for hybrid sterility following discovery of the unique candidate speciation locus, *DXZ4*, described in Chapter III. However, upon closer inspection of the candidate region we realized it was unresolved and likely collapsed in all current and previous domestic cat reference assemblies, complicating our analysis. Our first approach to resolving the locus involved long-read sequencing of bacterial artificial chromosomes (BACs) that overlapped or encompassed the entire *DXZ4* region as was done previously to resolve ampliconic sequence along the felid X and Y (Brashear et al., 2018; Brashear et al., 2021). While this approach yielded partial assembly of *DXZ4* and allowed observation of the first full repeat monomers in cats, previous estimates using short-read alignment and coverage pile-up around the assembly gap indicated the array was still underrepresented. Shortly after attempts to resolve the region using BACs, we sought to apply the trio binning method to the menagerie of felid hybrids available to us. The first was a female F1 Bengal, which resulted in the first felid single haplotype assemblies of both the domestic and Asian leopard cat and is documented in Chapter II. Much to our pleasure, *DXZ4* was fully resolved in both assemblies, as well as a long-read Jungle cat assembly. With a complete *DXZ4* locus, we

were able (in Chapter III) to generate *de novo* transcript annotations for the region in various stages of spermatogenesis, as well as in Jungle cat whole testes, representing the first observations of *DXZ4* activity during felid meiosis and allowing investigation of interspecific divergence between *DXZ4* transcripts, respectively. Following the revelation that *DXZ4* is likely involved in biological processes outside of female XCI and critical to the male-specific process of MSCI, in Chapter IV we sought to further investigate the structural evolution of *DXZ4* in a broader sampling of felid species. Additional comparisons included divergent mammalian lineages beyond the felidae in an effort to gain insight into the ancestral state of *DXZ4* and how its structural divergence between mammals is connected to biological function in distinct lineages.

CHAPTER II ULTRACONTINUOUS SINGLE HAPLOTYPE GENOME ASSEMBLIES FOR THE DOMESTIC CAT (*Felis catus*) AND ASIAN LEOPARD CAT (*Prionailurus bengalensis*)[*]

2.1 Introduction

The cat family Felidae is a speciose and geographically dispersed mammalian radiation, containing many of the most charismatic and endangered apex predators on Earth. Decades of genetic analysis of the species from this lineage have been driven by veterinary medical interest in the domestic cat, and its use as a biomedical model (O'Brien et al. 1982; 2002). Furthermore, wild felid species have benefitted from the advances in and applications of domestic cat genome assemblies to studying their conservation biology and evolutionary history (e.g., Johnson et al. 2006; O'Brien et al. 2006; Luo et al. 2008; O'Brien et al. 2017; Abascal et al. 2016; Zhang et al. 2019). Several attributes of felid genomes are ideal for comparative genetic analysis, including the strong chromosomal collinearity between all species (Wurster-Hill and Centerwall 1982; Modi & O'Brien 1988; Davis et al. 2009), and the highest reported rates of meiotic recombination within mammals (Menotti-Raymond et al. 2003; Segura et al. 2013; Li et al. 2016a). There is also an extensive body of literature describing prolific interspecific hybridization within and between the major clades of the cat family, both in free-ranging populations and through human controlled breeding (Gray 1972;

---

[*] Reprinted with permission from Bredemeyer KR, Harris AJ, Li G, Zhao L, Foley NM, Roelke-Parker M, O'Brien SJ, Lyons LA, Warren WC, Murphy WJ. 2021. Ultracontinuous Single Haplotype Genome Assemblies for the Domestic Cat (Felis catus) and Asian Leopard Cat (Prionailurus bengalensis). *Journal of Heredity* **112**:165–173 by Oxford University Press, Copyright 2021.

Schwartz et al. 2004, Homyack et al. 2008, Trigo et al. 2008, 2013; Davis et al. 2015; Li et al. 2016b; Figueiro et al. 2017; Li et al. 2019). Previous studies have highlighted the role of interspecific hybrids in the generation of essential genomic tools. In particular, the Bengal cat cross (*Felis catus* x *Prionailurus bengalensis*) was instrumental in the development of the first feline genetic maps (Menotti-Raymond et al. 1999, 2003).

Although the process and accuracy of genome assembly has progressed substantially in the past decade towards the capture of the most repetitive sequences, assemblies of diploid genomes still suffer from several problems: the absence or collapse of long repetitive DNA, failure to resolve sites of high allelic variation between haplotypes, and the pseudohaploid representations of diploid genomes are artifactual representations of the original parental haplotypes. Trio binning was developed to sort and independently assemble divergent parental haplotypes from F1 hybrids using a combination of short read Illumina and long read PacBio sequences, and has been used to generate haploid assemblies for the parent species of several bovid interspecific and subspecific hybrids (Koren et al., 2018; Rice et al., 2020; Low et al., 2020). This approach exploits the high heterozygosity found in interspecies F1 hybrids that was originally used to develop comparative genetic maps that allowed divergent mammalian genomes to be aligned and compared (Lyons et al. 1997; Menotti-Raymond et al. 2003). These same attributes greatly simplify the phasing of parental haplotypes when applied to parent-offspring trios, notably those based on F1 interspecific hybrids. In addition to generating novel genomes from closely related bovid species, the *de novo* assemblies produced from trio-binning have dramatically improved the existing reference genome for domestic cattle. Highly continuous assemblies like these allow gene discovery within and interspecific comparisons between large and complex regions that are

fragmented or lacking in both short-read and long-read diploid-derived genome assemblies (Hsieh et al., 2019; Vollger et al., 2019; Miga et al., 2020). These difficult to assemble regions are increasingly understood as playing important roles in disease biology, genome organization, gene regulation, and speciation. Here, we present two novel haploid *de novo* assemblies for two species of the Felidae, a domestic cat and an Asian leopard cat, generated by applying the trio-binning method to an F1 Bengal hybrid cat.

2.2 Methods

2.2.1 Biological Materials

The parent-offspring trio is composed of a random-bred domestic cat dam, an Asian leopard cat (*Prionailurus bengalensis euptilurus*) sire, and a female F1 Bengal cat offspring ("Amber", aka LXD-97). Fibroblast cell lines were established for the F1 female and the Asian leopard cat sire (Pbe-53). DNA for the domestic cat (Fca-508) dam was extracted from white blood cells. The F1 hybrid was generated at the National Cancer Institute animal colony as part of the generation of an interspecies mapping panel (Menotti-Raymond et al., 1999; 2003; Davis et al. 2015).

2.2.2 Nucleic Acid Library Preparation and Sequencing

Long-read library preparation and sequencing

High molecular weight genomic DNA was extracted using a modified salting out protocol (Miller et al., 1988) followed by length quantification using the Pippin Pulse pulse-field gel system (Sage Science). DNA was quantified via Qubit fluorometric quantification (Thermo Fisher Scientific). PacBio SMRT libraries were size selected (~20-kb) on the Sage Blue Pippin and sequenced across 20 SMRT cells on the Sequel I instrument (V3 chemistry) to yield approximately 90x coverage.

Short-read library preparation and sequencing

Standard dual indexed Illumina fragment libraries (~300-bp average insert size) were prepared for the parent samples using the NEBNext Ultra II FS DNA Library Prep Kit (New England Biolabs Inc.). Libraries were assayed with fluorometric quantification using the Qubit (Thermo Fisher Scientific) and electrophoresis using the TapeStation (Agilent). Samples were sequenced to ~40x genome-wide depth of coverage with 2×150-bp reads using the NovaSeq 6000 Sequencing System (Illumina).

Hi-C library preparation and sequencing

F1 Bengal fibroblasts from Amber were fixed as a monolayer using 1% formaldehyde for 10 minutes, divided into ~$4.2x10^6$ cell aliquots, snap frozen in liquid nitrogen and stored at -80°C as described (Ramani et al., 2016). Cells were lysed, resuspended in 200ul of 0.5x DNase I digestion buffer, and chromatin digested with 1.5 units of DNase I for 4 minutes. Downstream library preparation was performed as described (Ramani et al., 2016) and sequenced across one Illumina HiSeq X Ten lane.

2.2.3 Genome Assembly and Quality Control

All assembly tools used in this manuscript are listed in Table 1.

Haplotype Binning

A summary of software and versions used for each assembly step can be found in Table 3.1. All Illumina data was processed with *FastQC v0.11.8* (Andrews, 2010) followed by adapter trimming using *Trim Galore! v0.6.4*. Parental Illumina sequences were used to phase the raw F1 Bengal PacBio long reads into domestic and Asian leopard cat haplotype bins using the trio binning feature of *Canu v1.8* (*TrioCanu*)(Koren et al., 2017; Koren et al., 2018).

Table 1. Assembly Pipeline and Software Usage.

Software citations are listed in the text.

Assembly and Polishing	**Software**	**Version**
Haplotype Binning	Canu	v1.8
De novo Assembly	NextDenovo	v2.2-beta.0
Contig Polishing	NextPolish	v1.3.0
Benchmarking		
Basic Assembly Stats	QUAST	v5.0.2
Assembly Completeness	BUSCO	v4.0.6
Dotplot Generation	Nucmer	v4.0.0beta2
Dotplot Visualization	Dot	n/a
Scaffolding		
Hi-C Read Haplotyping	https://github.com/esrice/trio_binning	0.2.0
Hi-C Mapping for SALSA	https://github.com/esrice/slurm-hic/	n/a
Hi-C Scaffolding	SALSA2	v2.2
Ref-Based Scaffolding	RagTag	v1.0.1
Hi-C Contact Map Generation	Juicer	v1.5.7
Manual Assembly Inspection	Juicebox Assembly Tools	v1.11.08
Annotation		
Repeat Assessment	RepeatMasker	v4.0.9
Structural Variant Analysis	Assemblytics	v1.2.1
Annotation Liftover	Liftoff	v1.4.2

TrioCanu achieves this by identifying unique k-mers from the parental Illumina reads that are specific to each parental species. The greater the genetic divergence between the parents of the hybrid cross, the larger the number of species-specific k-mers that will be present within each PacBio long read to be classified as belonging to one parent or the other.

De novo Assembly

Haplotyped long reads for each species were assembled using *NextDenovo v2.2-beta.0* (github:Nextomics/Nextdenovo) with the configuration file (.cfg) altered for inputs: *minimap2_options_raw = -x ava-pb*, *minimap2_options_cns = -x ava-ont*. The *seed_cutoff=* option was adjusted to *8478* and *9777* for domestic and Asian leopard cat respectively.

Contig Polishing and QC

NextPolish v1.3.0 (Hu et al., 2019) and *NextDenovo* corrected long reads were used to polish the raw contigs. Notable changes to the *NextPolish* configuration file included: *genome_size=auto*, and *task=best*, which instructs the program to perform 2 iterations of polishing using the corrected long reads. The *sgs* option was removed as polishing with the parental diploid short reads could lead to conversion of consensus sequence to reflect the alternate haplotypes not present in the F1. The *lgs* options within the configuration file was left at default settings except for modification for PacBio long reads by adjusting *minimap2_options= -x map-pb*. Basic assembly stats were generated using *QUAST v5.0.2* (Mikheenko et al., 2018) with the *--fast* run option selected. To assess genome completeness, *BUSCO v4.0.6* (Simão, et al., 2015) was run using the *-m* genome setting with *-l mammalia_odb10* database selected (9,226 single copy genes). Visual assessment of the haploid assemblies was performed through alignment to the felCat9 reference (GCA_000181335.4) (Buckley et al. 2020) using *nucmer* (*mummer3.23* package; Marçais

et al., 2018) with default settings. The resulting delta file was used to generate a dot plot for genome comparison using *Dot: interactive dot plot viewer for genome-genome alignments* (DNAnexus).

Scaffolding

Polished contigs were scaffolded using Hi-C data generated from the F1 hybrid. Prior to scaffolding, F1 Bengal Hi-C reads were binned into parental haplotypes through alignment of the offspring reads to both polished parental assemblies using *bwa mem v0.7.17* (Li and Durbin, 2009) and the *classify_by_alignment*(https://github.com/esrice/trio_binning/ v0.2.0) program as described in Rice et al. (2020). Haplotyped reads were mapped to polished contigs using the pipeline and scripts described in Rice et al. (2020) (https://github.com/esrice/slurm-hic/) using *SALSA v2.2* (Ghurye et al., 2017; Ghurye et al., 2018) with parameters *-e none -m yes.* The haplotyped Hi-C reads were used to scaffold each assembly followed by visual inspection of the SALSA scaffolds using *QUAST*, *nucmer*, and Hi-C contact maps. Following *SALSA*, *RagTag v1.0.1* (Alonge et al., 2019) was used to align scaffolds to their respective position in the felCat9 reference (Buckley et al. 2020) to identify any misassemblies. Selected *RagTag* parameters included *–remove-small, -f 10000* and *-j unplaced.txt*, a text file of scaffolds for *RagTag* to ignore based on their small size and identification as repetitive sequence in the *nucmer* alignments. *RagTag* scaffolds were manually inspected with Hi-C maps generated using *Juicer v1.5.7* (Durand et al., 2016a) with option *-s none* selected for compatibility with DNase Hi-C libraries. Maps were visualized using *Juicebox v1.11.08* (Durand et al., 2016b) and *Juicebox Assembly Tools* with scripts from *3d-dna v.180922* (Dudchenko et al., 2017).

Assembly Quality Control

Assembly quality control was performed by mapping Illumina short-read data from the biological parents, 3 unrelated domestic cats, and 3 unrelated Asian leopard cats to both reference assemblies (Tables B2.1 and B2.2). Dictionaries were created for each reference fasta files using *Picard v2.21.6* CreateSequenceDictionary command. Short-read data was mapped using *bwa mem v0.7.17* (Li et al., 2009) and piped through *Samtools v1.3.1* (Li et al., 2009) view, sort, and index arguments. The sorted BAM files were processed in *GATK* with *v3.8.1* RealignerTargetCreator and IndelRealigner commands to fix indels. The realigned output sequences were then run through *ANGSD v0.925* (Korneliussen et al. 2014) to produce pseudo-haploid sequences from the diploid mappings and were subsequently split by chromosome using *pyfaidx v0.5.8* --split-files argument (Shirley et al., 2015). A multi-alignment file containing all mapped samples was created for each chromosome and parsed into 100kb windows using a custom script (see Data Availability). Pairwise uncorrected *p*-distance values were calculated per-window using a custom *p*-distance calculator script (see Data Availability). Assemblies were then evaluated through visual inspection of the *p*-distance traces across the reference genomes from both species to verify consistent separation of *p*-distances of the two species. Evidence of improper sorting would be indicated by a flip (high-to-low and low-to-high) in the *p*-distance signal of each respective species (all domestic cats and all Asian leopard cats, respectively).

Phased Haplotype Analysis

F1 interspecies hybrids are rare and, sometimes, acquisition of biological specimens from one or both biological parents may be difficult. Therefore, we sought to explore the prospects and limitations of using Illumina sequence data from non-biological parents with

the long-read phasing step in Trio-Canu. To evaluate how replacing one or both biological parents affected the haplotype sorting process, we developed a new script called *Phased Haplotype Analysis* (PHA) (see Data Availability). PHA takes the phased haplotype fasta files (maternal, paternal, and unknown) of a reference cross (biological x biological) and replacement cross (biological x non-biological or non-biological x non-biological) and compares the fasta files to identify correctly and incorrectly sorted reads. Correctly sorted reads are identified as reads that are phased to the same parental haplotype in both the reference and replacement crosses, whereas incorrectly sorted reads are identified as reads that were phased to a different parental haplotype in the replacement cross compared to the reference cross (Figure B2.1). Reads identified as incorrectly sorted in the replacement cross are organized into different subtypes (i.e. maternal-to-paternal, maternal-to-unknown, etc.) (Figure B2.1). PHA provides the number of reads correctly sorted into the same parental haplotype in both crosses, and provides a breakdown of the quantity of incorrectly sorted reads broken down into their respective subtypes. Further characterization of the incorrectly sorted reads was conducted with *RepeatMasker v4.0.7* (Smit et al., 2013-2015).

2.2.4 Genome Annotation

Repeat Sequence Annotation

We used *RepeatMasker v4.0.9* (Smit et al., 2013-2015) with *-excln* and *-species cat* selected to identify and annotate repetitive regions of both genomes while ignoring gap sequence.

Structural Variant Analysis

To estimate indel rates and quantify repeat expansion and contractions we ran *Assemblytics v1.2.1* (web-based) (Nattestad and Schatz, 2016) with a unique sequence length

requirement of 10,000 on nucmer alignments between domestic and leopard cat single haplotype assemblies.

felCat9.0 Gene Annotation Liftover

Because of the high sequence similarity between the domestic and Asian leopard cat genomes, we used *Liftoff v1.4.2* (Shumate and Salzberg, 2020) to perform an annotation liftover between the current felCat9 reference assembly (Buckley et al. 2020) and both *de novo* cat assemblies. Default parameters were used for all arguments except for calling *-copies* with *-sc 0.95* to identify extra copies of genes not previously annotated in felCat9.

2.3 Results

2.3.1 Sequencing and Assembly

All details pertaining to raw sequencing output are included in Table C2.3. Genome assembly and sequencing metrics for the Domestic and Asian leopard cat haploid assemblies are found in Table 2. The number of haplotyped long reads from both parental species was very similar (Fca-508: 49.37%, Pbe-53: 50.62%, Unknown: 0.01%), as would be expected from an F1 individual. The number of assembled contigs for the domestic cat (n=123) and Asian leopard (n=132) cat were also similar. Contig N50 size was 83.88 Mb and 83.70 Mb for the domestic and leopard cat, respectively, a 100% increase relative to the diploid felCat9 long read assembly (contig N50=41.9 Mb) that was based on a highly inbred domestic cat of the Abyssinian breed. The largest contig was generated by the Asian leopard cat assembly, where chromosome A1, the largest cat chromosome, was captured in a single contig spanning the centromere (Figure 6).

Table 2. Assembly Statistics and Benchmarks.

Species	**Domestic cat (2n=38)**	**Asian leopard cat (2n=38)**
Read Count	6,342,174	6,519,732
Base Count (bp)	109,251,556,255	112,023,028,516
Subread N50 (bp)	25,541	25,585
Contig Assembly		
Total Contigs	123	132
Largest Contig (bp)	205,171,639	240,846,738
Ungapped Assembly Length (bp)	2,422,283,418	2,435,689,660
N50 (bp)	83,875,697	83,696,501
BUSCO (mammalia_odb10)		
Single-Copy	8,563	8,589
Duplicated	20	21
Complete	8,583	8,610
Percent Complete	93.03%	93.32%
Fragmented	166	153
Missing	477	463
Percent Present (Comp+Frag)	94.83%	94.98%
Scaffold Assembly Stats		
Total Scaffolds	71	83
Primary Assembly Length (bp)	2,422,299,418	2,435,702,060
Total Gaps	60	56
N50 Scaffold (bp)	147,603,332	148,587,958

Contig alignments to felCat9 chromosomal sequences revealed that a majority of chromosome arms were captured in single contigs, and only 3 chimeric contigs were observed prior to scaffolding (1 in domestic cat, 2 in leopard cat) (Figure B2.2). In the domestic cat assembly, 56% of autosomal chromosome arms were captured in single contigs and 85% in fewer than two contigs. The leopard cat assembly was similarly continuous, with autosomal chromosome arms being captured by 1 (60%) or fewer than 2 contigs (94%). Centromeres were captured within a single contig on 9 domestic cat and 10 leopard cat chromosomes. BUSCO analysis revealed that 95% of the 9226 mammalian BUSCOs were represented in each assembly with most (98%) being complete single-copy. Using a combined Hi-C and reference-based alignment approach we were able to obtain 19 chromosome length scaffolds that represented the conventional felid karyotypic arrangement of 18 autosomes and X chromosome. In the domestic cat assembly, 52 small scaffolds remained unplaced, representing just 0.41% of the un-gapped assembly length. The leopard cat contained 64 unplaced scaffolds composing 0.5% of the un-gapped sequence length. Scaffold alignments to felCat9 revealed the previously observed interchromosomal chimeric contigs were properly resolved (Figure B2.3). Manual inspection using the Hi-C scaffolding data revealed no detectable misassemblies persisting for either assembly Figures A2.4 and A2.5). The total number of gaps introduced into each assembly was 60 (0.016 Mb) for the domestic cat and 56 (0.012 b) for the leopard cat (Tables B2.4 and B2.5).

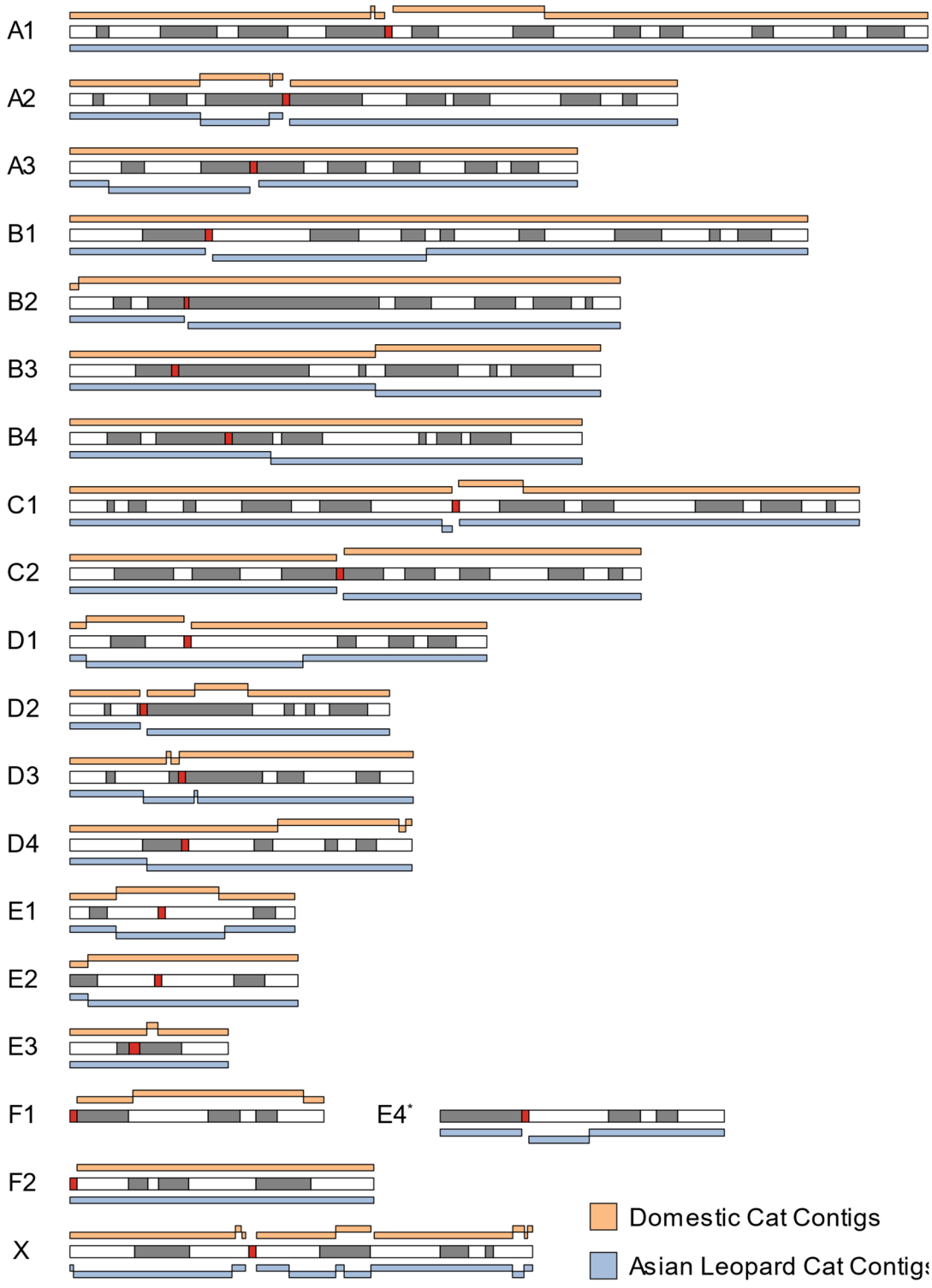


Figure 6. Alignment of domestic cat and Asian leopard cat single haplotype assembly contigs to felCat9.

All ideograms are based on the domestic cat (Cho et al. 1997; Davis et al. 2009) except for the modified F1 to E4 chromosome unique to the species of the genera *Prionailurus*, *Acinonyx* and *Puma* (Graphodatsky et al. 2020). G-banding is represented by dark bars and centromeres by red bars. Domestic cat contigs are depicted as orange bars above the ideogram, and Asian leopard cat contigs are depicted as blue bars below the ideogram.

The X chromosome in particular represented 28% (Fca) and 34% (Pbe) of all gaps, consistent with its enrichment for complex and ampliconic regions.

The scaffold N50 of the final domestic and leopard cat assemblies were 147.60 and 148.59 Mb, respectively, approaching the theoretical maximum based on the domestic cat's conventional chromosome lengths (Buckley et al. 2020). The leopard cat total genome length was 13.39 Mb longer than the domestic cat, which is likely due to variation in repetitive sequence amounts between the two species. This is supported by both RepeatMasker and Assemblytics structural variant analysis where we observed an 8.60 Mb increase in interspersed repeats and 9.19 Mb increase in gained sequence for the leopard cat when comparing the two assemblies (Tables B2.6 and B2.7). Gene liftover from the felCat9 reference assembly to the single haplotype assemblies yielded a total of 19,569 and 19,457 protein coding genes for domestic and leopard cat, respectively (Table C2.8).

Genome alignments revealed 97.3% pairwise sequence identity between the domestic cat and Asian leopard cat chromosomes, estimated from 348,732 alignments of mean length=6.8-kb spanning 99% of the domestic cat assembly. The alignments also revealed no large structural rearrangements (Figure B2.6). Two karyotypic differences were previously suggested to distinguish the two species: a pericentric inversion on Chr D2 and a putative pericentric inversion that distinguishes domestic cat Chr F1 (acrocentric) from Asian leopard cat Chr E4 (metacentric) (Wurster-Hill and Centerwall, 1982). Genome alignments demonstrated that the D2 and F1/E4 homologues are grossly colinear between the two species and the latter difference in centromere location between F1 and E4 is the result of a *de novo* centromere repositioning event.

2.3.2 Assembly Quality Control

To assess the phasing accuracy in our final haploid assemblies we used *p*-distance, the proportion of nucleotide sites where two sequences differ, to determine if any reads/regions of the genome were improperly sorted during the initial haplotype phasing step of *TrioCanu*. For example, any region of the genome where the mapped domestic cat reads were more or equally similar to the reference Asian leopard cat genome than the other leopard cat sequences would be considered evidence of improperly phased sequences, or alternatively, past episodes of introgression (e.g., Rice et al. 2020). The full genome *p*-distance plots for both the Domestic cat (Figure B2.7) and the Asian leopard cat (Figure B2.8) show consistent separation of the domestic cat and Asian leopard cat *p*-distance traces across all chromosomes (Figure 7A and B). This indicates that *TrioCanu*'s haplotype phasing step properly binned the long-read data into their respective parental haplotypes.

We also evaluated *TrioCanu*'s ability to accurately phase the F1 hybrid long-read data by replacing the Illumina data from either the biological mother (Fca-508, domestic cat), the biological father (Pbe-53, Asian leopard cat), or both biological parents with data from other individuals. We used three unrelated domestic cats and three unrelated Asian leopard cat samples (Table C2.1), one being the same subspecies (*P. bengalensis euptilurus*) as the Asian leopard cat sire, and two being from a different subspecies (*P. bengalensis bengalensis*). Our results produced nearly identical results from the haplotype sorting process with biological parents, with only a relatively small number of reads being phased to a different haplotype (Figure 2c). Further analysis revealed that the vast majority of incorrectly sorted reads (i.e., a read was sorted to a different haplotype in the replacement cross compared to the reference cross) were short in length (Figure B2.9), with 79-80% of

the reads being shorter than 10-kb in length (Table C2.9). We also analyzed the subtype distribution of the incorrectly sorted reads (i.e. incorrectly sorted from mother-to-father, father-to-mother, etc.) and found that the majority were switching between the maternal and paternal haplotypes (Table C2.10), comprising just 3.5% (<3X mean coverage) of the total F1 hybrid sequence data. However, when we performed replacement crosses with Asian leopard cats from a divergent subspecies (*P. b. bengalensis*) or closely related species (*P. javanensis* or *P. viverrinus*) the phasing of the parental reads were increasingly skewed towards one parent (Figure B2.10).

Finally, we reassembled the PacBio reads from LXD-97 after read phasing was performed with Illumina data from two different individuals (LilBub and Pbe-14) rather than the actual biological parents, Fca-508 and Pbe-53. The resulting domestic cat assembly aligned across 99.99% of the original Fca-508 assembly (Figure B2.11a), and differed in assembly length by only 0.34%, with an average 99.98% sequence identity and a SNP rate of 0.001%. The Asian leopard cat assembly produced comparable results, aligning across 100% of the original Pbe-53 assembly (0.24% length difference) with 99.98% sequence identity and a SNP rate of 0.001% (Figure B2.11b).

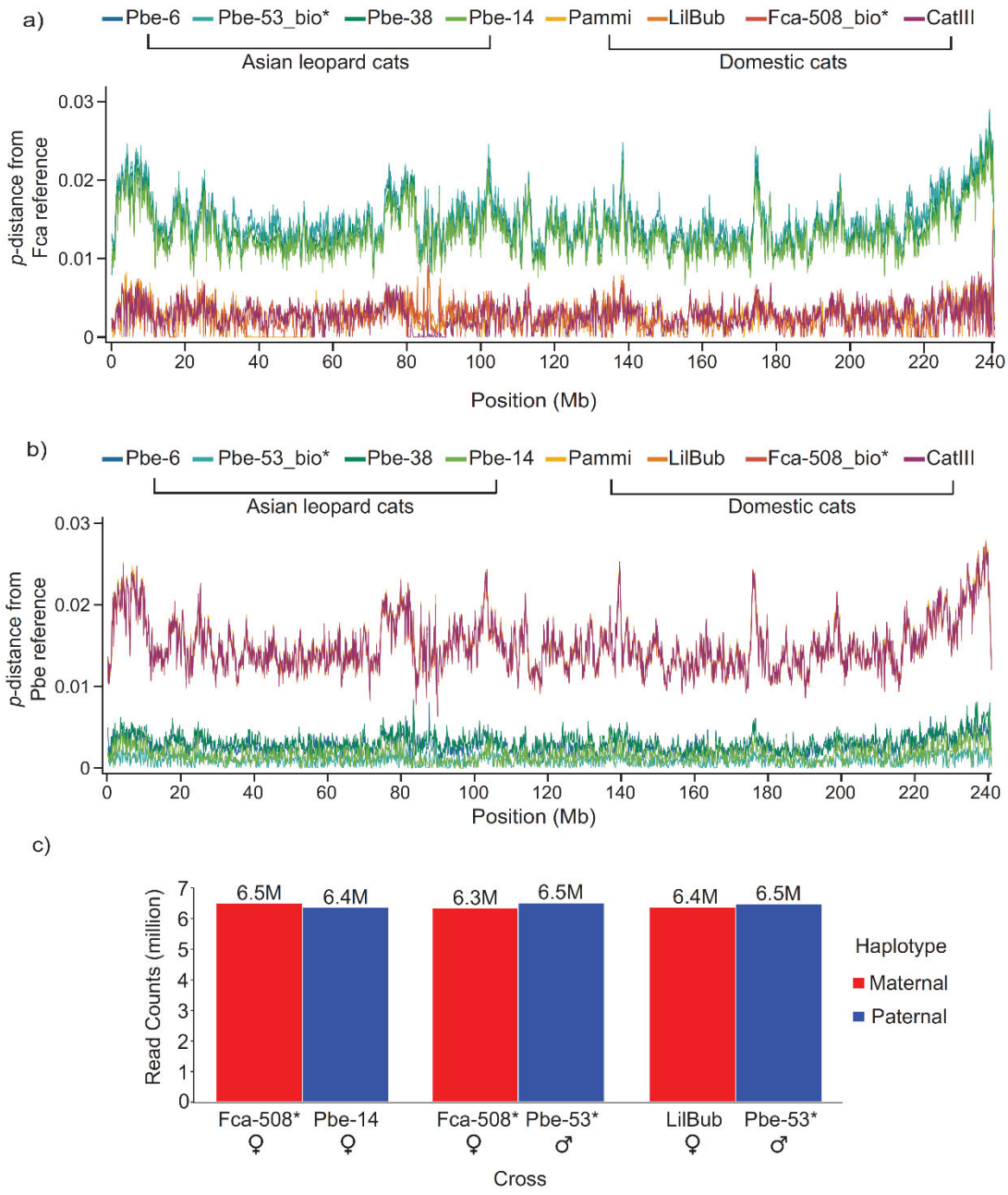


Figure 7. Read count distribution of single-replacement crosses and Chromosome A1 *p*-distance plots for both the domestic and Asian leopard cat reference sequences.

(a) p-distance traces for the biological parents and test sample short read data from both species mapped to the single haplotype domestic cat genome assembly. The Asian leopard cat samples show a clear separation from the traces of the domestic cat samples which lie close to 0. (b) *p*-distance traces for the biological parents and test sample short read data from both species mapped to the single haplotype Asian leopard cat genome assembly. In contrast to the *p*-distance traces for the domestic cat assembly, the Asian leopard cat traces lie close to 0 while the divergent domestic cat sample traces lie well above indicating uniformly elevated divergence from the Asian leopard cat assembly. The consistent separation of reads from the two species in both (a) and (b) demonstrates that *TrioCanu* has properly phased the F1-hybrid long-read data into their appropriate parental haplotypes. (c) Read count distributions of single-replacement crosses post-haplotype phasing. (*) = Biological parents. See Supplemental Tables 1 and 2 for individual sample IDs.

2.4 Discussion

We have produced two highly continuous genome assemblies for the domestic cat (*Felis catus*) and the Asian leopard cat (*Prionailurus bengalensis*) by applying the trio-binning approach to long sequence reads from a Bengal F1 hybrid. Sequence continuity for these two assemblies is twice that of the most recent diploid-based long read domestic cat reference (Buckley et al. 2020) and is equivalent to that of the most recent haploid human genome assemblies (Miga et al. 2020). Sequence improvements and gains relative to the diploid felCat9 assembly include complex repetitive regions previously un-spanned due to insufficient read lengths and/or high haplotype divergence, and resolution of multicopy gene families with high allelic diversity (i.e., Major Histocompatibility Locus, olfactory receptors). Furthermore, we have provided a genome assembly from a random-bred domestic cat, which is more representative of the domestic cat pet population.

In addition to improvements in the domestic cat reference genome, the simultaneous generation of a highly continuous Asian leopard cat genome will be a valuable resource for studying the population genetic diversity, subspecies delimitation and conservation with this species and other members of *Prionailurus*. This genome will also be valuable for health studies in closely related species, such as transition cell carcinoma in fishing cats and polycystic kidney disease in Pallas' cats. High resolution comparisons between a hybridizing pair of felid species will also be valuable for quantifying species-specific divergence across copy number variable regions, previously described in felids as being associated with hybrid sterility and speciation (Davis et al., 2015). The success of the trio-binning approach has stimulated the generation of other highly continuous genomes derived from additional felid F1 hybrids, like the Safari cat (domestic cat x Geoffroy's cat) and liger

(lion x tiger) (Bredemeyer et al. in prep.). Comparative genomic analyses from these high-quality assemblies will produce unprecedented insights into mechanisms underlying morphological divergence, adaptation and speciation within this enigmatic mammalian family.

F1 interspecies hybrids are rare biological resources, and in many cases, it may be logistically impossible to obtain the actual parents of the cross. This motivated us to explore the feasibility of applying short-read data from other conspecifics to phase the F1 long read sequences. We demonstrated that *TrioCanu*'s phasing is robust to the inclusion of Illumina short-read data from non-biological parents when one or both biological parents are missing, producing assemblies of virtually identical length and sequence identity with those produced from reads phased by the biological parents. However, under such circumstances we recommend phasing with reads from an individual of the same subspecies or derived from a genetically similar population, as phasing errors increase with divergence from the parental species.

Felid genomes are known to be highly conserved across the family, with G-banding and FISH analyses showing gross co-linearity across the majority of feline autosomes and the X chromosome (Wurster-Hill & Centerwall et al. 1982; Modi & O'Brien 1988; Davis et al. 2009). Our study provides the first demonstration that the genomes of *Felis* and *Prionailurus*, although karyotypically distinct, are grossly colinear and that cytogenetic differences do not correspond to chromosomal rearrangements. This confirms recent genomic comparisons between the domestic cat and lion that also demonstrated gross collinearity across the deepest divergence of the cat family (Armstrong et al. 2020), and

suggests an even more extreme level of karyotypic conservation within the Felidae than previously appreciated.

CHAPTER III RAPID MACROSATELLITE EVOLUTION PROMOTES X-LINKED HYBRID MALE STERILITY IN A FELINE INTERSPECIES CROSS

3.1 Introduction

Hybrids between mammalian species are often infertile or inviable as a result of genetic incompatibilities between parental haplotypes, as described by the "two rules of speciation" (Coyne & Orr, 1989; Coyne, 2018). The first, Haldane's Rule, is the long-standing observation of preferential sterility or inviability of the heterogametic sex resulting from an interspecific mating (Haldane, 1922). The second rule is the large X-effect (Coyne, 1992), which is the observation that the X chromosome is enriched for hybrid sterility factors and plays an increased role in post-zygotic isolation relative to autosomes (Coyne, 1992; Masly & Presgraves, 2007). Support for the large X-effect in mammals has come from genetic mapping studies of sterility phenotypes to the X chromosome using genome-wide association, QTL and eQTL studies (Good et al., 2008; Bhattacharyya et al., 2014; Turner et al., 2014a; Turner et al., 2014b; Schwahn et al., 2018; Lustyk et al., 2019).

Gene expression comparisons between fertile and sterile hybrid mouse testes revealed that X-linked genes in sterile testes were highly upregulated across the X chromosome (Good et al., 2010). Subsequent hybrid sterility studies employing enriched germ cell populations revealed X upregulation occurred in all stages of spermatogenesis but was most pronounced in cell stages that undergo meiotic sex chromosome inactivation (MSCI) and exhibit X-chromosome downregulation during normal spermatogenesis (Namekawa et al., 2006; Campbell et al., 2013; Larson et al., 2016). MSCI normally results in silencing and partitioning of both sex chromosomes into a heterochromatic XY body

during pachynema (Solari, 1974; Handel, 2004; Turner, 2007). MSCI is likely evolved in response to the extensive X-Y asynapsis, which is a consequence of the gradual increase in structural and sequence divergence of the differentiating X and Y gametologs during early stages of sex chromosome evolution (Lahn & Page 1999; Graves, 2006; Liu, 2019). MSCI is a specific version of a more general mechanism, termed meiotic suppression of unsynapsed chromatin (MSUC), which silences any asynaptic regions between homologous molecules (Schimenti, 2005; Turner et al., 2005). While both mechanisms occur during the pachytene stage of meiosis I, MSUC is often caused by high sequence divergence or structural variation between otherwise homologous chromosomes (Wang & Höög, 2006), while MSCI results from the heteromorphic nature of mammalian sex chromosomes specifically (McKee & Handel, 1993). The failure of the X chromosome to undergo proper conformational changes due to intrachromosomal structural variation has been hypothesized as an underlying mechanism explaining failure of MSCI, which manifests as X-chromosome upregulation during spermatogenesis of sterile hybrids (Lifschytz & Lindsley, 1972; Jablonka & Lamb, 1991).

Although previous studies have linked the failure of MSCI or post-meiotic sex chromosome repression (PSCR) to pachytene arrest and male sterility (Burgoyne et al., 2009; Good et al., 2010; Royo et al., 2010; Larson et al., 2016, Schwahn et al., 2018), and identified genes required for proper regulation of MSCI (Vernet et al., 2016), a genetic mechanism is currently lacking to explain how X-chromosome divergence triggers the failure of MSCI and X-linked gene upregulation in the context of hybrid sterility. Our previous studies demonstrated that two biomarkers of sterility in sterile mice testes, X chromosome-wide upregulation and meiotic arrest during pachynema, were conserved in cat

interspecific hybrids (Davis et al., 2015). This suggested sterile hybrids from divergent mammalian orders phenocopy each other and could conceivably result from a similar mechanism.

Here we explore hybrid male sterility in an interspecific hybrid cat breed, the Chausie. Chausies comprise an admixed cat breed population initially derived from a small number of foundational crosses between male Jungle cats (*F. chaus*) and female domestic cats (*F. silvestris catus*). We identify a novel mammalian X-linked candidate hybrid sterility locus, *DXZ4*, using a combination of GWAS, ancestry-based fine mapping, and genome-wide methylation analysis approaches. We used sorted germ cells from a fertile domestic cat to generate expression, methylation, and chromatin profiles of *DXZ4* and the X chromosome during various stages of spermatogenesis. Our results indicate that *DXZ4* plays a heretofore undescribed role in normal male meiosis, and that interspecific divergence across this locus likely contributes to disruption of MSCI, manifesting as hybrid sterility.

3.2 Results and Discussion

3.2.1 Biomarkers of Chausie Hybrid Male Sterility

We analyzed histology and RNA-Seq data from fertile and sterile testes from Chausies of fourth and fifth backcross generations containing similar pedigree-based estimated percentages (13-14%) of Jungle cat ancestry. Histological analysis of seminiferous tubule cross-sections from sterile Chausies revealed vacuolization, depletion of post-pachynema germ cells, and meiotic arrest at the pachytene spermatocyte stage, similar to observations from testes of sterile males in other hybrid cat breeds (Figure 8A) (Davis et al., 2015) and many other mammalian species (Moore et al., 1999; Thomsen et al., 2011; Bhattacharyya et al., 2013; Ishishita et al., 2015).

Transcriptional profiling of seminiferous tubule RNA isolates from sterile Chausies revealed upregulation in 22% of annotated protein coding genes along the X chromosome, a significant enrichment relative to autosomes (Fisher's exact test, *p*-value=1.5e-21). Expression of X-linked genes increased by a fairly uniform 2.45 log-fold average across the length of the chromosome (Figure 8B). There was no apparent clustering of differentially expressed genes along the length of the X chromosome (Figure 8C), consistent with a pattern of chromosome-wide misregulation as opposed to regional escape of meiotic gene silencing, concordant with previous observations from felid interspecific and rodent subspecific hybrids (Good et al., 2010; Davis et al., 2015; Larson et al., 2016). X upregulated genes were not significantly enriched for any biological processes. Twenty-two genes that were expressed in sterile Chausies showed zero expression in normal, fertile testes from domestic cats or Chausies (Figure 8C). These genes exhibited an average log-fold increase of 3.28, significantly larger than the fold change of genes normally expressed in fertile testes (student 2-tailed t-test, *p*-value=0.002). These results suggest the observed chromosome-wide misregulation in sterile Chausies was likely due to defective MSCI, and not an artifact of tissue composition bias (Good et al., 2010). Interestingly, we also observed significant numbers of genes downregulated across almost all autosomes in these testis samples (chi-squared test, *p*-value=2.2e-16) (Figure 8B, Figure B3.1), a feature not previously observed in other felid interspecific hybrids (Davis et al., 2015), despite sharing a similar meiotic arrest phenotype. Downregulated genes genome-wide were enriched for biological processes involved in meiotic division and spermatogenesis (Table C3.1), suggesting the

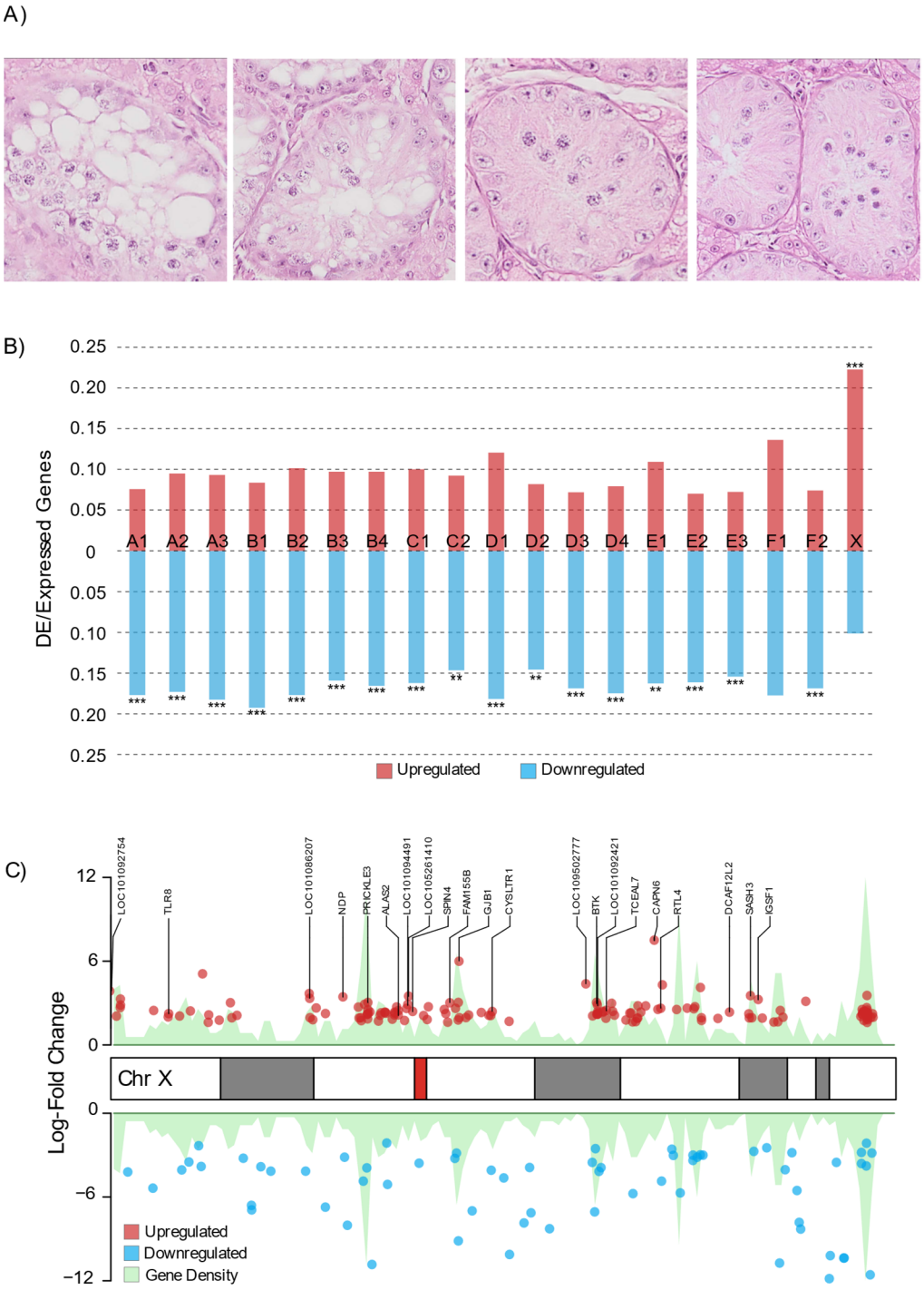


Figure 8. Chausie F1 males exhibit two biomarkers commonly associated with hybrid male sterility in mammals.

A) Histological cross sections of seminiferous tubules from sterile testes indicate arrest of spermatogenesis during pachynema. B) Differential expression analysis of RNA-Seq data from sterile and fertile testes reveal significant enrichment for upregulated genes on the X chromosome in sterile testes. P-values: *≤ 0.05, **≤0.01, ***≤0.001. C) Distribution of genes differentially expressed on the X chromosome in sterile testes. Labeled genes were those upregulated in sterile testes and lacking expression in fertile testes.

apparent downregulation of many of these genes likely stems from an absence of post-pachytene germ cells.

3.2.2 GWAS and Fine Mapping of Hybrid Sterility Locus DXZ4

To identify genetic variation associated with the hybrid male sterility phenotype, we performed a genome-wide association study (GWAS) by genotyping a cohort of 39 backcross hybrid males (previously phenotyped for sterility as described in Methods) on the Illumina 63K feline array. A single cluster of significant SNPs was identified on the X chromosome at ~94 Mb in the felCat8.0 reference assembly (Figure 9A). We performed fine-mapping of the candidate SNP region by genotyping 27 ancestry-informative short interspersed element (SINE) insertions identified from whole genome sequence alignments that distinguished the domestic cat and Jungle cat X chromosomes. This approach identified a 500 kb critical interval between 93.74 and 94.24 Mb in 94% of backcross hybrid males that possessed Jungle cat ancestry and were sterile. In contrast, 100% of male Chausies that inherited a domestic cat haplotype within this region were fertile (Figure 9B).

Three annotated functional loci reside within the critical interval region, two protein coding genes (*PLS3* and *AGTR2*) and a large macrosatellite repeat array (*DXZ4*) that has a well-documented role in primate and rodent female XCI (Deng et al., 2015; Darrow et al., 2016). The two protein coding genes within this interval were excluded from consideration based on the following functional genomic data. The first gene, *PLS3*, has a broad tissue expression profile with relatively low expression in testis reported in the NCBI genome browser and has no described role in spermatogenesis (Figure B3.2). The second gene, *AGTR2*, is not expressed in the feline testis. Therefore, we explored *DXZ4* as the candidate locus.

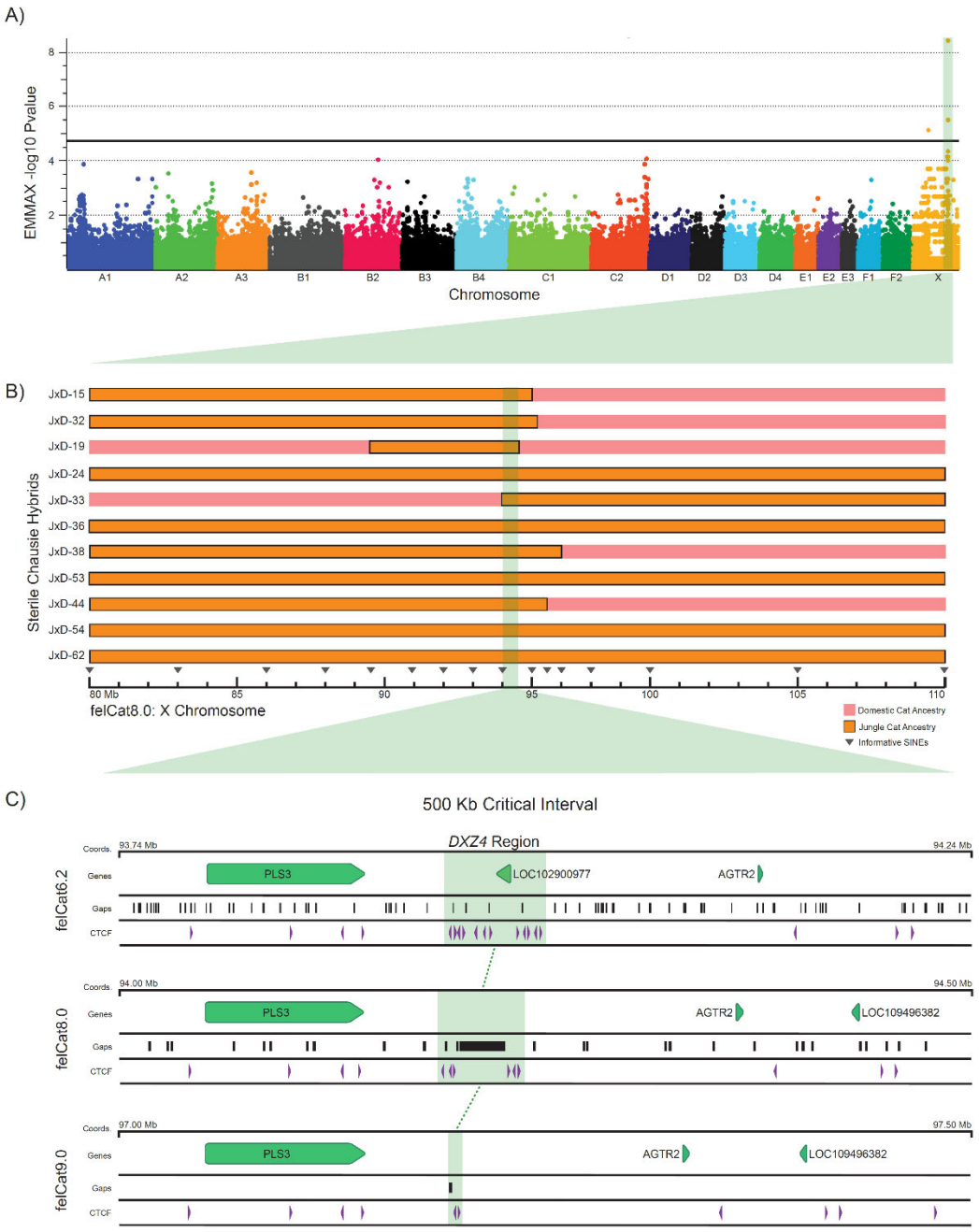


Figure 9. Identification of hybrid sterility locus *DXZ4* in a cohort of male Chausie backcross hybrids.

A) Hybrid sterility GWAS in a cohort of sterile and fertile Chausie hybrids identifies a significant SNP peak across the 94 Mb region of the felCat8 X chromosome. exceeding the Wellcome Trust recommendations for genome-wide significance ($P_{uncorrected}<5x10^{-5}$; $-\log_{10}P$ = 4.30) B) Fine mapping studies in Chausie backcross hybrids (n=80). A 500 kb critical interval was identified by associating regions of shared Jungle cat ancestry with the sterile phenotype (only the most informative hybrids are shown). Jungle cat ancestry for this region was observed in 14/16 sterile Chausie hybrids, while domestic ancestry was observed in all fertile hybrids (n=23). C) Macrosatellite *DXZ4* identified as candidate hybrid sterility locus with CTCF binding sites flanking assembly gaps and conserved position downstream of *PLS3*.

DXZ4 is a macrosatellite, a subcategory of variable number tandem repeat (VNTR) sequences distinguished by its large, multi-kilobase repeat unit (Giacalone et al., 1992; Dumbovic et al., 2017). *DXZ4* plays a well described role in structural aspects of female XCI which occurs in female somatic cells and results in the inactivation of a single X chromosome to balance gene dosage with hemizygous males (Lyon, 1961; Brockdorff & Turner, 2015; Galupa & Heard, 2018; Bansal et al., 2020). Given this previously described role in XCI, we were intrigued by the strong genotypic association based on *DXZ4* ancestry in males: 94% of sterile hybrids possessed a Jungle cat *DXZ4* allele within a mostly domestic cat genetic background, whereas all fertile hybrids possessed a domestic cat *DXZ4* allele. Prior to this observation, there was no evidence that *DXZ4* had a functional role in male meiotic silencing of sex chromosomes despite numerous parallels between X chromosome states resulting from male (i.e., MSCI) and female (i.e., XCI) silencing mechanisms. The final heterochromatic state is similar between the two silencing processes despite differences in certain histone modifications and differences in DNA methylation (McCarrey et al., 1992; Armstrong et al., 1997; Hoyer-Fender, 2003; Moretti et al., 2016), as is the segregation of the inactive X from autosomes through association with the nucleolus (Kierszenbaum & Tres, 1974; Knibiehler et al., 1981). During XCI, the shift in localization is governed by *DXZ4* and lncRNA from the associated locus *FIRRE*, which anchor the inactive X (Xi) to the nucleolar periphery (Deng et al., 2015; Yang et al., 2015; Fang et al., 2019). In XCI, nucleolar association is thought to govern the Xi epigenetic state (Zhang et al., 2007).

While localization of the X chromosome to the nucleolar periphery occurs during MSCI, its significance, as well as the role *DXZ4* might play in maintaining this state, is unknown. In addition to epigenetic and spatial similarities, distinct repertoires of miRNAs

escape X chromosome silencing in each instance, suggesting a common role for post-transcriptional regulation and the potential for shared means of gene escape between silencing mechanisms (Yan & McCarrey, 2009; Sosa et al., 2015). To evaluate *DXZ4* expression and methylation patterns during spermatogenesis, we profiled three different functional properties of *DXZ4* in whole testes and sorted germ cells from sexually mature domestic cats: 1) transcription using RNA-Seq, 2) methylation using reduced representation bisulfite sequencing (RRBS), and 3) long-range chromatin interactions of the silenced X chromosome using Hi-C.

3.2.3 DXZ4 *Assembly and Structural Assessment*

Because of its complex structure and polymorphic nature, *DXZ4* is incomplete or altogether missing in even the highest quality genome assemblies (Figure 9C) and has been poorly studied in most placental mammals. We first generated contiguous assemblies across the *DXZ4* macrosatellite in both parent species of the Chausie cat hybrids to enable a comparative analysis of locus structure and copy number, and facilitate examination of X-chromosome expression and methylation in the male germ line. *DXZ4* was annotated adjacent to an assembly gap in all previous iterations of the domestic cat reference genome assembly (Montague et al., 2014, Li et al., 2016b, Buckley et al., 2020) (Figure 9C). Comparisons between the different genome assemblies revealed substantial variation in the proportion of *DXZ4* sequence successfully incorporated into the X chromosome, likely due to the different sequence chemistries employed and their variable average read lengths (Figure B3.3). Fortuitously, the *DXZ4* locus was fully assembled in our recently published single haplotype genome assemblies for the domestic cat and Asian leopard cat (Bredemeyer et al., 2021) (Figure B3.4). We additionally sequenced and assembled the genome of a male

Jungle cat (where the X chromosome is effectively haploid) from long PacBio sequence reads (Text A3.1., Table C3.2). This Jungle cat assembly (FelChav1.0) was highly contiguous and contained within 106 contigs, totaling 2.43 Gb, with a contig N50=91 Mb and scaffold N50=148.6 Mb (Table 3).

In all three felid assemblies, *DXZ4* was embedded within a single contig and exhibited sequence gain relative to the gapped reference felCat9, suggesting successful assembly of the locus (Figure B3.4). *DXZ4* maintained its position downstream of *PLS3*, consistent with other mammalian genomes, despite partial assembly of the macrosatellite (Horakova et al., 2012a) (Figure B3.5). In each of our felid genome assemblies, *DXZ4* is composed of a compound satellite repeat with two divergent tandem repeats, divided by a conserved spacer sequence (Figure 10). The repeat array proximal to *PLS3*, hereby referred to as Repeat A (RA), contains a single CTCF site in the reverse direction, a conserved characteristic of repeat units that comprise the monomeric human *DXZ4* repeat array (Miga et al., 2020). The more distal Repeat B (RB), on the other hand, possesses two CTCF sites facing away from one another. The intervening spacer sequence was 32.15, 34.98 and 33.97 kb long in domestic cat, Jungle cat and Asian leopard cat assemblies, respectively. Alignments between the regions revealed five conserved CTCF binding motifs of varying directionality, which accounted for approximately half the sequence divergence observed between species RA and RB monomers. (Figure B3.6, Table C3.3).

Table 3. FelCha1.0 assembly statistics.

Species	**Jungle Cat (2n=38)**
Read Count	7,594,421
Base Count (bp)	122,681,343,250
Subread N50 (bp)	25,928
Contig Assembly	
Total Contigs	106
Largest Contig (bp)	205,710,267
Ungapped Assembly Length (bp)	2,428,281,414
N50 (bp)	91,188,488
BUSCO (mammalia_odb10)	
Single-Copy	8,559
Duplicated	25
Complete	8,584
Percent Complete	93.04%
Fragmented	180
Missing	462
Percent Present (Comp+Frag)	94.99%
Scaffold Assembly Stats	
Total Scaffolds	52
Primary Assembly Length (bp)	2,428,287,114
Total Gaps	61
N50 Scaffold (bp)	148,552,997

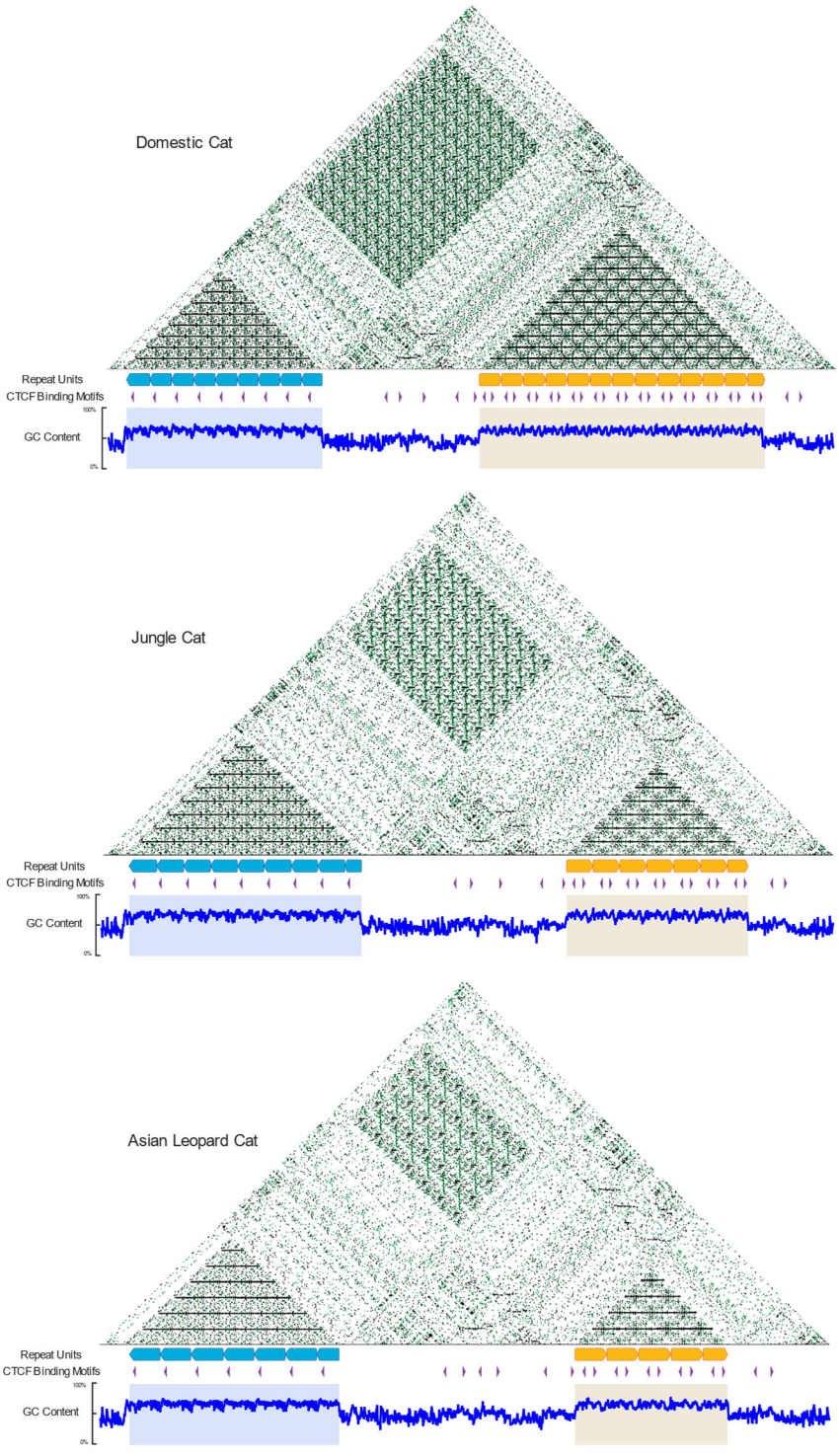


Figure 10. Dichotomous structure of *DXZ4* revealed in 3 cat species.

Self-dot plots and CTCF motif annotations reveal that macrosatellite *DXZ4* is composed of two distinct tandem repeats divided by a conserved spacer sequence in cats. Enrichment for CpG islands as suggested by high GC content across repeat units is consistent with previous observations of the locus in human and mouse.

In addition to differences in motif profiles, we observed instances of copy number variation between the tandem repeats in the three felid species. The average interspecific length of RA and RB was 4,554 bp and 4,607 bp, respectively, with a standard deviation (SD) of only 46 bp between all repeat units (Table C3.4). A maximum likelihood phylogeny of the repeat units resolved RA and RB into reciprocally monophyletic groups, with a large, between-group mean *P*-distance=0.424. Mean within-group genetic distances were 15-30-fold smaller: 0.028 between RA units and 0.013 between RB units. Within RA and RB arrays, the repeat sequences grouped by species with the exception of the most proximal repeat, RA-1, which formed a divergent clade that also follows the species tree (Li et al., 2019) (Figure 11).

Despite capturing *DXZ4* within single contigs and overall structural conservation across all three cat assemblies, repeat array lengths (RA: 31-41 kb, RB: 22-59 kb, Full Array: 105-150 kb) far exceeded the PacBio mean read length (16-17 kb). We subsequently estimated the copy number with an *in-silico* approach that utilized short-read mapping and collapsed repeat arrays (Lucotte et al., 2018) (Figure B3.7). This approach was used to validate copy number for the assembly and investigate copy number variation across multiple individuals. We confirmed that copy number of RA and RB differed between all three species (Table C3.5). Copy number of *DXZ4* repeat arrays within domestic cats varied dramatically, especially across RA (SD = 6.81), which had twice the standard deviation of RB (SD = 2.76). Standard deviation for domestic cat total copy number was also very high (SD = 8.45), but was expected based on the hypervariability of the locus described in human (Tremblay et al., 2011; Schaap et al., 2013). Subdivision of domestic cats based on breeding history reveals outbred domestic shorthairs display increased variation in copy number

relative to defined breeds. Despite variability across individuals, the relative relationship between copy number of RA and RB remained constant within each species, with domestic cat RA copy number being less than RB on average while the Jungle cat and Asian leopard cat average copy number of RA exceeded that of RB.

3.2.4 DXZ4 *Expression in Male Germ Cells*

We next asked how interspecific variation at *DXZ4*, a locus associated with a seemingly unrelated process in female somatic cells (XCI), could contribute to hybrid sterility in males. XCI requires the presence of the lncRNA *XIST*, as well as a host of long-range chromatin interacting loci (including *DXZ4*) and epigenetic modifiers (Jégu et al., 2017; Bansal et al., 2020). These loci work together to alter the structural, epigenetic, and transcriptional landscape of the inactivated X chromosome (Xi), ultimately resulting in the condensation and nucleolar association of the heterochromatic Barr body (Barr & Bertram, 1949; Bourgeouis et al., 1985; Dyer et al., 1989; Chadwick & Willard, 2003). We hypothesized that feline *DXZ4* has a structural and/or functional role during MSCI, specifically in the formation or maintenance of the XY body formed from meiotic silencing of unsynapsed chromatin between the X and Y chromosome. This XY body is analogous to the Barr Body that forms during XCI. If true, we secondarily posited that interspecific variation at *DXZ4* might perturb XY body formation as a result of altered three-dimensional chromatin interactions or altered gene expression that lead to the observed upregulation of X linked genes and meiotic arrest in sterile Chausies.

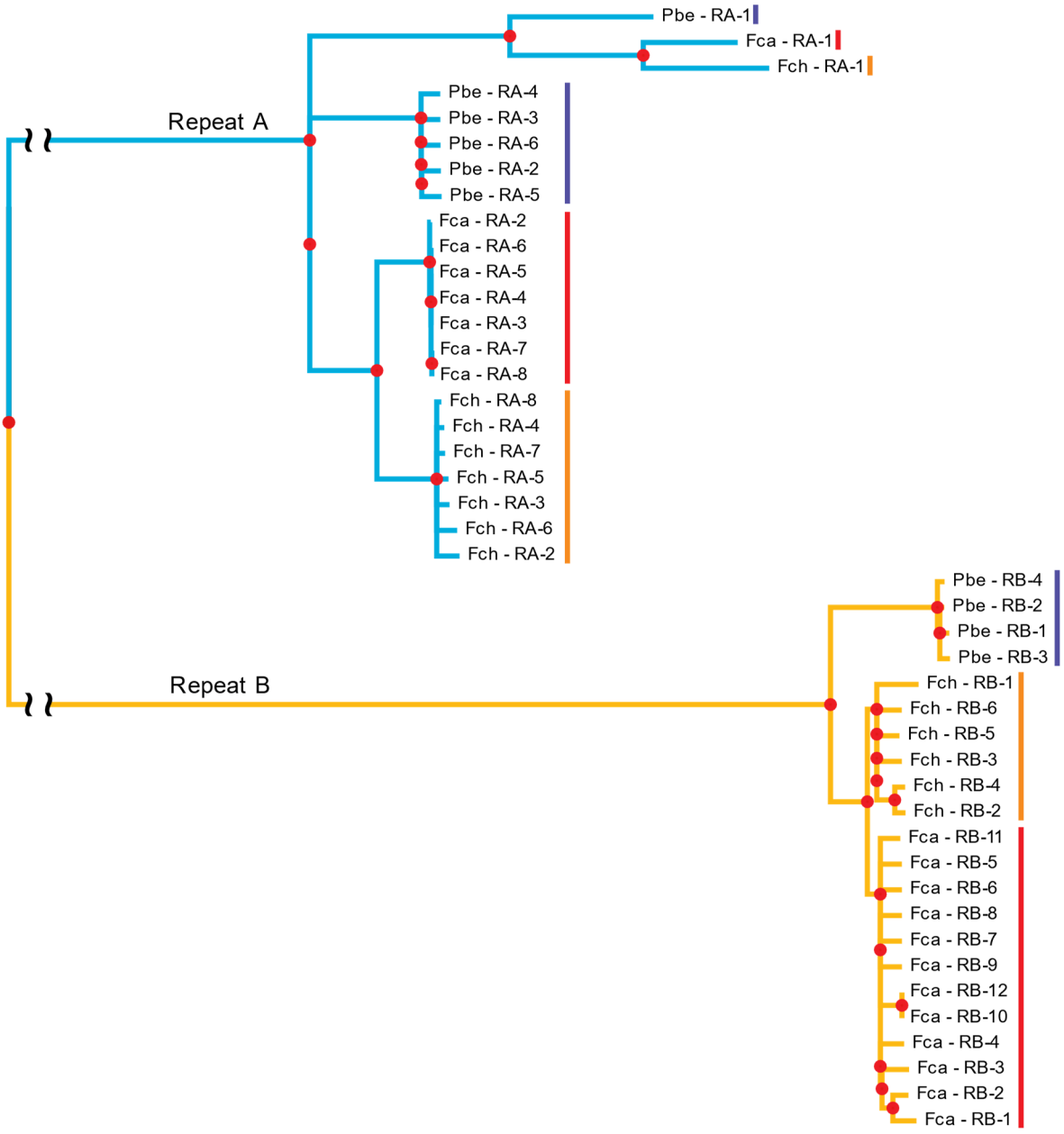


Figure 11. *DXZ4* repeat unit phylogenetic analysis.

Neighbor-Joining tree generated from an alignment of felid *DXZ4* repeat units. Repeats A and B are represented from domestic (Fca), Jungle (Fch), and Asian leopard (Pbe) cats with a *DXZ4* repeat unit from the human telomere to telomere assembly (Miga et al., 2020) serving as an outgroup to root the tree (Not shown). Red dots represent a bootstrap support value ≥90.

Using our fully assembled *DXZ4* loci, we first investigated transcriptional activity during spermatogenesis in fertile domestic cats. In human XCI, each *DXZ4* repeat unit is capable of transcribing smallRNAs, while *DANT1* and *DANT2* genes both transcribe multiple isoforms described as either short- or array-traversing transcripts (ATTs), with the latter spanning the entire macrosatellite from flanking promotors (Figueroa et al., 2015). Although the precise function of the transcribed lncRNAs and smallRNAs are still poorly understood, they are thought to contribute to regulation of the inactive chromatin state (Pohlers et al., 2014).

Analysis of rRNA-depleted and smallRNA-enriched RNA-Seq data from sorted germ cells and seminiferous tubules, revealed transcription of *DXZ4* in both pachytene spermatocytes and round spermatids, indicating that the locus escapes X silencing after MSCI and during the formation of post-meiotic sex chromatin (PMSC) (Figure 12). The largest peaks of transcriptional activity occurred across the spacer region adjacent to the RA array and not within individual repeat units of either array **(**Figure 12A). *De novo* transcript assembly identified multiple transcripts spanning the *DXZ4* locus that vary across cell types with lengths between 364 and 1,321 bp. A single 426 bp repeat-A-spanning transcript (*RAST*) similar in orientation and positioning to the human *DXZ4* array-traversing *DANT1-ATT* isoform was annotated in domestic cat whole testes, pachytene spermatocytes and round spermatids. Annotation of the RB region also revealed a 601 bp repeat-B-spanning transcript (*RBST*) in all cell types. The orientation and positioning of *RBST* appeared orthologous to the *DANT2-ATT* isoform in humans (Figure B3.8). Orthology between array spanning transcripts in cat and human was further supported by ungapped pairwise alignment identities of 46% and 47% between *RAST*/*DANT1-ATT* and *RBST*/*DANT2-ATT* mRNAs,

respectively. While this sequence similarity would be considered low for protein-coding sequences, these results are not unexpected for lncRNAs. LncRNAs evolve rapidly between species and exhibit signatures of conservation outside of overall sequence identity and that vary according to the functional role of the RNA within the nucleus (Kutter et al., 2012, Dhanoa et al., 2018, Ramírez-Colmenero et al., 2020). We observed many islands of consecutive conserved bases across our alignments, indicative of evolutionary constraint in DNA interacting motifs/domains or lncRNA structural conformation (Ramírez-Colmenero et al., 2020) (Figure B3.9).

Additional cell type-specific transcripts were observed within and directly downstream of RB. In sorted germ cells and whole testes, we observed coverage peaks spanning ~250 bp in each of the first four (most proximal) RB units (Figure 12A). Discontiguous mega blast analysis of the sequences underlying the peaks revealed no significant matches to the nucleotide collection database, suggesting this transcript type is unique to domestic cat. Annotation of the smallRNAs revealed two smallRNA clusters that vary only slightly in lengths and positioning between the two cell types across the spacer sequence that separates RA and RB (Figure 12B). SmallRNA annotations were absent from both RA and RB repeat arrays, despite visible coverage peaks of varying length across individual RA units. However, these patterns were lacking in RB units, suggesting differences in smallRNA transcription between the two arrays.

Expression across *DXZ4* was also detected in Jungle cat seminiferous tubule RNA-seq data. Similar to the domestic cat, RA and RB spanning transcripts were detected, with one additional RA and two additional RB-spanning species-specific transcript variants annotated (Figure B3.10).

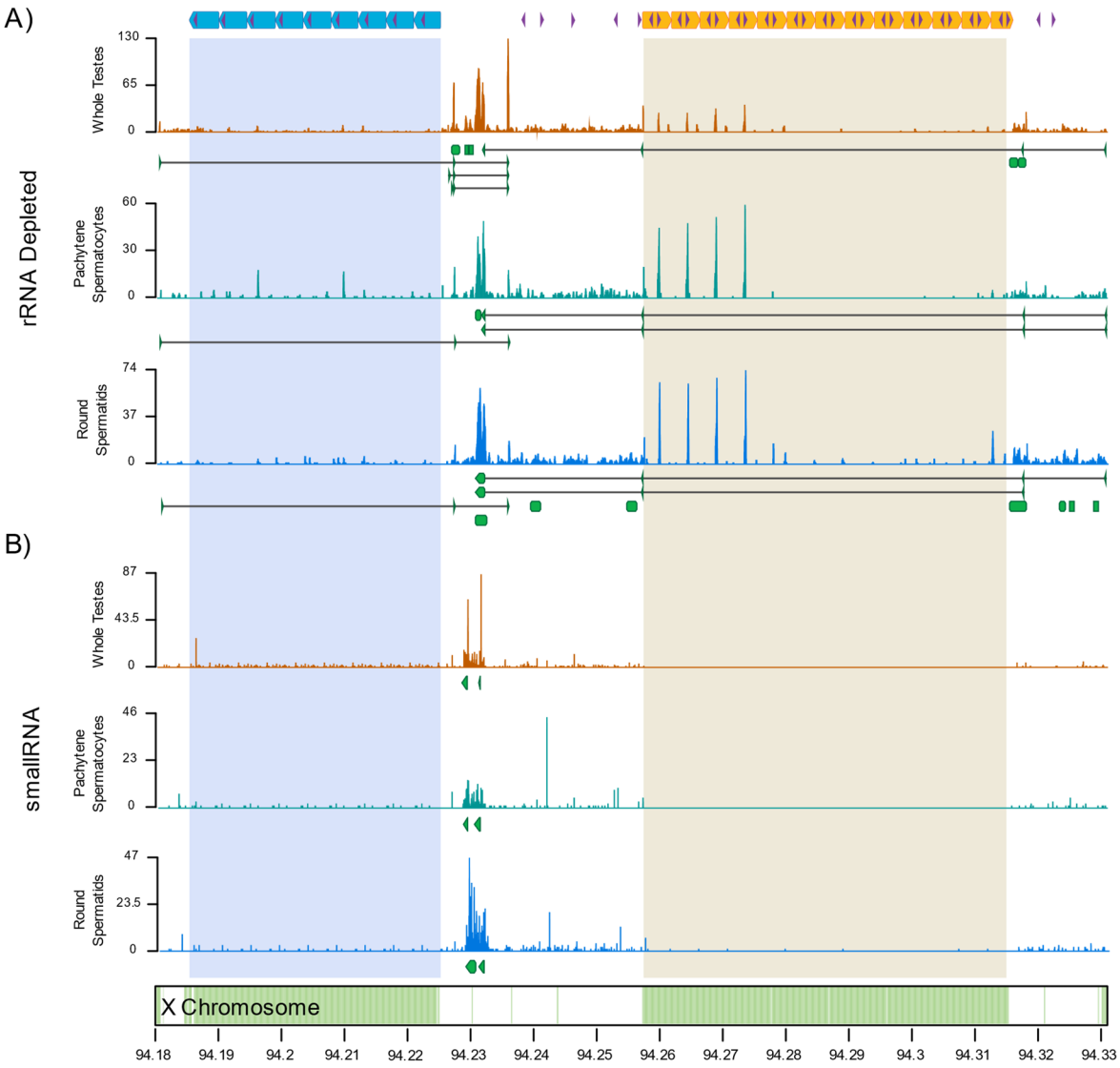


Figure 12. Transcriptional activity of *DXZ4* during domestic cat male meiosis.

RNA-Seq data from A) rRNA depleted and B) smallRNA libraries generated from seminiferous tubules, sorted pachytene spermatocytes, and sorted round spermatids from a fertile domestic cat. Green regions along the X chromosome ideogram (bottom) represent CpG islands. Shaded regions across tracks represent the boundaries of the *DXZ4* repeat arrays (top). The x-axis corresponds to X chromosome coordinates (in Mb) in the single haplotype domestic cat genome assembly. The y-axis represents raw coverage scaled by maximum coverage on a per track basis.

The domestic cat *RAST* and Jungle cat RA-Spanning Transcript 1 (*RAST1*) were highly similar, while the Jungle cat specific RA-Spanning Transcript 2 (*RAST2*) differed by a single exon in the region upstream of repeat A, suggesting divergence of exon usage between the two species (Figure B3.11). We observed 96% sequence identity across alignments between conserved exons in the domestic cat *RAST* and Jungle cat *RAST1/2*. Unlike the two Jungle cat *RAST* isoforms, all RB Spanning Transcripts (*RBST1*, *RBST2* and *RBST3*) exons were conserved between the two species and differed only in exon length, resulting in 99%, 98% and 96% sequence identity. Despite similarities between the larger, repeat-spanning transcripts, smallRNAs were not annotated across the Jungle cat *DXZ4* region. However, we cannot rule out sample-specific bias as only a single testis was analyzed. The Jungle cat lacks annotations across the spacer region, although it appears to be depleted for smallRNA across the entire *DXZ4* locus except for RB (Figure B3.12). Closer inspection of the RB-1 repeat unit in both species also revealed differences in expression peaks for the rRNA-depleted libraries. In the Jungle cat, we observed large coverage peaks within the two most proximal repeat units (Figure B3.13). Although similar to the ~250 bp peaks expressed across RB 1-4 in domestic cat, peaks in the Jungle cat are larger (~350 bp) and differ in the total number of repeats exhibiting expression, as well as the position of transcription within each repeat unit.

In summary, we observed transcriptional activity across *DXZ4* in both Jungle cat and domestic cat whole testes, as well as domestic cat sorted germ cell populations. The felid *DXZ4* locus transcribes conserved lncRNAs that are orthologous to human *DANT1* and *DANT2*. We observed several clear differences in ncRNA and smallRNA expression across the RB repeat unit between the two cat species. However, we cannot rule out that these

apparent differences are the result of assaying expression from Jungle cat whole seminiferous tubules and not sorted germ cells. Nonetheless, our transcriptional analyses suggest that feline *DXZ4* normally escapes MSCI and plays an unknown functional role during male meiosis, possibly related to transcriptional silencing of the sex chromosomes.

3.2.5 **DXZ4** *Methylation and Expression in Backcross Hybrids*

Like other macrosatellites, *DXZ4* is subject to regulation via direct DNA methylation of enriched CpG sites. *DXZ4* was first associated with XCI when Giacalone (1992) reported that the hypomethylated state of CpG islands on the human inactive X chromosome (Xi) contrasted with the surrounding hypermethylated heterochromatin, suggesting *DXZ4* escaped silencing. Further investigation by Chadwick (2008) verified that these epigenetic differences within the Xi influenced both transcriptional activity and binding of CTCF proteins critical to normal XCI (McLaughlin & Chadwick, 2011; Horakova et al., 2012b; Bonora et al., 2018).

A PCA based on methylation frequency (MF) for six testes samples revealed distinct clustering of the two sterile Chausies (JXD-019, JXD-061) separate from fertile Chausies (JXD-049, JXD-080) and domestic cats (FCA-4048, FCA-4415) (Figure B3.15). Methylation varied across chromosomes and fertility phenotypes (Figure B3.16), with fertile felids significantly hypermethylated genome-wide relative to sterile felids (Average MF: fertile=0.220, sterile=0.194, t=2.3, df=7.8, p=0.0484). We found variable levels of methylation across the entire *DXZ4* candidate locus as well as a suggestive, but non-significant, trend that sterile Chausies with Jungle cat *DXZ4* ancestry (n=2) were hypermethylated relative to fertile individuals with the domestic *DXZ4* haplotype (n, fertile Chausies=2, domestic cat=2) (t=4.8, df=1.5, p=0.0727) (Table C3.6). Mean MF of the sorted

pachytene spermatocytes and round spermatids revealed greater hypomethylation across the *DXZ4* region than fertile whole testis, consistent with a relaxed, open chromatin state during these stages of meiosis. Within *DXZ4*, we did observe significant hypermethylation of RA, but not RB, in sterile felids (RA: t=7.1, df=1.7, p=0.0318; RB: t=-1.1, df=1.1, p=0.4659) (Figure 13, Table C3.7). Expression profiles from fertile and sterile Chausie seminiferous tubules showed decreased expression across the entire locus (Figure B3.14) and significant downregulation (logFC = 2.9) of the domestic *RAST* in sterile hybrids, implicating RA interspecific variation in the failure of MSCI through *DXZ4* misregulation.

In summary, we observed reversal of *DXZ4* to a hypermethylated state in sterile seminiferous tubules, associating *DXZ4* activity with the fertility status of hybrid individuals. The observed hypermethylated state of the locus in sterile males is comparable to the hypermethylated and inactivated state of human *DXZ4* on the active X chromosome (Xa) in female somatic cells. Taken together, these different lines of evidence suggest that interspecific divergence at *DXZ4* leads to transcriptional and epigenetic misregulation of *DXZ4* in the testes of sterile hybrids, and contributes to the observed biomarkers of sterility: X chromosome-wide upregulation and meiotic arrest at pachynema.

3.2.6 Structural Conformation of the X-chromosome in Male Germ Cells

Previous studies revealed that the human, macaque, and mouse Xi exhibit a unique structural arrangement composed of two large super-domains forming a bipartite structure, with *DXZ4* functioning as the hinge region (Rao et al., 2014; Deng et al., 2015, Darrow et al., 2016). Knockout studies of *DXZ4* demonstrated that deletion of the locus was sufficient to disrupt the unique structural organization of the Xi in female somatic cells (Darrow et al., 2016; Giorgetti et al., 2016; Bonora et al., 2018, Bansal et al., 2020).

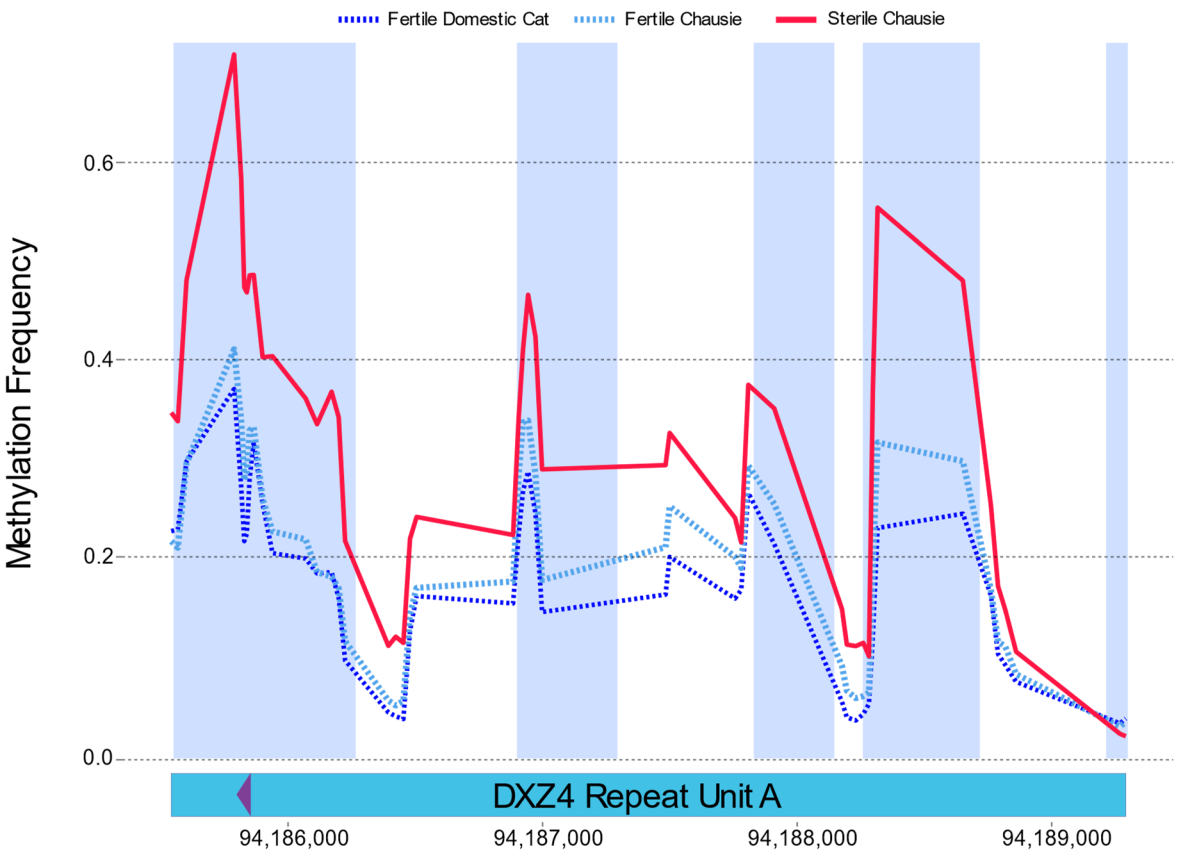


Figure 13. Methylation profiles across the *DXZ4* RA region in sterile and fertile hybrid testes.

Sliding window of 20-cytosine averages of methylation frequency (MF) with a 10-cytosine step comparing fertile felids (n=4, domestic cat=2, Chausie=2) and two sterile Chausies across tandem repeat A of candidate gene *DXZ4*. Regions highlighted in blue possess a window averaged p-value of ≤ 0.05.

In domestic cat, this bipartite structure is also maintained and clearly visible in Hi-C maps generated from female fibroblasts, suggesting *DXZ4* is functionally conserved in XCI of carnivores, in addition to primates and rodents (Figure B3.17) (Brashear et al., submitted).

Because of the interesting parallels between silenced X chromosomes resulting from MSCI and XCI, we sought to compare changes in chromatin conformation resulting from each process. We used previously published Hi-C data from a female F1 hybrid cell line (Bredemeyer et al., (2021) where the domestic cat haplotype exhibited features of chromatin organization characteristic of the inactive X, while the alternative haplotype maintained a structure analogous to the active X of male fibroblast cells. We hypothesized that this reflects skewing of X inactivation in the domestic cat haplotype of the female Bengal F1 hybrid, a phenomenon previously described in interspecific rodent crosses that were used to generate phased Xa and Xi Hi-C maps (Deng et al., 2015; Darrow et al., 2016). Thus, we used the domestic cat and Asian leopard cat X chromosomes to represent the female Xi and Xa state, respectively (Figure B3.18). We detected depletion of A/B compartmentalization between the Xa and Xi state, akin to mouse and human (Rao et al., 2014; Darrow et al., 2016), in addition to degeneration of both topologically associated domains (TADs) and formation of a bipartite structure (Figure B3.19). While Hi-C data from domestic cat pachytene spermatocytes and round spermatids did not reveal formation of an analogous bipartite structure, we did observe depletion of intrachromosomal interactions and TADs as spermatogenesis progressed, a feature originally described in the mouse (Figure 14A) (Alavattam et al., 2019; Patel & Kang, 2019). X chromosome A/B compartmentalization showed clear changes across stages of domestic cat spermatogenesis, suggesting that broader

features of large-scale nuclear organization may be conserved between silenced X chromosomes in both sexes (Figure 14B).

3.2.7 DXZ4 *is a Rapidly Evolving Macrosatellite*

The hypervariable nature of *DXZ4* observed here between different cat species, and described in humans, implicates copy number variability as a likely source of genetic incompatibility between closely related species. We postulate that this interspecific variation in copy number drives the aforementioned shift in methylation and transcriptional profiles across *DXZ4* in the testes of sterile hybrid males. Coincidentally, copy number dependent misregulation has been reported previously in another macrosatellite *D4Z4*. Like *DXZ4*, *D4Z4* maintains an inactive, hypermethylated state when surrounding chromatin is otherwise hypomethylated and active (Chadwick, 2009). However, this state is reversed when the total number of satellite repeat units falls below a certain threshold. A shift from hyper- to hypomethylation as a result of fewer repeat units leads to upregulation of genes within each *D4Z4* repeat unit, resulting in facioscapulohumeral muscular dystrophy (Hewitt et al., 1994, van Overveld et al., 2003).

Copy number dependent regulation of the *D4Z4* macrosatellite is just one example of a conserved regulatory mechanism referred to as repeat-induced gene silencing (RIGS) (Ogaki et al., 2020). The process of RIGS, originally observed in *Drosophila* and *Arabidopsis* transgenes, was proposed to protect the genome from transposons and other foreign sequences that rapidly duplicate and disperse throughout the genome (Garrick et al., 1998; Henikoff, 1998). Studies in mouse revealed that decreasing the copy number of certain transgenes would substantially increase the level of expression per copy, while increasing copy number would lead to increased suppression of the transgene.

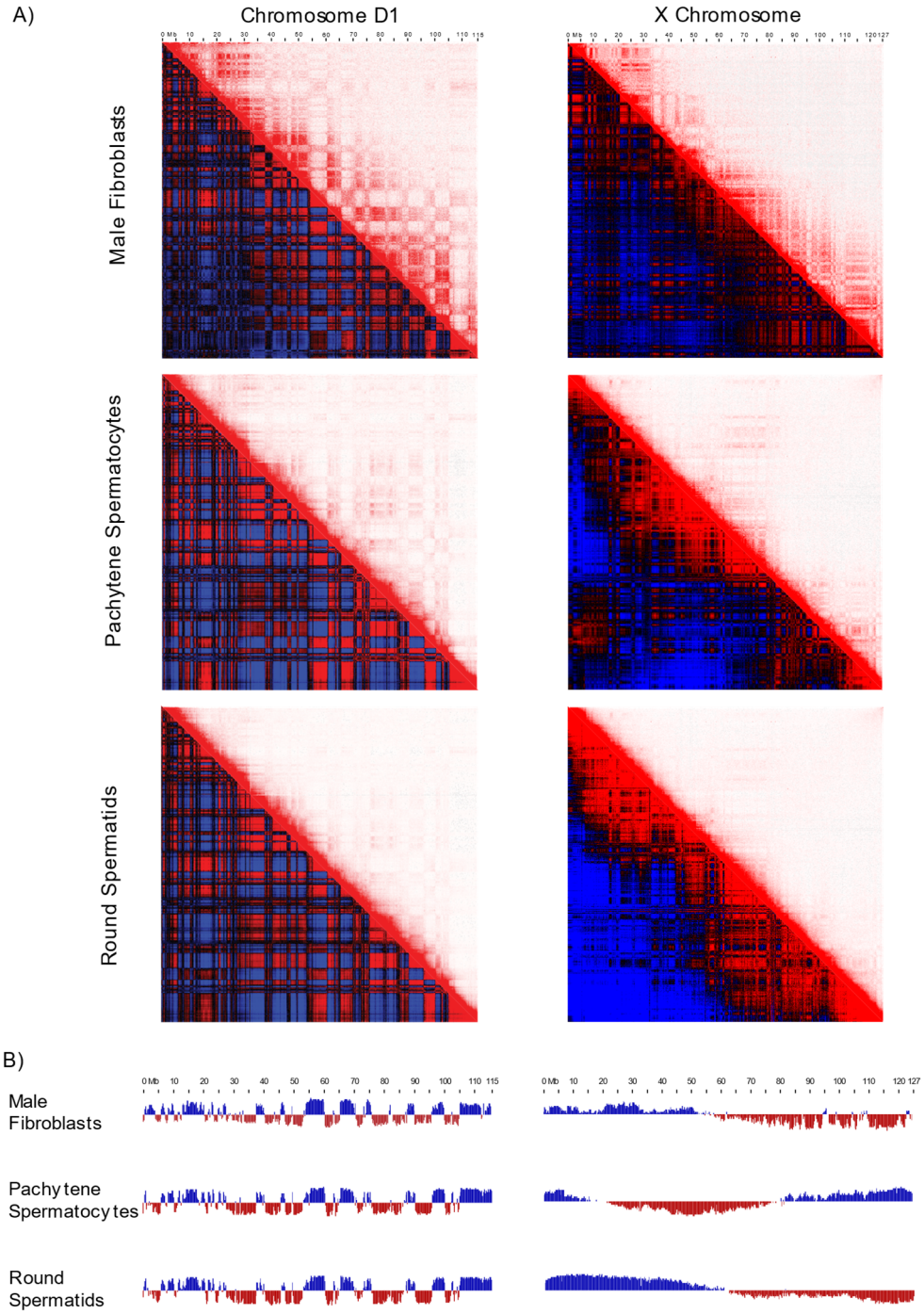


Figure 14. Chromatin conformation of the X chromosome in 3 different male domestic cat cell types.

A) Pearson's correlation (bottom left)/raw contact (top right) maps of a representative autosome, Chr. D1, and the X chromosome in fibroblasts, pachytene spermatocytes, and round spermatids (Pearson's and raw contact maps calculated at 250 kb resolution). B) Eigenvector tracks revealing changes in compartmentalization during felid meiosis. The association between eigenvector directions and A/B compartments were not established using histone modification profiles for either chromosome. Blue compartments along the X chromosome are assumed to be A (active) due to escape of the pseudoautosomal region from silencing during MSCI/PSCR.

Misregulation resulting from copy number variation might provide an intriguing general mechanism for how interspecific divergence might destabilize other complex macrosatellite regions with roles in developmental or reproductive processes.

Until the release of the human X chromosome telomere-to-telomere assembly (Miga et al., 2020), the mouse was the only mammal where *DXZ4* was fully represented in the genome assembly. This is likely due to the smaller and less complex structure of murine *Dxz4*, which is composed of 7 repeat monomers of varying length (3.8 and 5.7 kb), while human *DXZ4* contains between 12 and 120 repeat monomers of very similar length (3.0 kb) (Tremblay et al., 2011; Horakova et al., 2012a, Schaap et al., 2013) (Figure B3.20). Despite this dramatic structural divergence, *DXZ4* in both human and mouse maintains its putative organizational role forming the bipartite structure of the Xi generated by XCI (McLaughlin & Chadwick, 2011; Horakova et al., 2012b). Our assemblies of *DXZ4* loci from two cat species surprisingly revealed that while the felid structure was much more similar to human, it also differed significantly from both human and mouse orthologs by being composed of two divergent tandem repeats separated by a spacer sequence. The felid RA unit is likely orthologous to the human *DXZ4* repeat unit based on greater sequence identity and similar CTCF binding motif patterns. The second repeat array in felids, RB, is absent in both the human and mouse assemblies. The number and orientation of inter-repeat CTCF binding motifs in the mouse ortholog did not resemble those in either felid repeat array.

The compound nature of *DXZ4* repeat arrays in cats complicates our previous hypothesis that copy number incompatibility is the likely underlying cause of misregulation in the testes of sterile Chausies. *In silico* copy number estimates indicate that RA expansion in Jungle cat matches the expectations of RIGS, but RB does not. This relationship is

especially pronounced in established breed domestic cats like the Egyptian Mau, which were used in founding the Chausie hybrids used in the study. In Jungle cat, copy number of RA is always larger than RB, with the situation being reversed in domestic cat. Additionally, RA alone showed significant methylation and transcriptional differences between sterile and fertile Chausies, further supporting our hypothesis that variation at RA contributes to the hybrid sterility phenotype in Chausies.

We postulate that misregulation of RA could result in failure of MSCI through a number of mechanisms. The first implicates incompatibility between transcripts produced by the Jungle cat *DXZ4* allele and regulatory machinery utilized within the domestic cat background. Our investigation of Jungle cat and domestic cat transcript annotations revealed minimal divergence between conserved exons of Jungle and domestic Repeat A and B Spanning Transcripts (*RAST*, *RBST*) orthologs, however, we also identified unique Jungle cat isoforms that differ in exon usage (*RAST*) and length (*RBST*). Because the domestic cat *RAST* was downregulated in sterile backcross Chausies, it is possible that the Jungle cat *DXZ4* allele is either failing to express *RAST* at all, or expressing the exon variable *RAST2* isoform that is potentially ineffective within hybrids with a predominantly domestic cat genomic background. This hypothesis provides a plausible explanation for *DXZ4* hypermethylation within the sterile Chausie, as silencing of the orthologous human *DANT1-ATT* isoform was also shown to correspond with a hypermethylated, constitutive heterochromatin state in human embryonic stem cells (Figueroa et al., 2015).

A second conceivable mechanism for failure of MSCI through RA misregulation is through epigenetic interference between DNA bound methyl groups and CTCF proteins (Filippova et al., 2007). Hypermethylation of RA in hybrids could prevent binding of CTCF

proteins to sequence motifs present within each repeat unit, a situation comparable to *DXZ4*-embedded CTCF motifs on the human Xa (Chadwick 2008). While inspection of methylation data revealed a significant peak in methylation frequency across the CTCF binding motif of RA, future comparisons between sorted germ cells from fertile and sterile Chausies that apply ChIP-seq and Hi-C experiments are required to properly test these different hypotheses.

3.2.8 Conclusions

We identify the *DXZ4* macrosatellite as a novel mammalian X-linked candidate hybrid-sterility locus in an interspecific cross between domestic cats and Jungle cat, two species that diverged approximately 3-4 million years ago (Johnson et al., 1996; Li et al., 2016 & 2019). *DXZ4* is a compelling candidate hybrid sterility locus based on its structural complexity, known role in sex chromosome silencing, and peculiar biology. The rapid mutation rate of *DXZ4* fits the theoretical requirements of a 'speciation gene', where evolutionary labile satellite DNA represents a common template upon which genetic incompatibilities rapidly arise. The implication that rapid evolution of sex-linked satellite elements may contribute to infertility and inviability, and thereby promote speciation, has received support in previous studies of *Drosophila* interspecific hybrids (Bayes & Malik, 2009; Ferree & Barbash, 2009; Ferree & Prasad, 2012). The satellite-encoded hybrid sterility gene *OdsH* from *D. mauritiana* ChrX acts as a sterilizing factor in male *Drosophila simulans/mauritiana* hybrids by associating with the heterochromatin of the *D. simulans* Y chromosome, whereas the *D. simulans* ortholog does not (Bayes & Malik 2009). Similarly, it was predicted that interspecific satellite divergence in mammals would also play a significant role in reproductive dysfunction and speciation (Yunis & Yasmineh, 1971).

The discovery of the unique compound *DXZ4* repeat structure in felids immediately raises many questions regarding its function during XCI, and potentially MSCI. Whereas RB is absent in humans and mice, what is its significance in felids? Is RB shared by all cat species? Is the possession of both RA and RB the ancestral or derived state for placental mammals? In this study, we limited our investigation of *DXZ4* transcription to expression in male germ cells. To fully assess the function of both *DXZ4* repeat arrays, future work will require transcriptional analysis of the locus in other cell types, particularly female embryonic stem cells undergoing XCI. It is also unclear how differing CTCF binding motifs of each repeat unit in felids might correspond to the long-range physical interactions formed between X-linked macrosatellite repeat (MSR) loci, *DXZ4*, *FIRRE* and *ICCE* in human XCI. Although conservation of the bipartite structure on the felid Xi suggests that these *DXZ4* interactions are maintained, our Hi-C resolution was incapable of detecting superloops between the MSR loci in cats and will require improved Hi-C resolution or more sensitive, targeted 4C-approaches. Additional comparative sequencing and functional studies are necessary to more fully understand the full repertoire of roles played by *DXZ4* in mammalian biology.

3.3 Materials and Methods

3.3.1 Chausie Hybrids

The Chausie cat breed was originally established by hybridization between the domestic (*Felis catus*) and Jungle cat (*Felis chaus*), two species that diverged ~3 million years ago (Mya) (Li et al., 2019). Like Bengals and Savannahs, early generation hybrid Chausies follow Haldane's rule by exhibiting hybrid male sterility, and require F1 females to be backcrossed to fertile male domestic cats, or to late generation fertile backcross Chausie males, as explained below. The number of generations required for rescue of male fertility

is dependent on parent species divergence times and is correlated to the percentage of wildcat ancestry maintained in male individuals (Davis et al., 2015; Allen et al., 2020). The Chausie is derived from the least diverged interspecific cross of the previously described breeds (Li et al., 2016, 2019). After 2-3 generations, litters begin producing fertile males, allowing breeding between hybrid individuals. Today's Chausie breed represents a population of breeding hybrid males and females from generations spanning F1s (typically females) to hybrids more than six generations past F1. Later backcross generation hybrid males, typically F3-F5, exhibit variable fertility between individuals. Determination of fertility for all Chausies was performed using one or both of two methods: breeding records and histopathology as described in Davis et al. (2015). Sterility was defined as repeated, confirmed matings with multiple proven breeder females over 1 or more years with no conception, whereas fertile individuals were defined by documented productive breeding with validation via pedigree records.

3.3.2 Histopathological Evaluation of Backcrossed Chausie Testes

Histopathological evaluation was performed on testes and epididymides from sexually mature males that underwent orchidectomy. Testes were laterally bisected and stored in Bouin's fixative and later transferred to formalin. Testes and epididymides were embedded in paraffin, sectioned, and stained with H & E. Histology was evaluated to determine the presence/absence of germ cells, stage of meiotic progression, and the presence/absence of normal sperm. Fertile individuals possessed seminiferous tubules and caput epididymides with large numbers of sperm displaying normal morphology. Histological data was available for all individuals utilized in RNA-Seq analyses.

3.3.3 RNA-Seq and Differential Expression Analysis

Testes from two sterile and two fertile backcrossed Chausies were obtained from sexually mature cats that underwent orchidectomy. Testes were bisected and seminiferous tubules isolated by dissociation with collagenase. RNA and DNA were extracted using Trizol Reagent (Applied Biosystems, 2010). Extracts were assessed for quality using the Agilent 2200 Tape Station System. Extracted RNA was used to generate both RNA-depleted and small RNA libraries (NEXTFLEX, Perkin Elmer) and sequenced on the Illumina HiSeq 2500 to obtain an average of 40 million paired end reads per sample. Examination of sequencing data included *FastQC v0.11.8* (Andrews, 2010) followed by adapter trimming using *Trim Galore! v0.6.4* (Babraham Bioinformatics, 2020). Trimmed reads were mapped to the single haplotype domestic cat assembly Fcat_Pben_1.0_maternal_alt (Fca-508: GCA_016509815.1) using *STAR v2.7.7a* (Dobin et al., 2016) with default settings. We used *SAMtools v1.9* (Li et al., 2009) to process the alignment into a sorted bam file for downstream analysis. Raw read counts were calculated from bam files using *HTSeq-count v0.13.5* (Anders et al., 2013). For comparison of genes expressed between sterile and fertile hybrids differential expression analysis was performed using the Bioconductor package *edgeR v3.32.1* (Robinson et al., 2009; McCarthy et al., 2012). For assessing X chromosome misregulation in sterile testes, we first established an expression profile for fertile testes and identified upregulated genes based on whether or not they were previously expressed in fertile testes. Chromosome-wide misexpression was tested for significance using a chi-squared test. Difference in upregulation between the X and autosomes was tested for significance using Fisher's exact test. Statistical tests for chromosomal enrichment of differentially expressed (DE) genes was performed based on expectations generated using

Markov chain Monte Carlo simulations. Gene Ontology enrichment analysis for biological processes was performed using *PANTHER v.14* (Mi et al., 2019, Ashburner et al., 2000).

3.3.4 Genome-Wide Association Study (GWAS) & Fine Mapping

A binary case-control genome wide association study was performed on a cohort of 23 fertile and 16 sterile backcross Chausie hybrids that were genotyped on the Illumina 63K Feline SNP array. Hybrids possessing a genotype call rate <0.9 were removed from further study. We searched for marker-based-associations meeting or exceeding the Wellcome Trust recommendations ($P_{uncorrected}$=5x10^{-5}; -log$_{10}$ P = 4.30) (Wellcome Trust Case Control Consortium 2007). Notably, this significance threshold is conservative considering the polygenic nature of hybrid sterility and the modest SNP density of the Illumina feline SNP array, with the Wellcome Trust recommendations developed for a much higher density SNP array (Human Affymetrix 500 K GeneChip; see Wellcome Trust Case Control Consortium 2007). All marker-based association analyses were carried out using a mixed linear model, as described and implemented in EMMAX (Kang et al., 2010; Segura et al., 2012), and were executed in the SVS environment (Golden Helix, Version 7.7.6) as described (Davis et al., 2015, Seabury et al., 2017).

Fine mapping of a critical interval was performed using bi-allelic SINE insertion marker assays that distinguish the domestic cat from Jungle cat X chromosome. SINE INDELS homozygous in the reference female domestic cat and homozygous null in the female Jungle cat pseudo-reference were identified by aligning Jungle cat Illumina short reads (SRA accession no: SRX1058146) to the v8.0 “Cinnamon” reference genome (Li et al., 2016b). 35 candidate fixed SINE insertions in the domestic cat were used to design primers for PCR-testing (following Murphy & O’Brien, 2007) in 10 random-bred domestic

cats and eight Jungle cats (six captive animals from zoos in Central Asia and two from Thailand) and resolved on 2.5% high-resolution agarose gels. A final set of 27 SINE markers that were informative for ancestry inference (i.e., homozygous insertions in all domestic cats and absent (null) in all Jungle cats) were used to assay hybrid backcross Chausies.

3.3.5 Jungle Cat Genome Assembly

To create a Jungle cat genome reference assembly, we extracted DNA from an early passage, primary fibroblast cell line, derived from a captive-born male Jungle cat, housed originally at the Blijdorp Zoo, Rotterdam. To resolve assembly issues resulting from large, repetitive, and highly polymorphic regions, a male individual was selected to generate an effectively haploid (excluding the pseudoautosomal region) X chromosome assembly. The same cell line was used to generate PacBio, Illumina, and Hi-C libraries. The PacBio SMRT library was sequenced on the PacBio Sequel platform, resulting in 7,594,421 reads with an N50 subread length of 25.89 kb, corresponding to ~50x coverage relative to the length of the domestic cat reference assembly, felCat9 (Buckley et al., 2020). Illumina sequence libraries were generated using the NEB-Ultra II kit and sequenced on the Illumina NovaSeq6000, yielding ~33x coverage, and were used for polishing of the PacBio contigs. Sequence adapters were trimmed using *Trim Galore!* and read quality assessed using *FastQC*. Hi-C libraries were prepared following the Ramani (2016) protocol for *in situ* DNase Hi-C. Libraries were sequenced on the Illumina NovaSeq6000, yielding ~1.5 billion paired end reads (180x Coverage). Hi-C reads were trimmed using *Trim Galore!* with additional commands *--clip_R1 10 --clip_R2 10 --three_prime_clip_R1 15 --three_prime_clip_R2 15* selected to remove 10 and 15 bp from the 5' and 3' end of each read, respectively. Contig assembly was performed with *NextDenovo v2.2-beta.0*

(github:Nextomics/Nextdenovo) with the configuration file (.cfg) altered for inputs: *minimap2_options_raw = -x ava-pb*, *minimap2_options_cns = -x ava-ont*, and *seed_cutoff=* 13944.

Mitochondrial Genome Assembly

Prior to polishing raw contigs with short reads, we screened the assembly using BLAST to detect the presence of a complete Jungle cat cytoplasmic mitochondrial (*cymt*) sequence. We performed this check because previous studies had verified the presence of nuclear mitochondrial (*numt*) sequences within the domestic cat and *Panthera* nuclear genomes (Lopez et al., 1996; Kim et al., 2006). Despite residing in distinct cellular environments, *numt* and *cymt* sequences are highly similar, with *cymt* reads being far more abundant due to higher copy number per cell. Failure to account for *numt* sequence during the assembly polishing step with short-read data would potentially result in conversion of *numt* to *cymt* sequence if a more similar *cymt* "bait" sequence is excluded from the assembly (Rhie et al., 2020). Our BLAST analysis of the Jungle cat assembly identified a single significant hit to contig ctg000098, a chimer of tandemly duplicated jungle cat *cymt* sequences and chromosome D2, which harbors the domestic cat *numt* sequence. (Antunes et al., 2007). To isolate the full-length assembly of the Jungle cat *cymt* sequence we performed an alignment between ctg000098 and the previously published Jungle cat *cymt* sequence (Li et al., 2016) using LastZ (Harris 2007). We then used changes in percent identity across the alignment to distinguish and extract the jungle cat *cymt* sequence from surrounding chromosome D2 *numt* sequence.

Contig Polishing, Purge-Dups, and Quality Control

We polished the raw assembly contigs with *NextPolish v1.3.0* (Hu et al., 2020), using the *NextDenovo* corrected long reads, and Illumina short reads. Notable changes to the *NextPolish* configuration file included: *genome_size=auto*, and *task=best*, which instructs the program to perform 2 iterations of polishing. We used default settings for both *sgs* and *lgs* read mapping options except for indicating PacBio input with *minimap2_options= -x map-pb*. Following polishing, *purge-dups v.1.0.1* (Guan et al., 2020.) was used to remove haplotigs and smaller, low coverage contigs. Basic assembly stats were generated using *QUAST v5.0.2* (Mikheenko et al., 2018) with the minimum contig length set to *-m 1* and the *--fast* run option selected. To assess genome completeness, *BUSCO v4.0.6* (Simão, et al., 2015) was run using the *-m* genome setting with *-l mammalia_odb10* database selected (9,226 single copy genes). Visual assessment of the haploid assemblies was performed through alignment to the single haplotype domestic cat assembly Fca-508 using *nucmer* (*mummer3.23* package; Marçais et al., 2018) with default settings. The resulting delta file was used to generate a dot plot visualized in *Dot: interactive dot plot viewer for genome-genome alignments* (DNAnexus).

Y Chromosome Contig Identification and Isolation

To identify Y chromosome contigs within the Jungle cat assembly, we used two parallel approaches. The first is based on mapping female Jungle cat Illumina reads to the male Jungle cat PacBio contigs. This approach relies on the expectation that female reads will lack Y chromosome sequence and thus allows for selection of contigs based on zero or limited read coverage across their length. Illumina sequence reads from a female Jungle cat were previously generated by Li et al., (2016), accession number SRR2062187, and aligned

to the contigs using *bwa mem v0.7.17* (Li et al., 2009) with default settings. The alignment output was piped into *Samtools v1.9* (Li et al., 2009), where it was converted into bam format, sorted, and indexed. To annotate coverage across the assembly we used the *genomecov* tool of the *BEDTools suite v2.29.2* (Quinlan et al., 2010) with the *-d* (for per base coverage) and *-bga* (for regional coverage as bedgraph) options selected. The results were output in bedgraph format and all contig nucleotide positions identified as having a female read coverage threshold of 15x, 10x, 5x or 0x, and extracted into separate lists. For each coverage list *R v.3.5.1* (R Core Team) was used to calculate the percent of positions in each contig at, or below, the coverage threshold. This was done by summing the total number of positions annotated from *BEDTools* for each contig, and dividing by the contigs total length. Next, to begin screening for Y contigs, we extracted contigs with 70-100% of their nucleotide positions within the threshold coverage. The first step in determining this cutoff was to identify Y contigs using known domestic cat Y chromosome sequences. To do this we used NCBI's basic local alignment search tool *(BLAST) v2.9.0* (Altschul et al., 1990) command line application with options *-culling_limit 1, -evalue 1e-25, and -perc_identity 85* specified to align domestic cat mRNA sequences (Murphy et al., 2006; Pearks Wilkerson et al., 2008; Li et al., 2013) and ampliconic Y chromosome BAC clones (Brashear et al., 2018) to the Jungle cat assembly custom BLAST database. Because the Jungle and domestic cat diverged fairly recently (~3mya) (Li et al., 2016a) the *megablast* algorithm was used to generate alignments. *GNU parallel* (Tange, 2011) was used to BLAST multiple queries simultaneously. Finally, the average percent positional contig coverage identified by single copy mRNA sequences was averaged and +/-25% added to determine the percent position coverage cutoff. This threshold was further justified by plotting a histogram for the data

where we observed a bimodal distribution for the number of contigs either lacking or possessing female read coverage across most of their lengths**.** This result suggests the cutoff was sufficient for capturing Y sequence contigs while avoiding autosome/X linked contigs. Finally, *megablast* was used to align contigs with 70% of their nucleotide identity possessing 15x coverage or less to the felCat9.0 assembly (Buckley et al., 2020). For this step, only the top blast hit was output using command line options *-num_alignments 1*, *-max_hsps 1*. To avoid false hits to repetitive elements both the felCat9.0 custom database and Jungle cat contigs were repeat-masked using *RepeatMasker v4.0.9* (Smit et al., 2013-2015) with default settings and option *-species felidae* selected. Using these results, we manually selected additional contigs presumed to be Y linked based on a combination of alignment percent identity to fca-9.0 autosomal sequences, percent nucleotide positions covered by female reads, and total sequence length. All contigs identified in this way were merged into a single list and removed from the assembly using *seqTK subseq v1.3* (Heng, 2018). These extracted sequences were then scaffolded using HiC libraries using two different approaches. The first was a more automated approach using *SALSA v2.2* (Ghurye et al., 2017; Ghurye et al., 2019) with reads preprocessed and mapped following the esrice slurm-hic pipeline run manually (github:esrice/slurm-hic). *SALSA* was run on the resulting bedfile with default options except *-e DNASE, -m yes*. This was followed by manual curation of Hi-C read mapping using *Juicer v1.5.7* (Durand et al., 2016b), followed by scaffolding with *3d-dna v180922* (Dudchenko et al., 2017) and *Juicebox Assembly Tools* (*JBAT*, *Juicebox v1.11.08*) (Durand et al., 2016a).

Scaffolding

Polished contigs (excluding those removed by purge-dups or identified as Y) were first scaffolded using Hi-C data and *SALSA v2.2* with parameters *-e none -m yes.* Inspection

of the SALSA scaffolds was performed using *QUAST*, *nucmer*, and *Juicebox*. Following *SALSA*, *RagTag v1.0.1* (Alonge et al., 2019) was used to localize scaffolds to their respective position in the chromosome length single haplotype domestic cat assembly Fca-508. Selected *RagTag* parameters included *–remove-small, -f 10000* and *-j unplaced.txt*, a text file of scaffolds for *RagTag* to ignore based on their small size and identification as repetitive sequence in the *nucmer* alignments. *RagTag* scaffolds were manually inspected with Hi-C maps generated using *Juicer v1.5.7* with option *-s none* selected for compatibility with DNase Hi-C libraries. Maps were visualized using *Juicebox v1.11.08* and *Juicebox Assembly Tools* with scripts from *3d-dna v.180922*.

3.3.6 Genome Annotation

Repetitive sequence annotation was performed with *RepeatMasker v4.0.9* with *-excln* and *-species cat* selected to identify and annotate repetitive regions of both genomes while ignoring gap sequence. To estimate indel rates and quantify repeat expansion and contractions we ran *Assemblytics v1.2.1* (web-based) (Nattestad and Schatz, 2016) with a unique sequence length requirement of 10,000 on nucmer alignments between domestic single haplotype assemblies. Because of the high sequence similarity between the domestic and Jungle cat genomes, we used *Liftoff v1.4.2* (Shumate and Salzberg, 2020) to perform an annotation lift over between the current felCat9 reference assembly (Buckley et al., 2020) with single copy Y chromosome sequence (Li et al., 2013) and the Jungle cat *de novo* assembly. Default parameters were used for all arguments except for calling *-copies* with *-sc 0.95* to identify extra copies of genes not previously annotated in felCat9.

3.3.7 **DXZ4** ***Repeat Unit Analysis and*** **in silico** ***Copy Number Estimation***

Identification and isolation of *DXZ4* repeat units was performed manually using GC content traces, CTCF motif annotations, and self-self dotplots for the region using *Geneious Prime v2021.0.3*. CTCF motifs were annotated using the *Geneious Annotate & Predict* tool with a sequence motif of GAGTTTCGCTTGATGGCAGTGTTGCACCACGAAT, based on the Horakova (2012a) conserved CTCF motif logo, with the most prevalent nucleotide representative of each position. A max mismatch of 13 was selected to allow for interspecific ambiguity within the motif. CTCF sites annotated using this method corresponded to the approximate location within human *DXZ4* repeat units originally described by Chadwick (2008). Once annotated and extracted, independent repeats were aligned using the *Mafft Multiple Aligner v1.4.0*. Neighbor-joining consensus trees were generated using the *Geneious Tree Builder* plugin and maximum likelihood trees generated using the Geneious *RAxML v8.2.11* (Stamatakis, 2014) plug-in with nucleotide model: *GTR GAMMA*, Algorithm: *Rapid hill-climbing* and Replicates: *500* selected. Mean within and between group distances for masked (10% gaps masked) *DXZ4* repeat unit alignments were calculated using *Mega-X v10.0.5* (Kumar et al., 2018).

In silico estimations of copy number were performed using short read mapping across collapsed tandem repeats (Lucotte et al., 2018). A representative unit from each of the *DXZ4* Repeat A and Repeat B arrays was selected from each of the three assemblies (domestic, Jungle, and Asian leopard cat) based on pairwise identity visualized using a neighbor joining tree and inserted in place of the full repeat array in the X chromosome of each assembly. We also included the first (most proximal) copy of RA (RA-1) due to its divergence from other RA units. lllumina short read data from 12 domestic cats (representing

both outbred and established breeds), 1 Jungle and 1 Asian leopard cat (**Table C3.**8) were mapped to their respective *DXZ4* modified genome assemblies using *bwa mem v0.7.17* (Li et al., 2009). Male individuals were selected when available to avoid confounding of copy number estimates across two haplotypes, which occurs in females. Alignment files were processed using *samtools fixmate*, *sort*, *markdup*, and *view* with *-q 20* and *-bh* specified (*v.1.10*; Li et al., 2009). Coordinates for the collapsed *DXZ4* regions and the single copy control gene *DMD* used in Lucotte et al. (2018) were recorded in a BED file and used to calculate the mean across feature coverage using *bedtools coverage v2.30.0* with *-mean* called (Quinlan &Hall, 2010). *DMD* was verified as single copy in the three genome assemblies using the lift over GFF file. Average coverage across the entire genome of each individual was generated from the filtered and sorted BAM file using *bedtools genomecov* with *-d* selected, and used to calculate copy number across each repeat unit and *DMD*. We observed an average coverage of 0.5 across *DMD* for all X hemizygous male individuals, as expected. Female individuals also exhibited the expected *DMD* coverage of 1.0 and had repeat estimates divided by 2 to account for diploidy.

3.3.8 Reduced Representation Bisulfite Sequencing (RRBS)

We obtained testes from six felids (fertile domestic cat males=2, Chausie backcross males=4, two fertile, two infertile) and two domestic cat germ cell populations (pachytene spermatocytes=1, round spermatids=1), which were sequenced at an average depth of 22.2- and 20.2-fold for autosomes and the X chromosome, respectively. Chausies were selected on the basis of having near-identical estimated % Jungle cat ancestries (**Table C3.**9). Genomic DNA libraries were prepared with a reduced representation bisulfite sequencing (RRBS) approach using the Msp1 restriction enzyme (Boyle et al., 2012) and the NEBNext

sample preparation kit (New England Biolabs) (**Table C3.**10). Each library was spiked with 1ng of enterobacteria phage lambda DNA as a non-methylated internal control for estimating bisulfite (BS) conversion (e.g. Lea et al., 2015). For all purification or size selection (100-400 bp) steps, we used AMPure XP beads (Beckman Coulter). Fragmented DNA was treated with bisulfite to convert unmethylated cytosines following the low DNA input protocol in the Qiagen EpiTect Fast Bisulfite Conversion kit (Qiagen USA). We enriched the converted DNA for adapter-ligated fragments with 12 cycles of PCR amplification and MyTaq Mix (Bioline Inc). Simultaneously, unique sequence tags were included in the amplification to barcode each library to enable pool of seven samples per lane of single-end (1x100nt) sequencing on an Illumina NovaSeq 6000.

Sequence pools were demultiplexed based on perfect sequence matches between expected and observed barcode sequence tags. We trimmed reads to remove low quality bases (Q<20), clipped remnant adapter sequences, and discarded reads that were <20 bp in length using cutadapt 1.8.1 (Martin 2011). To prevent loss of signal due to multimapping across the *DXZ4* locus we collapsed the Fca-508 genome assembly RA and RB repeat arrays into a single representative repeat unit for each array. The modified Fca-508 assembly was subsequently prepared with bowtie2 in BS-Seeker2 for read lengths bounded from 50-500 bp (Langmead 2010; Langmead & Salzberg 2012; Chen et al., 2010). We aligned processed reads to the built reference with bowtie2 and called methylation in BS-Seeker2. We calculated the methylation frequency (MF) per cytosine as the proportion of methylated cytosines from the total read depth per site (Chen et al., 2010). Bisulfite conversion efficiency was estimated by mapping each genomic library to the 48,502 bp phage lambda linear genome (NC_001416.1) and assessing the MF of the lambda-mapped data from

cytosines with at least 10x sequence coverage. Conversion rates were estimated as [1-average MF across the phage lambda genome]. We used the unite function in R v3.6.0 (2019) MethylKit (Akalin et al., 2012) package to apply a coverage filter to retain cytosines with a depth of coverage between 10x and below the 99.9% percentile. We included all methylation motifs (CG, CHH, and CHG) for analysis. We constructed a methylation matrix across all 8 samples. To assess library quality, we used the *prcomp* function in *R v3.6.0* (R Core Team 2019) to conduct a principal component analysis (PCA).

3.3.9 X-chromosome Candidate Region Analysis

We further scrutinized the *DXZ4* gene region (94,183,053-94,228,160 Mb) to evaluate methylation trends by plotting simple moving averages with the geom_ma function from tidyquant package. We also constructed sliding window plots using a cubic smoothing spline with the R packages GenWin and pspline to fit per-cytosine MF estimates (Craven & Wahba 1978; Beissinger et al., 2015). We used the generalized cross-validation (GCV) smoothing (λ) method to identify the inflection points of the spline to define the window boundaries. We averaged MF in the RA and RB *DXZ4* regions across groups of felids and conducted simple t-tests of differences per cytosine in sliding windows with 20 cytosines per window and a 10-cytosine step. We filtered the data to include only cytosines sequenced in all 8 felid samples to reduce intra-window disparities in sample size.

3.3.10 Domestic Cat Sorted Germ Cell RNA-Seq

Testes for germ cell sorting were collected from adult male domestic cats that underwent orchidectomy at the Texas A&M small animal hospital. Assumptions of fertility were based on maturity of animal and relative testes size. Testes were collected and immediately placed in cell culture media w/FBS prior to cell sorting. Target populations of

pachytene spermatocytes and round spermatids were collected using the STA-PUT method of sedimentation velocity (Go et al., 1971, Wang et al., 2001), snap frozen in liquid nitrogen and stored at -80°C. Purities of recovered populations of pachytene spermatocytes and round spermatids were ≥90% based on morphological analysis under phase optics.

For each germ cell population both RNA and DNA were extracted using Tri-Reagent (Applied Biosystems, 2010). An aliquot of DNA was used for RRBS sequencing and extracted RNA was used to create two different RNA-Seq libraries from each of the sorted populations. Additionally, RNA-Seq libraries were generated from whole seminiferous tubules of two domestic cats and one Jungle cat. For each sorted germ cell population, a technical replicate RNA-Seq library was generated for both protocols. The first RNA-Seq library was pre-processed using the NEBNext rRNA Depletion Kit and subsequently converted into an RNA-Seq library using the NEBNext Ultra Directional RNA Library Prep Kit (Illumina). The second library was generated using the NEBNext Multiplex Small RNA Library Prep Set (Illumina). The rRNA depleted libraries were sequenced on the Illumina HiSeq 4000. The smallRNA library was first size selected for 105-160 bp fragments using the Pippin-Prep and sequenced on the Illumina HiSeq 2500v4 in rapid mode for generation of 50 bp single end reads. Sequencing data was checked for quality and post-processed using *FastQC* and *Trim Galore!*.

3.3.11 RNA-Seq Read Mapping and Analysis

Trimmed reads were mapped to Fca-508 and *de novo* Jungle cat assembly (FelChav1.0) using *STAR* with default settings and *--outSAMstrandField intronMotif* to enable downstream compatibility with *Cufflinks*. Technical replicates were merged using *samtools merge* prior to annotation. *Cufflinks v2.2.1* (Trapnell et al., 2012) was used to

generate *de novo* annotations previously absent from the felCat9 lift over annotation as a result of germ-cell-specific expression or increased sensitivity to detection of lowly transcribed small and lncRNAs afforded by our RNA-Seq library protocols. SmallRNA libraries were annotated using *ShortStack v 3.8.5* with *-dicermax 31* and *-mincov 0.5rpm* specified (Axtell, 2013). Transcripts, annotations and read alignments were visualized and assessed using IGV and Geneious.

3.3.12 In situ *DNase Hi-C*

Hi-C libraries were prepared following the Ramani (2016) protocol for *in situ* DNase Hi-C. Fibroblasts from male domestic cat and sorted germ cell populations were fixed, converted into Hi-C libraries and sequenced to approximately 50x coverage. Previously published Hi-C data haplotype-phased from an F1 Bengal (Bredemeyer et al., 2021) suggested skewing of XCI towards the domestic cat X, based on comparisons between the haplotyped X chromosomes. The domestic cat X Hi-C map exhibits features characteristic of an inactive X, while structural conformation of the Asian leopard cat X was more similar to autosomes and the active X of male fibroblast cells. Thus, we used the domestic cat X and Asian leopard cat X to represent the female Xi and Xa state, respectively. All cell types were selected because they were representative of the X chromosome in a haploid state. The domestic cat Xi, pachytene and round spermatid cell types were selected to observe the inactive or partially inactive X chromosome in the two sexes. For comparison, male fibroblasts were selected to represent a single haplotype active X state. Maps were generated using *Juicer v1.5.7* with *-s none* selected for compatibility with reads from libraries generated using DNase as the fragmenting enzyme. Hi-C maps were visualized using *Juicebox v1.11.08*. Finer resolution Pearson's plots were generated using *juicer-tools*

pearsons with *-p KR* selected for normalization and *BP 250000* selected to set a bin size of 250 kb.

CHAPTER IV COMPARATIVE GENOMICS OF DXZ4 IN PLACENTAL MAMMALS

4.1 Introduction

Complex and highly repetitive regions are absent from nearly all genome assemblies and are often referred to as genomic "dark matter" (Sedlazeck et al. 2018; Ahmad et al. 2020). Variable number tandem repeats (VNTRs) often fall into this category as they harbor large copy-number differences within and between species, confounding their assembly from a diploid genome and eluding our understanding of their biological properties (Richard et al. 2008). Macrosatellites are a class of VNTR defined by particularly large repeat units, often several kilobases in length, high GC content and highly polymorphic copy number (Warburton et al. 2008; Tremblay et al. 2010 & 2016; Schaap et al. 2013). Macrosatellites have been implicated in key cellular processes through transcriptionally or spatially controlled chromatin remodeling and are often regulated through direct DNA methylation that can be disrupted by significant deviations in copy number that often contribute to disease (Hewitt et al. 1994; Chadwick 2009; Dumbovic et al. 2017).

In Chapter III we mapped a major effect X-linked locus associated with male hybrid sterility in an interspecific hybrid cat breed, and identified the macrosatellite *DXZ4* as a candidate hybrid-sterility locus (Bredemeyer et al. Submitted). However, until recently, the *DXZ4* locus was not completely or accurately assembled in the human genome and nearly all other mammalian genomes, limiting insights into its structure and functional evolution. Despite fragmentation in all previous domestic cat assemblies, *DXZ4* was fully represented in the domestic cat and Asian leopard cat single-haplotype assemblies (SHA) generated from an F1 Bengal (Chapter II) using the trio-binning method (Koren 2018; Rice 2020;

Bredemeyer et al. 2021) and a diploid male Jungle cat assembly where the X chromosome is hemizygous. In all three assemblies *DXZ4* was composed of a compound tandem repeat with two distinct and highly divergent repeat arrays (A and B). This compound structure differed from the simple, tandemly duplicated monomer previously reported in mouse and human genome reference assemblies (Horakova et al. 2012; Tremblay et al. 2011; Miga et al. 2020). The felid repeat A (RA) and the human monomer were described as likely orthologs, while repeat B (RB) was reported as absent in human and mouse. Despite this dramatic structural divergence, *DXZ4* maintains its role in establishing the unique chromatin conformation of the inactive X chromosome resulting from XCI in all three species (Deng et al. 2015; Darrow et al. 2016; Brashear et al. 2021) and was transcriptionally active during felid male meiosis (Bredemeyer et al. Submitted).

These observations raised a number of questions concerning *DXZ4* structure and evolution across placental mammals. Here, we broadened our previous characterization of *DXZ4* in felids through the generation of four additional, ultra-continuous felid single haplotype assemblies derived from long read sequencing of two F1 hybrids, the Liger and Safari cat. Our analysis of the resulting lion, tiger, Geoffroy's cat, and domestic cat assemblies demonstrate that the compound tandem repeat structure is conserved across the felid lineage, with interspecific copy number variability observed in both RA and RB. To determine whether the felid *DXZ4* arrangement is the exception or the rule, we expanded our exploration of the region to include high-quality genome assemblies spanning the mammalian phylogeny. Results of this comparison reveal rapid divergence in both the presence and organization of the two felid *DXZ4* repeat arrays across mammalian orders, with some species composed exclusively of repeat unit A, and others comprised of repeat

unit B, or a mixture of both. This level of large-scale repeat unit structure turnover together with repeat copy number evolution suggests a far greater scope of complexity and divergence than previously appreciated, especially for a locus involved in a conserved developmental process like XCI. As ultracontinuous genomes become available for a wider variety of organisms, "dark matter" regions previously missing from genomes might hold the key to answering pervasive questions in disease biology, genome organization, gene regulation, and speciation.

4.2 Results

4.2.1 Single Haplotype Assemblies

All details pertaining to raw sequencing output are included in Table C4.1. Genome assembly and sequencing metrics for the four felid single haploid assemblies (SHA) are found in Table 4. Haplotyped bases from each parental species were very similar for the Liger (Lion: 49.3%, Tiger: 50.7%) when using Illumina data from the original parents of the hybrid. However, parents for the Safari cat F1 were unavailable, restricting us to previously generated Illumina data from non-parental individuals. Safari cat haplotype phasing was therefore slightly skewed (Domestic cat: 46.2%, Geoffroy's cat: 53.8%), a phenomenon reported in our previous single haplotype assemblies from an F1 Bengal (Chapter II, Bredemeyer et al., 2021). Despite this skewing, the domestic cat assembly from the F1 Safari cat was captured in 103 contigs, (N50=92.7 Mb), a 10.5% increase in continuity relative to the single-haplotype domestic cat assembly generated from the F1 Bengal hybrid. The Geoffroy's cat was captured in 88 contigs (N50=104.5 Mb), making it the most contiguous felid SHA generated so far. The Geoffroy's cat assembly contained the largest contig of the four assemblies, with complete capture of chromosome A1, similar to the previously

published Asian leopard cat SHA (Figure 15). Contig alignment to one of the domestic cat single haploid assemblies (Fca-508: GCA_016509815.1) revealed that a majority of chromosome arms (Domestic Cat-126=72%, Geoffroy's Cat=63%, Tiger=61%, Lion=57%) were captured in single contigs. A small number (two domestic, two Geoffroy's cat, Figure B4.1; two tiger, four lion, Figure B4.2) of chimeric contigs were observed prior to scaffolding. Centromeres were captured within a single contig on 12 domestic cat, 9 Geoffroy's cat, 7 tiger and 7 lion chromosomes. The lion Y pseudoautosomal and single-copy Y chromosome regions were captured within a single 8.6 Mb contig. An additional 14 contigs totaling 4.3 Mb were also identified as belonging on the Y chromosome that comprise highly repetitive, ampliconic sequence (Table C4.2).

Using Hi-C and reference-based alignment approaches, we were able to obtain chromosome length scaffolds for all four assemblies, with scaffold N50s of exceeding 147 Mb in all cases (Table C4.3-6). The X chromosome was the least continuous chromosome within each of the female haplotype assemblies, accounting for 22-25% of assembly gaps.

Table 4. Assembly statistics for the single haploid assemblies generated from the Safari Cat and Liger.

Species	**Domestic Cat**	**Geoffroy's Cat**	**Tiger**	**Lion**
Chromosome Number (Auto, Sex)	18, X	17, X	18, X	18, Y
Raw Read Stats				
Read Count	8,218,109	10,389,709	10,489,759	10,570,078
Base Count (bp)	182,737,769,633	212,437,089,269	194,491,582,536	189,010,631,741
Subread N50 (bp)	32,222	31,036	28,447	28,059
Contig Assembly Stats				
Total Contigs	103	88	146	103
Largest Contig (bp)	172,124,406	239,106,607	166,130,000	166,870,000
Ungapped Assembly Length (bp)	2,425,722,928	2,426,362,316	2,408,668,598	2,297,542,863
N50 (bp)	92,686,623	104,474,415	74,360,613	77,781,637
BUSCO (mammalia_odb10)				
Single-Copy	8,589	8,580	8,585	8,372
Duplicated	27	22	29	33
Complete	8,616	8,602	8,614	8,405
Percent Complete	93.39%	93.24%	93.37%	91.10%
Fragmented	156	160	152	150
Missing	454	464	460	671
Percent Present (Comp+Frag)	95.08%	94.97%	95.01%	92.73%
Scaffold Assembly Stats				
Total Scaffolds	70	46	74	53
Primary Assembly Length (bp)	2,425,730,028	2,426,370,816	2,408,678,698	2,297,552,363
Total Gaps	39	45	65	55
N50 Scaffold (bp)	148,491,486	152,606,360	146,942,463	147,402,474

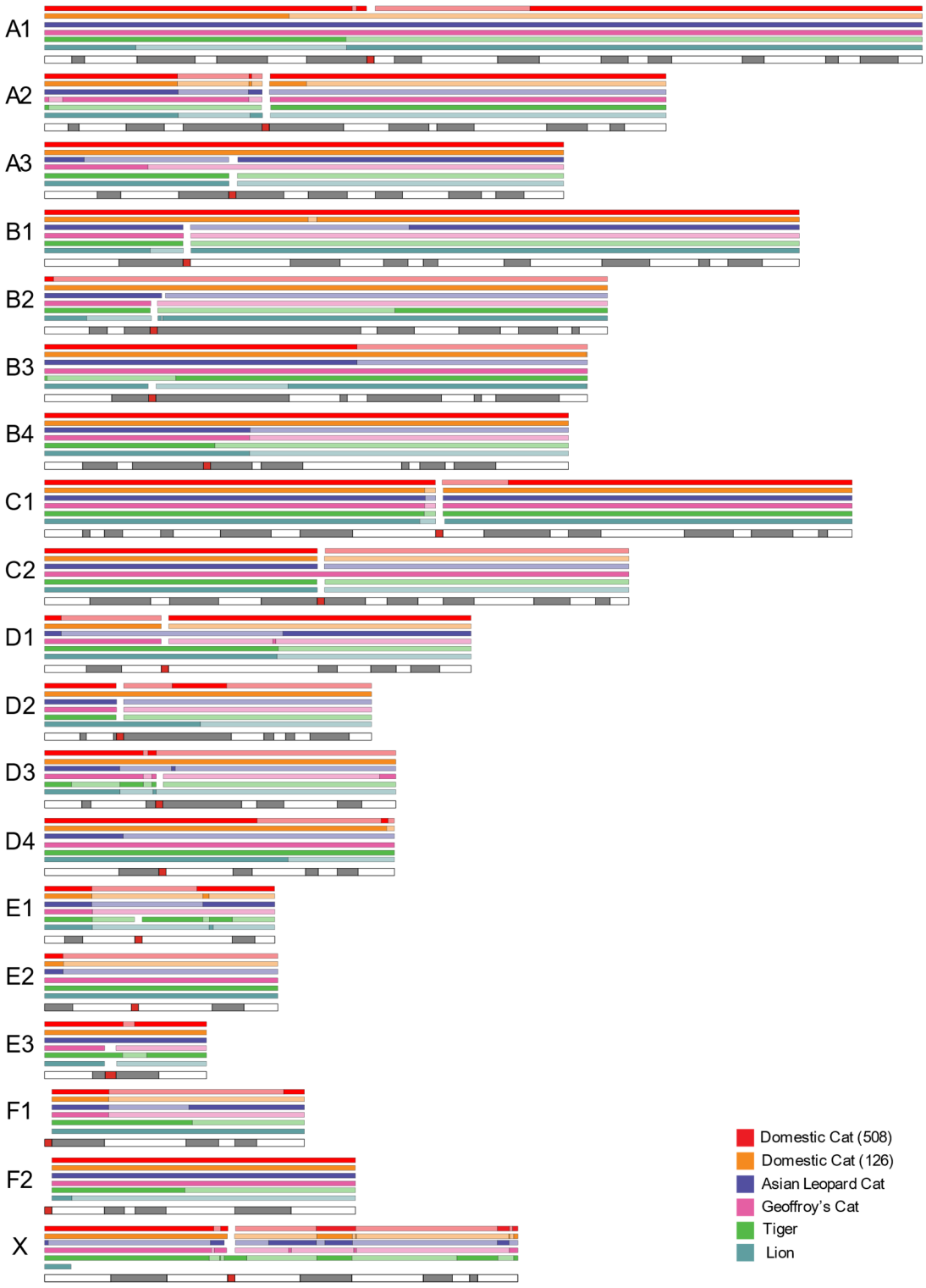


Figure 15. Contig alignment of six felid single haplotype assemblies to the felCat9 reference genome assembly.

All ideograms are based on the domestic cat (Cho et al. 1997; Davis et al. 2009). G-banding is represented by dark bars and centromeres by red bars. Bars above ideograms are colored by species and represent contigs > 1 Mb. Breaks between contigs are indicated by a black line and shift in color contrast.

While the felid karyotype is highly conserved, South American cats belonging to the Ocelot lineage possess one fewer chromosome than other felid species due to fusion of acrocentric chromosomes F1 and F2 into a single metacentric, C3 (O'Brien 2020, Atlas of Mammalian Chromosomes, 2nd Edition). Presence of C3 was validated in the F1 Safari Cat cell lines using g-banding karyotype analysis (Figure B4.3). Chromosome C3 was assembled into a single scaffold with Hi-C data.

4.2.2 DXZ4 *in Felids*

Expansion of our analysis to include the *DXZ4* locus from a larger sampling of the felid phylogeny revealed conservation of its canonical X chromosome location downstream of *PLS3* and upstream of *AGTR2* (Horakova et al, 2012) and capture within a single contig (Figure B4.4). Closer inspection of these regions revealed conservation of the RA/RB *DXZ4* array structure and CTCF motif orientation reported in our previous felid assemblies (Figure 16). The interspecific copy number variation reported previously was also observed with the inclusion of the new assemblies. A maximum likelihood phylogeny based on 10% masked alignments from repeat units of all SHAs and the Jungle cat recapitulated previous results (Figure 17). RA and RB form reciprocally monophyletic groups with an even larger between-group mean *P*-distance (0.67) than reported previously (0.42) (Table C4.7). The mean *P*-distance across all RA monomers was much lower: 0.0372, while RB monomer *P*-distance was 0.0255. The most proximal RA monomer (RA-1) once again formed a monophyletic clade apart from other RA monomers, which grouped by species. Most of the divergence driving monophyly of the RA-1 repeat is located towards the proximal end of the monomer, encompassing the CTCF binding site. Together these characteristics suggest unique selective constraints on RA-1. The average length of RA and RB across felid species

was 4,615 bp +/- 119 bp (standard deviation, SD) and 4,644 bp +/- 32 bp, respectively (Table C4.8 & 9). Interestingly, between the two arrays the average monomer length differs by only 29 bp despite the large genetic divergence between them, suggesting both felid *DXZ4* monomer types may be under similar physical constraints. Closer inspection of the tiger RA units showed that they contained a copy number variable microsatellite sequence (GGGAA) embedded within each monomer. Copy number ranged between 64 and 124 copies and accounted for the observed increase in length relative to other felids (Table C4.10, Figure B4.5). Sequence divergence of monomers within RA and RB arrays was low (RA (excluding RA-1) *P*-distance=0.0010, SD = 0.0006, RB *P*-distance=0.0015, SD =0.0007). Each species' repeat unit was monophyletic, and as a whole tracked the felid species tree, consistent with concerted evolution in both repeat arrays homogenizing sequence variation (Table C4.7).

The intervening spacer sequence was also quite similar in length across felid species (Table C4.11). Multispecies sequence alignments between the regions revealed 5 conserved CTCF binding motifs, and several lineage specific motifs (Figure B4.6). The spacer sequences also exhibited much greater sequence conservation than observed for monomers of RA/RB (Table C4.7).

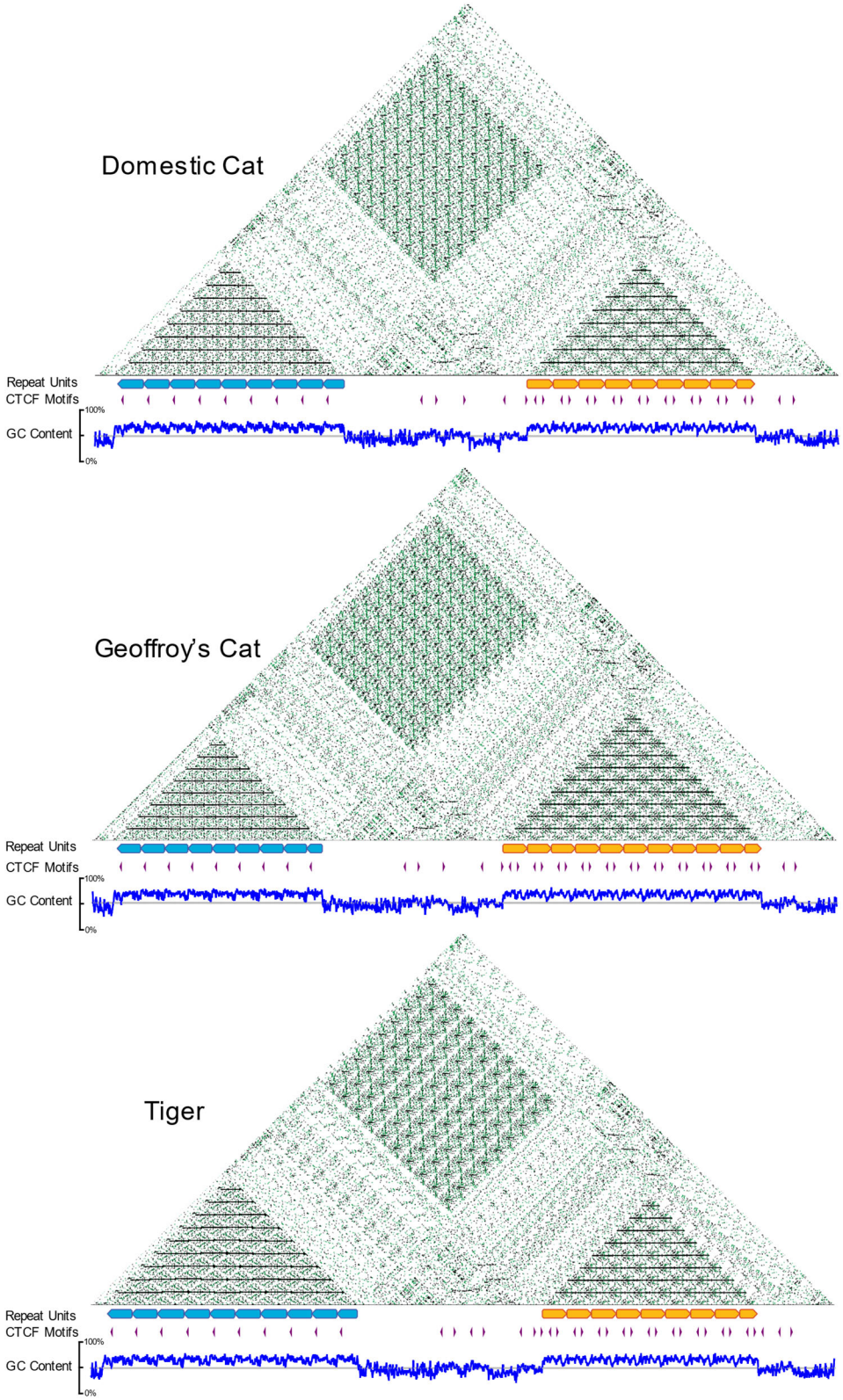


Figure 16. Dichotomous structure of *DXZ4* is conserved across additional felids and the Panthera lineage.

Self-dot plots and CTCF motif annotations reveal that macrosatellite *DXZ4* is composed of two distinct tandem repeats divided by a conserved spacer sequence in all cat assemblies.

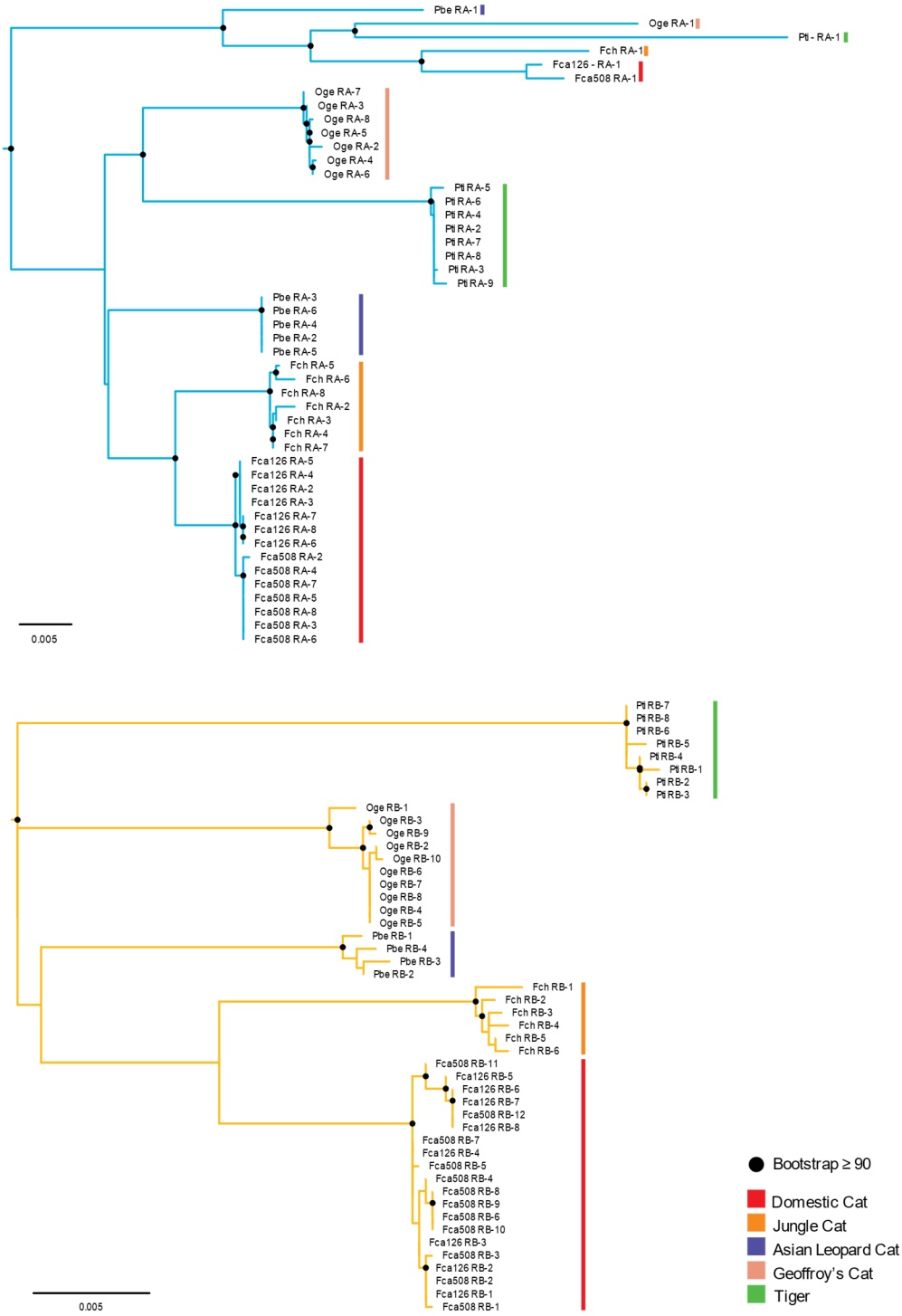


Figure 17. Expanded phylogenetic analysis of felid *DXZ4* repeat monomers.

Maximum likelihood phylogeny generated from an alignment of felid DXZ4 monomers. RA and RB trees were pruned from a larger tree rooted using a *DXZ4* repeat monomer from the human telomere to telomere assembly (Miga et al., 2020) (Not shown). Monomers are represented from domestic cat assemblies derived from both F1 Bengal and Safari hybrids (Fca508 and Fca126, respectively) and both associated with red. Additional assemblies are color coded accordingly. Black dots represent a bootstrap support value ≥90.

4.2.3 DXZ4 Across Placental Mammals

The selected mammalian assemblies provide a diverse set of X chromosomes with large differences in length and centromeric positionings both within and between superordinal clades. We defined canonical *DXZ4* positioning as residing between *PLS3* and *AGTR2* and further classified the status of *DXZ4* as complete, incomplete or absent (Figure 18). The canonical *DXZ4* region is present in all species excluding mouse, cow and yak where in all three instances *AGTR2* is located upstream of *PLS3* by 53 Mb, 120 Mb and 60 Mb, respectively. In mouse the *DXZ4* ortholog has been described previously and is immediately upstream of *Pls3*, similar to human and felids (Horokova 2012, Bredemeyer et al., Submitted). In cow and yak we observed a small cluster of CTCF binding motifs upstream of *AGTR2* that we identified as *DXZ4*-derived using BLAST alignment (Figure B4.7A). In sheep and goat, the canonical *DXZ4* region was resolved but divided by the embedded *KLHL4* gene. Like the other bovids, sheep and goat possessed *DXZ4*-derived, non-repetitive CTCF clusters upstream of *AGTR2* (Figure B4.7B). In all bovid assemblies, regions surrounding both *PLS3* and *AGTR2* were fully resolved but lacked tandem repeat structure, therefore we identified *DXZ4* as 'Absent' and excluded them from subsequent analysis.

Self-alignment dotplots generated from remaining assemblies revealed satellite repeat structure and the presence of distinct monomers in all species but the chimpanzee, where a gap occupied the region corresponding to the entire human macrosatellite sequence (Figures B4.8-11, Figure B4.12). We identified the remaining assemblies as 'Unresolved' if the *DXZ4* region was broken by at least one gap (Figures B4.13 and B4.14) and 'Resolved' if unbroken (Figure B4.15). We were able to annotate repeat monomers for each of the

assemblies except sloth, elephant and rabbit (Table C4.12), due to nested repeat structure and variation in monomer length. The rabbit in particular was highly expanded, with a large inversion of the macrosatellite bordered by two gaps, suggesting possible misassembly of the region and likely underrepresentation of monomers (Figure B4.16). While repeat units were not obtained for these assemblies, the orientation of CTCF binding motifs resembled the felid RA in elephant and sloth, and felid RB in rabbit.

Previous monomer alignments and CTCF profiles suggested the human repeat monomers were more similar to felid RA than RB (Bredemeyer et al., 2021). Rhesus, rat and horse also exhibited an RA-like CTCF profile, while dog, pig and both bat species were more similar to the felid RB monomers (Table C4.13). While we only observed representation of both arrays in felids, it is possible that assemblies classified as incomplete may harbor alternate arrays collapsed or unassembled within the unassigned sequence.

Repeat monomer length varied across lineages with the shortest residing in rat (2872 bp) and longest in pig (6660 bp). Standard deviation between repeat units of the same species was consistently small, varying by an average of 31 bp (0.23%-1.36% of total monomer length) when discounting the tiger, mouse and horse outliers. While increased variation in mouse and tiger monomer length was previously attributed to divergent microsatellite alleles between repeat units, the horse differed in that it exhibited unusually high sequence divergence across the entire length of the monomer (Figure B4.17). Within species *DXZ4* monomer length is highly conserved, likely attributed to either selection pressure or fidelity of the macrosatellite expansion process.

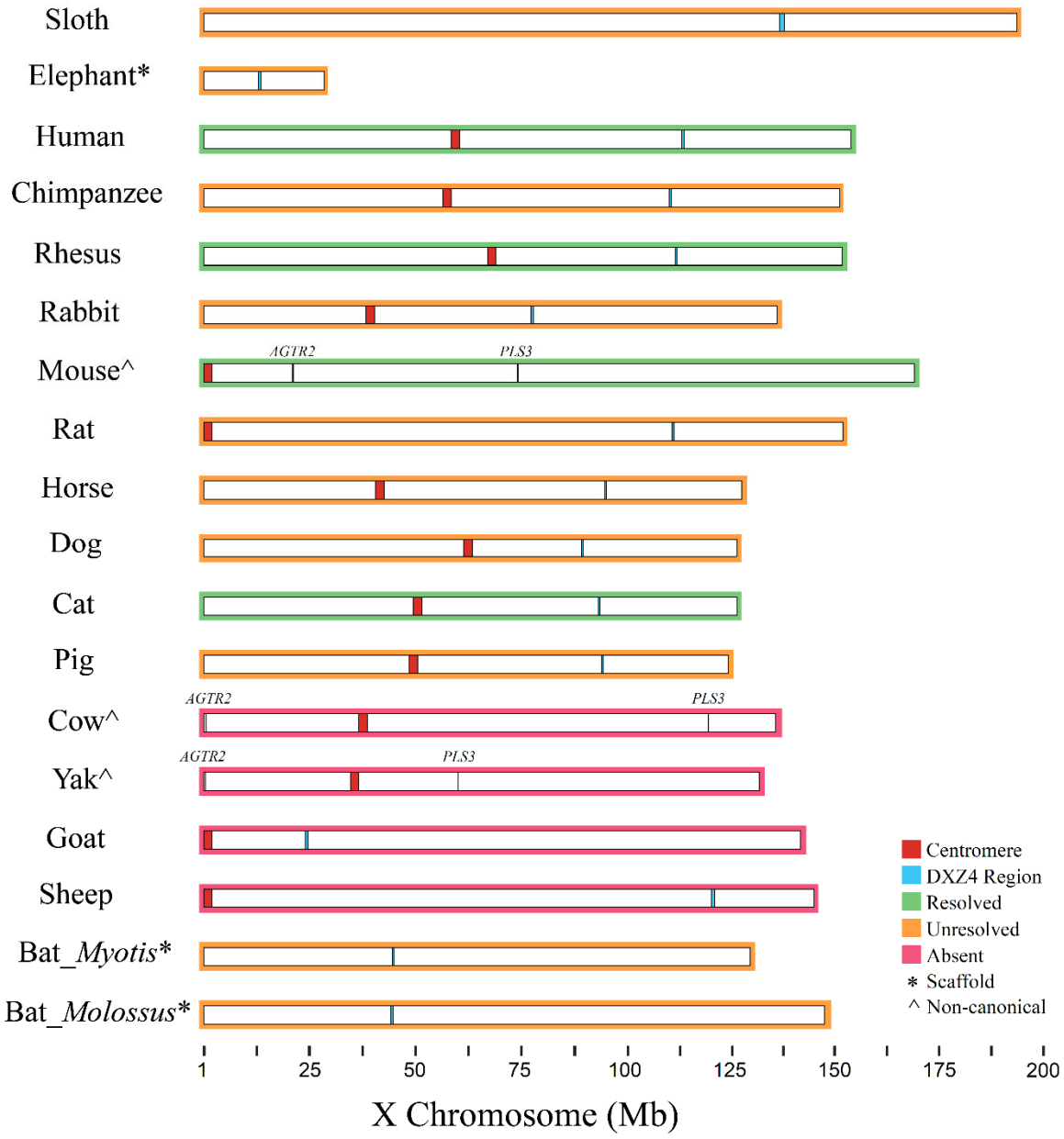


Figure 18. X Chromosome diagrams showing location and state of the *DXZ4* locus.

X Chromosomes are sorted based on mammalian phylogeny. Outlining colors represent state of *DXZ4* region (Resolved, Unresolved or Absent). Blue lines indicate canonical *DXZ4* region between *PLS3* and *AGTR2*. Species where the canonical region is disrupted are indicated by a caret and have *PLS3* and *AGTR2* labeled. Centromeres are indicated by a red rectangle. Species not assembled to chromosome level are indicated with a star.

A maximum likelihood phylogeny based on 10% masked alignment of all annotated repeat monomers revealed species but not superordinal level grouping (Figure 19, Figure B4.18). This break in organization results from insertion of the rodent lineage within Laurasiatheria as opposed to Euarchontoglires. Mean within-group *P*-distance revealed very low intraspecific divergence in all but the horse (Table C4.14). Unlike RA-1 in all felids, the most proximal repeat monomer of other species was not significantly diverged from other monomers within the array. Mean between-group *P*-distance revealed unexpected relationships between multiple repeat monomers and their felid equivalent (based on CTCF profiles) (Table C4.14). When comparing relationships derived from sequence identity vs those extrapolated from CTCF binding profiles, we notice the two features do not always agree. For example, dog, pig and *M.molossus* were more similar to the felid RA monomer, despite containing a felid RB-like CTCF motif profile. The horse was opposite, with increased similarity to felid RB and an RA-like CTCF motif profile. We hypothesize that discrepancy between CTCF motif profiles and overall sequence identity is heavily influenced by the primary biological function of *DXZ4* which may vary between species. A maximum likelihood phylogeny of CTCF binding motifs from different species revealed that relationships differed from the phylogeny built from whole repeat monomers. We observed CTCF motifs grouped primarily by general position within repeat monomers as opposed to by species, supporting our previous hypothesis suggesting independent evolution of repeat monomers and CTCF binding motifs based on *DXZ4* functionality (Figure B4.19).

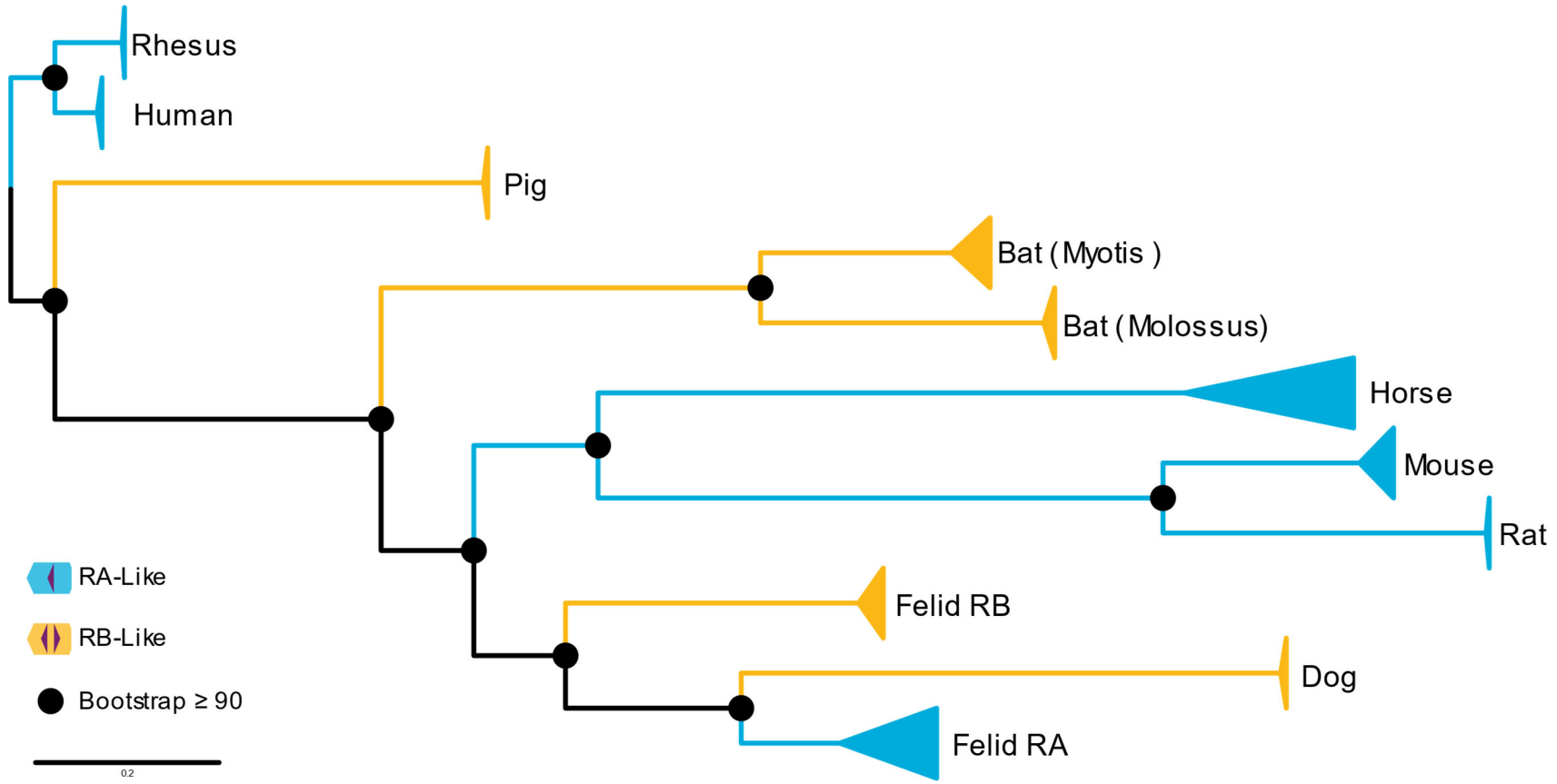


Figure 19. Phylogenetic analysis of DXZ4 monomers from divergent mammal species.

Maximum likelihood phylogeny generated from an alignment of mammalian DXZ4 monomers and rooted by primates (Rhesus and Human). Major groupings in the tree were collapsed to aid visualization. Branches representing monomers containing CTCF profiles of felid RA are colored blue. Branches representing felid RB CTCF profiles are colored orange. Black dots represent a bootstrap support value ≥90.

4.3 Discussion

The application of the trio-binning sequencing and assembly approach (Koren et al., 2018) to two felid F1 hybrids, Safari cat (*Felis catus* x *Oncifelis geoffroyi*) and Liger (*Panthera tigris* x *Panthera leo*), increased the number of highly continuous, single haplotype assemblies from the family Felidae to six, representing four out of eight extant felid lineages (Johnson et al., 2006). Together, these assemblies represent a powerful resource for future investigations into genes within complex and highly repetitive regions that are normally absent from diploid assemblies, including large immune gene clusters (e.g., MHC, IGH, TCRs), highly duplicated gene families, ampliconic sequence, and centromeres. Furthermore, many satellite elements of different sizes were resolved, including the macrosatellite *DXZ4* that has been implicated in complex developmental and disease processes, reproductive incompatibilities and speciation (Giacalone et al., 1992; Dumbovic et al., 2017; Bredemeyer et al., Submitted).

In humans, *DXZ4* is composed of between 12 and 120 repeat monomers, each containing a CTCF binding site (Tremblay et al., 2011). In all felid species *DXZ4* is composed of two distinct repeat monomers rather than one (Bredemeyer et al., Submitted). Like human, each monomer contains CTCF binding sites but differ in their number and orientation, likely important for CTCF binding affinity and loop extrusion directionality (Rao et al., 2017; Bonora et al., 2018). While the number of monomers composing each array differed between felids, the length of monomers, and CTCF profile did not, suggesting these features are important to *DXZ4* functionality.

To determine whether the felid macrosatellite structure is the exception or the rule in mammals, we expanded our investigation of *DXZ4* to include additional mammalian genome

assemblies representing species from divergent orders. *DXZ4* structure was highly divergent both in monomer sequence identity and large-scale organization between mammals. Despite divergence in repeat monomers we observed conservation of CTCF binding motif patterns across almost all species, with some resembling the felid RA, RB, or both. Unfortunately, most of the more informative basal or divergent mammalian assemblies were the least resolved, making accurate interpretation of the ancestral *DXZ4* state impossible at this time. In all four bovid species, we failed to find evidence of a tandem duplication and instead observed only a small, non-repetitive cluster of CTCF binding motifs identified as *DXZ4*-derived. While the cause of *DXZ4* macrosatellite degradation within this lineage is unclear, clustering of CTCF binding motifs further support their importance in a conserved *DXZ4* function.

To date, *DXZ4* is primarily affiliated with female XCI, where it is required for establishment and maintenance of the inactive X (Xi) bipartite structure (Deng et al., 2015; Giorgetti et al., 2016). Through binding with the nucleolar lamina, *DXZ4* forms physical interactions with *ICCE* and *FIRRE* macrosatellites harboring clusters of CTCF binding motifs (Darrow et al., 2016; Jégu et al., 2017; Bansal et al., 2020). On the Xi, cohesin molecules that normally govern topological association through loop-extrusion are depleted, requiring clustered CTCF sites for maintenance of higher-order chromatin organization (Rao et al., 2014; Minajigi et al., 2015; Kentepozidou et al., 2020). However, multiple studies have shown that ablation of *DXZ4* had no significant impact on the silenced state of the Xi despite loss of the exclusive bipartite structure (Bonora et al., 2018; Froberg et al., 2018). This startling observation suggests that super-domains formed by *DXZ4* on the Xi are not necessary for achieving or maintaining the inactive X state and therefore could be related to

some other cellular requirement or represent vestigial features of a former, but now lost, ancestral function. This idea is supported by the observed structural divergence between *DXZ4* of divergent mammals utilized in this study. If *DXZ4* were necessary for success of a biological process as highly conserved and developmentally critical as XCI we would expect a greater level of conservation between repeat units or array structure, both of which vary widely across mammals. These observations together with observations of highly conserved CTCF binding motifs, support chromatin structural organization as the present role for *DXZ4* in mammals.

Our previous investigations of *DXZ4* in male domestic cat meiotic germ cells revealed transcriptional activity during a second, male-specific instance of X chromosome silencing termed meiotic sex chromosome inactivation (MSCI). MSCI occurs during pachynema of male meiosis I and results in an X chromosome state reminiscent of the Xi in female somatic cells. In the testes of sterile male hybrids this process is disrupted as a result of *DXZ4* misregulation, leading to X chromosome-wide gene upregulation and subsequent meiotic arrest. These observations suggest that the presence and proper regulation of *DXZ4* is likely critical for a process outside of female XCI, suggesting the ancestral role of the locus might actually relate to male meiotic stability as opposed to dosage compensation.

Unfortunately, *DXZ4* was variably resolved and not uniformly present in most of the assemblies in our study *DXZ4*, limiting our insights into structural evolution of the locus. Additionally, *DXZ4* function has only been described in human, rhesus, mouse and cat, making correlation between divergence that was observed, and biological relevance across mammalian lineages difficult and further confounding our understanding of the locus. Future investigation of the relationship between Xi structural conformation and *DXZ4* organization

across additional species, especially those apparently lacking *DXZ4*, should lend additional insights into the function of *DXZ4* in, and exclusive of, XCI.

Recent advances in sequencing technology have made generation of high-quality genomes achievable for a larger proportion of the genomics community. Pacific Biosciences high-fidelity (PacBio HiFi) reads are highly accurate long reads (>20kb) that negate the need for Illumina polishing, enabling accurate assembly of duplications composed of monomers with high sequence identity like segmental duplications and macrosatellites (Nurk et al., 2020). Additionally, concentrated efforts have been and are being made by multiple groups to generate gapless or near-gapless assemblies for a wider array of species (Rhie et al., 2021; Miga et al., 2020; Logsdon et al., 2020). As the number of resolved macrosatellites increases so will our understanding of *DXZ4* and its ancestral function within placental mammals. Currently, conservation of CTCF motifs, expendability in XCI and recent association with male meiosis suggest a more varied role of the macrosatellite than previously appreciated.

4.4 Methods

4.4.1 Biological Materials

The parent-offspring trio of the Safari cat was composed of a random-bred domestic cat (*Felis silvestris catus*) dam, a Geoffroy's cat (*Oncifelis geoffroyi*) sire, and a female F1 offspring (GXD-2). Fibroblast cell lines were established for the F1 female at the National Cancer Institute animal colony as part of the generation of an interspecies mapping panel (Menotti-Raymond et al., 1999; 2003; Davis et al. 2015). Tissue from the parents of the F1 Safari cat were unavailable but cell lines were karyotyped to confirm species identity and F1 status. The parent-offspring trio of the Liger was composed of a Tiger (*Panthera tigris*) dam,

a Lion (*Panthera leo*) sire, and a male F1 offspring (LxT-3). A fibroblast cell line was established for the F1 male liger.

4.4.2 Nucleic Acid Library Preparation and Sequencing

Long-read library preparation and sequencing

For the F1 Safari cat and Liger, high molecular weight genomic DNA was extracted using a modified salting out protocol (Miller et al., 1988) followed by length quantification using the Pippin Pulse pulse-field gel system (Sage Science). DNA was quantified via Qubit fluorometric quantification (Thermo Fisher Scientific). PacBio SMRT libraries were size selected (~20-kb) on the Sage Blue Pippin and sequenced on the Sequel II instrument to yield approximately 158x and 153x coverage for the Safari and Liger F1, respectively.

Short-read library preparation and sequencing

For the F1 Liger, standard dual indexed Illumina fragment libraries (~300-bp average insert size) were prepared for the parent samples (tiger and lion) using the NEBNext Ultra II FS DNA Library Prep Kit (New England Biolabs Inc.). Libraries were assayed with fluorometric quantification using the Qubit (Thermo Fisher Scientific) and electrophoresis using the TapeStation (Agilent). Samples were sequenced to ~26-28x genome-wide depth of coverage with 2×150-bp reads using the NovaSeq 6000 Sequencing System (Illumina).

Hi-C library preparation and sequencing

F1 Safari cat fibroblasts from GXD-2 were fixed as a monolayer using 1% formaldehyde for 10 minutes, divided into ~$4.2x10^6$ cell aliquots, snap frozen in liquid nitrogen and stored at -80°C as described (Ramani et al., 2016). Cells were lysed, resuspended in 200ul of 0.5x DNase I digestion buffer, and chromatin digested with 1.5 units

of DNase I for 4 minutes. Downstream library preparation was performed as described (Ramani et al., 2016) and sequenced on the Illumina NovaSeq 6000 to ~78x coverage.

4.4.3 Genome Assembly and Annotation

Haplotype Binning

Software and versions used for each assembly step were the same ones used to generate the previous single-haplotype assemblies (Bredemeyer et al., 2020). All Illumina data was processed with *FastQC v0.11.8* (Andrews, 2010) followed by adapter trimming using *Trim Galore! v0.6.4*. Illumina sequences were unavailable for the actual parents of the F1 Safari cat and substituted with Illumina data from the parental species, including the F1 Bengal domestic cat parent (Fca-508) and previously generated Geoffroy's cat Illumina data (Oge-3: SRR6071645) (Li et al. 2019). Illumina sequences from the Liger and Safari cat parent species were used to phase the raw F1 Liger and Safari cat PacBio long reads into haplotype bins using the trio binning feature of *Canu v1.8* (*TrioCanu*) (Koren et al., 2017; Koren et al., 2018).

De novo Assembly

Haplotyped long reads for each species were assembled using *NextDenovo v2.2-beta.0* (github:Nextomics/Nextdenovo) with the configuration file (.cfg) altered for inputs: *minimap2_options_raw = -x ava-pb*, *minimap2_options_cns = -x ava-pb*. The *seed_cutoff=* option was adjusted to 32k for all assemblies.

Contig Polishing and QC

NextPolish v1.3.0 (Hu et al., 2019) and *NextDenovo* corrected long reads were used to polish the raw contigs. Notable changes to the *NextPolish* configuration file included: *genome_size=auto*, and *task=best*, which instructs the program to perform two iterations of

polishing using the corrected long reads. The *sgs* option was removed as polishing with the parental diploid short reads could lead to conversion of consensus sequence to reflect the alternate haplotypes not present in the F1. The *lgs* options within the configuration file was left at default settings except for modification for PacBio long reads by adjusting *minimap2_options= -x map-pb*. Basic assembly stats were generated using *QUAST v5.0.2* (Mikheenko et al., 2018) with the *--fast* run option selected. To assess genome completeness, *BUSCO v4.0.6* (Simão, et al., 2015) was run using the *-m* genome setting with *-l mammalia_odb10* database selected (9,226 single copy genes). Visual assessment of the haploid assemblies was performed through alignment to the single-haplotype domestic cat assembly Fcat_Pben_1.0_maternal_alt (Fca-508: GCA_016509815.1) (Bredemeyer et al. 2021) using *nucmer* (*mummer3.23* package; Marçais et al., 2018) with default settings. The resulting delta file was used to generate a dot plot for genome comparison using *Dot: interactive dot plot viewer for genome-genome alignments* (DNAnexus).

Lion Y Chromosome Contig Identification and Isolation

To identify Y chromosome contigs within the lion assembly we used two parallel approaches. The first was based on mapping female lion Illumina reads to the lion haplotype contigs and the second was based on BLAST identification of lion contigs using previously described domestic cat Y sequences. This approach relies on the expectation that female reads will lack Y chromosome sequence and thus allows for selection of contigs based on zero or limited read coverage across their length. Illumina sequence reads from a female lion (Brooke) were previously generated by Armstrong et al. (2020), accession number SRR10009886, and aligned to the contigs using *bwa mem v0.7.17* (Li et al, 2009) with default settings. The alignment output was piped into *Samtools v1.9* (Li et al, 2009), where

it was converted into bam format, sorted, and indexed. To annotate coverage across the assembly we used the *genomecov* tool of the *BEDTools suite v2.29.2* (Quinlan et al, 2010) with the *-d* (for per base coverage) and *-bga* (for regional coverage as bedgraph) options selected. The results were output in bedgraph format and all contig nucleotide positions identified as having a female read coverage threshold of 15x, 10x, 5x or 0x, and extracted into separate lists. For each coverage list *R v.3.5.1* (R Core Team) was used to calculate the percent of positions in each contig at, or below, the coverage threshold. This was done by summing the total number of positions annotated from *BEDTools* for each contig, and dividing by the contigs total length. Next, to begin screening for Y contigs, we extracted contigs with 70-100% of their nucleotide positions within the threshold coverage. The first step in determining this cutoff was to identify Y contigs using known domestic cat Y chromosome sequences. To do this we used NCBI's basic local alignment search tool *(BLAST) v2.9.0* (Altschul et al, 1990) command line application with options *-culling_limit 1, -evalue 1e-25, and -perc_identity 85* specified to align domestic cat mRNA sequences (Murphy et al, 2006; Pearks Wilkerson et al, 2008; Li et al, 2013) and ampliconic Y chromosome BAC clones (Brashear et al, 2018) to the lion assembly custom BLAST database. Because lion and domestic cat diverged ~11mya (Li et al, 2016) the *dc-megablast* algorithm was used to generate alignments. *GNU parallel* (Tange, 2011) was used to BLAST multiple queries simultaneously. Finally, the percent positional contig coverage identified by single copy mRNA sequences was averaged and +/-25% added to determine the percent position coverage cutoff. This threshold was further justified by plotting a histogram for the data where we observed a bimodal distribution for the number of contigs either lacking or possessing female read coverage across most of their lengths**.** This result suggests the cutoff

was sufficient for capturing Y sequence contigs while avoiding autosome/X-linked contigs. Second, *megablast* was used to align contigs with 70% of their nucleotide identity possessing 15x coverage or less to the felCat9.0 assembly (Buckley et al. 2020). For this step, only the top blast hit was output using command line options *-num_alignments 1*, *-max_hsps 1*. To avoid false hits to repetitive elements both the felCat9.0 custom database and lion contigs were repeat-masked using *RepeatMasker v4.0.9* (Smit et al, 2013-2015) with default settings and option *-species felidae* selected. Using these results, we manually selected additional contigs presumed to be Y linked based on a combination of alignment percent identity to fca-9.0 autosomal sequences, percent nucleotide positions covered by female reads, and total sequence length. All contigs identified in this way were merged into a single list and removed from the assembly using *seqTK subseq v1.3* (github:seqtk).

Scaffolding

Polished contigs from the domestic and Geoffroy's cat were scaffolded using Hi-C data generated from the F1 Safari cat hybrid. Prior to scaffolding, Safari cat Hi-C reads were binned into parental haplotypes through alignment of the offspring reads to both polished parental assemblies using *bwa mem v0.7.17* (Li and Durbin, 2009) and the *classify_by_alignment* (https://github.com/esrice/trio_binning/ v0.2.0) program as described in Rice et al. (2020). Haplotyped reads were mapped to polished contigs using the pipeline and scripts described in Rice et al. (2020) (https://github.com/esrice/slurm-hic/) using *SALSA v2.2* (Ghurye et al., 2017; Ghurye et al., 2018) with parameters *-e none -m yes.* Prior to scaffolding, all Y associated contigs were removed to prevent incorporation of repetitive Y chromosome contigs into highly similar, but paralogous autosomal regions (Brashear et al., 2018) during scaffolding. Previously published Hi-C data for tiger (SRR8616865) and

lion (SRR10075807/SRR10075808) were available and used to scaffold their respective assemblies with *SALSA* parameters *-e GATC -m yes* (DNA Zoo Dudchenko et al. 2017; Armstrong et al. 2020). The resulting scaffolds were inspected using *QUAST*, *nucmer*, and Hi-C contact maps. Following *SALSA*, *RagTag v1.0.1* (Alonge et al., 2019) was used to align scaffolds to their respective position in the chromosome-length single-haplotype domestic cat assembly. Selected *RagTag* parameters included *–remove-small, -f 10000* and *-j unplaced.txt*, a text file of scaffolds for *RagTag* to ignore based on their small size and identification as repetitive sequence in the *nucmer* alignments. *RagTag* scaffolds were manually inspected with Hi-C maps generated using *Juicer v1.5.7* (Durand et al., 2016a) with option *-s none* selected for compatibility with DNase Hi-C libraries. Maps were visualized using *Juicebox v1.11.08* (Durand et al., 2016b) and *Juicebox Assembly Tools* with scripts from *3d-dna v.180922* (Dudchenko et al., 2017).

4.4.4 Genome Annotation

To identify conserved genes across each assembly we used *Liftoff v1.4.2* (Shumate and Salzberg, 2020) to perform an annotation liftover between the current felCat9 reference assembly (Buckley et al. 2020) and all four *de novo* felid assemblies. Default parameters were used for all arguments except for calling *-copies* with *-sc 0.95* to identify extra copies of genes not previously annotated in felCat9.

4.4.5 **DXZ4** *Annotation and Alignment in Felids*

Identification and isolation of *DXZ4* repeat units was performed manually using GC content traces, CTCF motif annotations, and self-self dotplots for the region using *Geneious Prime v2021.0.3* and *FlexiDot v1.06* (Seibt et al., 2018). CTCF motifs were annotated using the *Geneious Annotate & Predict tool* with a sequence motif of

GAGTTTCGCTTGATGGCAGTGTTGCACCACGAAT, based on the Horakova (2012a) conserved CTCF motif logo, with the most prevalent nucleotide representative of each position. A max mismatch of 13 was selected to allow for interspecific ambiguity within the motif. CTCF sites annotated using this method corresponded to the approximate location within human *DXZ4* repeat units originally described by Chadwick (2008). Once annotated and extracted, independent repeats were aligned using the *Mafft Multiple Aligner v1.4.0* and maximum likelihood trees generated using the Geneious *RAxML v8.2.11* (Stamatakis, 2014) plug-in with nucleotide model: *GTR +I+G*, Algorithm: *Rapid hill-climbing* and Replicates: *500* selected. Maximum likelihood bootstraps were generated using *IQ-TREE web server* (Trifinopoulos et al., 2016) with *ultrafast* selected and set for 1000 bootstrap alignments and maximum iterations. Trees were pruned using *Mesquite v3.61* (Maddison & Maddison, 2019) and visualized using *FigTree v1.4.4* (Rambaut, 2018). Mean within- and between-group distances for masked (10% gaps masked) *DXZ4* repeat unit alignments were calculated using *Mega-X v10.0.5* (Kumar et al., 2018).

4.4.6 Investigation of DXZ4 *in Placental Mammals*

In an effort to determine the ancestral state and further describe the evolution of *DXZ4* we sampled assemblies from all mammalian superordinal clades (Murphy et al. 2021). Assemblies were downloaded from NCBI, with chromosome length versions derived from long-reads and male individuals preferentially selected. Based on these criteria we were able to obtain an additional 15 assemblies, increasing the total number of represented species to 22 (Table C4.15). Of these the elephant, horse, rabbit and myotis bat were female. The three assemblies not assembled into full chromosomes were the elephant, and both bat assemblies. In assemblies alternate to the species reference and lacking annotation, a liftover was

performed using *Liftoff* as described in the genome assembly section. This allowed identification of *PLS3* and *AGTR2* genes which have previously acted as a proxy for *DXZ4* (Horakova 2012a). CTCF binding motifs were annotated as described for Felid assemblies. Centromere positions were identified using a combination of NCBI annotations, interspecific alignments and the Atlas of Mammalian Chromosomes, 2nd Edition (Table C4.16). Dotplots of each region was generated using *FlexiDot*. We determined presence/absence of *DXZ4* based on presence of repeat structure, CTCF binding motifs and location relative to *PLS3* and *AGTR2*. For assemblies lacking repeat structure human, cat and pig repeat monomers were queried against the X-chromosome using the discontiguous-megablast BLAST algorithm. *DXZ4* repeat unit alignments within and between species was performed using *MAFFT* with trees generated using *RAxML*. Similar to the felid analysis, bootstraps were calculated using *IQ-Tree*. Mean within and between-group *P*-distances were calculated using *Mega-X*.

CHAPTER V CONCLUSIONS AND FUTURE WORK

5.1 *DXZ4* in Male Meiosis and Hybrid Sterility

In chapter III we presented genetic evidence for an association between hybrid male sterility in the Chausie hybrid cat breed and structural and functional variation at the *DXZ4* macrosatellite. These observations indicated, for the first time, a function for the locus outside of its traditional role on the inactive X (Xi) of female somatic cells. We hypothesized that *DXZ4* was critical to male meiosis through participation in meiotic sex chromosome inactivation (MSCI), a process specific to pachytene spermatocytes, and whose failure results in male sterility phenotypes observed in other mammals. One limitation of our study was that the functional genetic profiles in hybrid cats were generated from seminiferous tubules of whole testes, which may be confounded by tissue composition bias. Tissue composition bias stems from the mosaic nature of testes tissue, which is composed of multiple classes of somatic support cells and germ cells at various stages of meiosis and spermatogenesis. As primordial male germ cells progress through meiosis they exhibit dramatic changes in chromatin organization and structural morphology accompanied by cell-stage-specific epigenetic, transcriptional, and chromatin conformation profiles (Djureinovic et al., 2014). As a result, testes express far more tissue specific genes than any other organ, further complicating interpretation of gene expression and activity (Uhlén et al., 2015). To mitigate this problem and allow unbiased profiling of *DXZ4* on the male inactivated X, we isolated (via cell sorting) populations of germ cells exhibiting MSCI and post-meiotic sex chromatin (PMSC) (pachytene spermatocytes and round spermatids, respectively) from fertile domestic cat testes. To determine whether *DXZ4* function is explicitly linked to

initiation and maintenance of meiotic X-chromosome silencing we also initially intended to include pre-meiotic spermatogonia cells in the analysis. This proved unachievable with the StaPut cell sorting method we employed because the number of spermatogonial cells in a testis is insufficient for traditional library preparation methods. To circumvent this constraint and further increase resolution across meiotic stages, future experiments should focus on utilizing FACS or single-cell technologies, both of which are capable of isolating distinct germ cell populations, including spermatogonia (Nagano et al., 2015; Larson et al., 2016a, 2016b; Jung et al., 2019). Application of these techniques to testes of fertile and sterile Chausie hybrids would enable more confident associations between stage specific differences in expression, methylation and chromatin organization of the X Chromosome and meiotic disruption and hybrid sterility.

5.2 Investigating the Biological Function of *DXZ4*

Following resolution of the *DXZ4* locus, we primarily focused on its functionality in male germ cells and how it could be contributing to the observed hybrid male sterility phenotype in Chausies. However, we did not focus on its activity in female somatic cells. In humans, *DXZ4* transcribes lncRNAs and smallRNAs of unknown function and is bound by CTCF proteins exclusively on the Xi, prompting formation of its unique bipartite structure (Tremblay et al., 2011; Pohlers et al., 2014; Rao et al., 2014). This same structure is also formed on the Xi of domestic cat, implicating *DXZ4* exists in a euchromatic state akin to the human (Brashear et al., 2021). To validate whether this is true would require female RNA-Seq and methylation data to determine whether the locus was transcriptionally active in domestic cat. One complicating issue with investigating the role of *DXZ4* in females of most species, including felids, is that alleles exist in simultaneously active and inactive states,

dependent on the condition of their native chromosome, making haplotype specific analysis necessary for unbiased investigation. In human and mouse, this issue is resolved by read haplotype phasing, with mouse making use of hybrid cell lines exhibiting a high rate of heterozygosity and skewing of XCI towards a single haplotype (Rao et al., 2014; Darrow et al., 2016). This same concept could be applied to a number of felid F1 hybrids, whose parent species are divergent enough for efficient phasing. In chapter II we generated single haploid assemblies for an F1 Bengal involving the phasing of long reads using k-mer analysis. Additionally, we were able to efficiently phase Illumina Hi-C libraries from the same F1 Bengal to generate species specific Hi-C maps. In this hybrid we observed what we suspect is heavy skewing of XCI in the domestic cat X homolog. This suspicion was based on observation of a bipartite structure and TAD attenuation on the domestic cat X chromosome only, with the Asian leopard cat X exhibiting a normal organization more consistent with an active X or autosomes.

In human and mouse, the Xi also exhibits enrichment of long-range chromatin interactions between *DXZ4* and the additional CTCF enriched regions of other X-linked macrosatellite loci, *ICCE* and *FIRRE* (Darrow et al., 2016). Unlike the bipartite structure, these interactions have not been observed in felids, likely due to insufficient Hi-C map resolution. This deficiency could be addressed by increasing the number of captured contacts either through additional sequencing or generation of more complex libraries, though it is possible that felids simply lack these interactions, as they appear to be dispensable to bipartite structure formation in both human and mouse (Froberg et al., 2018; Barutcu et al., 2018). Knockouts of the *DXZ4* region in human and mouse also revealed that despite its connection to XCI in females, the presence or absence of *DXZ4* has little to no impact on

establishment or maintenance of the silenced, heterochromatic Xi state (Bonora et al., 2018, Froberg et al., 2018). However, loss of *DXZ4* did correspond to loss of the bipartite structure. Together these observations offer strong support for structural organization as the primary role for *DXZ4* but in biological processes independent of XCI.

Chapter III, describes associations between *DXZ4* and hybrid male sterility, implicating the locus in male meiosis, specifically the process of MSCI. Transcriptional activity and hypomethylation across the locus in pachytene spermatocytes supported this association and suggest presence and proper regulation of *DXZ4* is critical to successful male gametogenesis. While this implicates *DXZ4* in a role outside of XCI, little data exists to support its function during male X inactivation. In domestic cat pachytene spermatocytes and round spermatids, we did not observe any large-scale structural changes resembling the Xi bipartite structure, but did observe TAD attenuation and A/B compartmentalization changes across the X chromosomes, indicative of large-scale chromatin reorganization. We did not observe any of the long-range chromatin interactions reported on the female Xi, but likely lacked the resolution required to detect them. Expendability of both the bipartite structure and long-range chromatin interactions for proper XCI and their absence from the male inactivated X chromosome suggest that these characteristics are either temporally regulated, specific to an alternative biological process, or are simply a consequence of other features governing the female Xi. In the future, mapping of transcriptional, epigenetic and structural data to assemblies with resolved *DXZ4* regions will lead to a greater understanding of the mechanistic underpinnings and functional significance of *DXZ4* in mammalian biology.

5.3 Implications of *DXZ4* Structure in Felids

In chapters III and IV, we demonstrated how single-haplotype genome assembly was crucial to resolution of the *DXZ4* macrosatellite. We also revealed a novel, complex structure not previously observed in mouse and human, that was conserved across all five felid species for which we generated single haplotype genome assemblies. We revealed that *DXZ4* exists as a compound repeat of two divergent repeat arrays, Repeat A and Repeat B (RA/RB). The orthologous human and mouse *DXZ4/Dxz4* tandem repeats are composed of a single duplicated monomer exhibiting similarities with only felid RA monomer. The felid RA repeat was more similar to the human monomer, both in terms of CTCF binding motifs as well as at the sequence level and that RB was completely absent. This raised the question that RA and RB might have distinct structural and functional roles. Currently, functional analysis of felid RA and RB is limited to male meiosis in domestic and Jungle cat. Future studies utilizing RNA-Seq data from additional cat species, particularly those represented here with single-haplotype assemblies, as well as RNA-Seq data from additional tissue types, would lend insight into conservation of transcription across the felid lineage in germ cells and might uncover distinct functionalities of the RA and RB repeat arrays outside of male meiosis.

5.4 Evolution of *DXZ4* in Mammals

Differences between *DXZ4* structure in human and felids raised several questions regarding the ancestral state of the macrosatellite in placental mammals. In chapter IV we investigated the conservation of these two repeat arrays across a multitude of placental mammal species for which long-read genomes were available. We observed several clade specific differences in *DXZ4* structure, as well as absence of the macrosatellite region in

bovids. In cow and yak, the region between the canonical boundary genes was broken and the ancestral *DXZ4* region translocated to the distal end of the p-arm in both species. How this X chromosome rearrangement might correspond to *DXZ4* loss is an interesting topic for future studies that will require additional data to unravel. In human, macaque, mouse and rat we observed only RA-like *DXZ4* motifs suggesting potential loss of RB-like monomers in the Euarchontoglires clade. Conversely, in Laurasiatherians we only observed *DXZ4* monomers exhibiting RB-like CTCF binding motifs. Unfortunately, every species exhibiting only RB-like monomers was unresolved across the *DXZ4* region, and potentially could harbor RA-like monomers in unassembled or unplaced genomic sequence. Interestingly, we also observed RB-like CTCF motifs in monomers of the *DXZ4*-derived *ICCE* macrosatellite, which we found present as a simple tandem duplication in all primates, cat species, dog and pig, and lacking from bovids and rodents (Westervelt & Chadwick et al., 2018). Lack of structural change in *ICCE* relative to *DXZ4* in these species, as well as its absence from others suggests a conserved role for the macrosatellite in mammals and merits future comparative analysis.

5.5 The Future of Felid Single-Haplotype Assemblies

In our analysis of *DXZ4* across mammals we recognized that assemblies for many species still lack the continuity required for comparisons of complex, highly repetitive regions. Fortunately, recent advances in sequencing technology and assembly methods, as well as genome consortia dedicated to generating gapless or near-gapless assemblies, are contributing to an ever-increasing number of high-quality genomes (Koren et al., 2018; Logsdon et al., 2020; Miga et al., 2020; Nurk et al., 2020; Rhie et al., 2021). In chapter II we illustrate the genome quality achievable with application of these modern techniques and

long read sequencing through application of the trio binning method to an F1 Bengal hybrid. Assembly contiguity and quality for the resulting domestic cat and Asian leopard cat genomes were equivalent (or superior) to those of the human genome assembly and showed considerable improvement over the previous domestic cat reference genome published only a few years ago (2017). Subsequent single-haplotype assemblies of a second domestic cat, Geoffroy's cat, lion and tiger using improved long-read data yielded even higher continuity and genome completeness, positioning the domestic cat assembly to act as the new felid reference. As a group, these assemblies allow unprecedented insight and comparison between historically complex regions like macrosatellites, ampliconic sequences and centromeres, which until recently were missing from even the human assembly. With these genome assemblies, four out of the eight felid lineages are already represented, spanning the deepest split of the felid phylogeny (Domestic Cat and Panthera Lineages); and because additional hybrids exist within and between previously unsampled clades, expansion of phylogenetic breadth to capture the full range of felid diversity remains a target for future endeavors. As genome assembly quality has accelerated over the past decade, so too has our understanding of highly complex regions that orchestrate many aspects of biology and evolution.

REFERENCES

Abascal F, Corvelo A, Cruz F, Villanueva-Cañas JL, Vlasova A, Marcet- Houben M, Martínez-Cruz B, Cheng JY, Prieto P, Quesada V, et al. 2016. Extreme genomic erosion after recurrent demographic bottlenecks in the highly endangered Iberian lynx. Genome Biol. 17:251.

Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, Mason CE. 2012. MethylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles.

Alavattam KG, Maezawa S, Sakashita A, Khoury H, Barski A, Kaplan N, Namekawa SH. 2019. Attenuated chromatin compartmentalization in meiosis and its maturation in sperm development. *Nat. Struct. Mol. Biol.* [Internet] 26:175–184.

Allen R, Ryan H, Davis BW, King C, Frantz L, Irving-Pease E, Barnett R, Linderholm A, Loog L, Haile J, et al. 2020. A mitochondrial genetic divergence proxy predicts the reproductive compatibility of mammalian hybrids: A mitochondrial proxy of hybridibility. *Proc. R. Soc. B Biol. Sci.* 287.

Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, Lippman ZB, Schatz MC. 2019. RaGOO: Fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.* 20:1–17.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–410.

Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, Robinson MD. 2013. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat. Protoc.* 8:1765–1786.

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Antunes A, Pontius J, Ramos MJ, O'Brien SJ, Johnson WE. 2007. Mitochondrial introgressions into the nuclear genome of the domestic cat. In: Journal of Heredity. Vol. 98. p. 414–420.

Armstrong SJ, Hultén MA, Keohane AM, Turner BM. 1997. Different strategies of X-inactivation in germinal and somatic cells: Histone H4 underacetylation does not mark the inactive X chromosome in the mouse male germline. *Exp. Cell Res.* [Internet] 230:399–402.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene Ontology: tool for the unification of biology. *Nat. Genet.* [Internet] 25:25–29.

Axtell MJ. 2013. ShortStack: Comprehensive annotation and quantification of small RNA genes. *Rna* 19:740–751.

Babraham Bioinformatics. 2020. Trim_Galore, http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

Balcova M, Faltusova B, Gergelits V, Bhattacharyya T, Mihola O, Trachtulec Z, Knopf C, Fotopulosova V, Chvatalova I, Gregorova S, et al. 2016. Hybrid Sterility Locus on Chromosome X Controls Meiotic Recombination Rate in Mouse. *PLoS Genet.* [Internet] 12:1–16.

Bansal P, Kondaveeti Y, Pinter SF. 2020. Forged by *DXZ4*, *FIRRE*, and *ICCE*: How Tandem Repeats Shape the Active and Inactive X Chromosome. *Front. Cell Dev. Biol.* 7:1–10.

Barr ML and Bertram EG. 1949. A morphological distinction between neurones of the male and female, and the behaviour of the nucleolar satellite during accelerated nucleoprotein synthesis. Nature. 163:676–677.

Barutcu AR, Maass PG, Lewandowski JP, Weiner CL, Rinn JL. 2018. A TAD boundary is preserved upon deletion of the CTCF-rich Firre locus. *Nat. Commun.* [Internet] 9.

Bayes JJ, Malik HS. 2009. Altered heterochromatin binding by a hybrid sterility protein in *Drosophila* sibling species. *Science.* [Internet] 326:1538–1541.

Beissinger TM, Rosa GJ, Kaeppler SM, Gianola D, De Leon N. 2015. Defining window-boundaries for genomic analyses using smoothing spline techniques. *Genet. Sel. Evol.* 47.

Bhattacharyya T, Gregorova S, Mihola O, Anger M, Sebestova J, Denny P, Simecek P, Forejt J. 2013. Mechanistic basis of infertility of mouse intersubspecific hybrids. *Proc. Natl. Acad. Sci. U. S. A.* 110:468–477.

Bhattacharyya T, Reifova R, Gregorova S, Simecek P, Gergelits V, Mistrik M, Martincova I, Pialek J, Forejt J. 2014. X Chromosome Control of Meiotic Chromosome Synapsis in Mouse Inter-Subspecific Hybrids. *PLoS Genet.* 10.

Bonora G, Deng X, Fang H, Ramani V, Qiu R, Berletch JB, Filippova GN, Duan Z, Shendure J, Noble WS, et al. 2018. Orientation-dependent *Dxz4* contacts shape the 3D structure of the inactive X chromosome. *Nat. Commun.* [Internet] 9.

Bourgeois CA, Laquerriere F, Hemon D, Hubert J, Bouteille M. 1985. New data on the *in situ* position of the inactive X chromosome in the interphase nucleus of human fibroblasts. *Human* Genetics. 69:122-129.

Boyle P, Clement K, Gu H, Smith ZD, Ziller M, Fostel JL, Holmes L, Meldrim J, Kelley F, Gnirke A, et al. 2012. Gel-free multiplexed reduced representation bisulfite sequencing for large-scale DNA methylation profiling. *Genome Biol.* [Internet] 13.

Brashear WA, Bredemeyer KR, Murphy WJ. 2021. Structural and functional constraints dominated placental mammal X Chromosome evolution. Forthcoming.

Brashear WA, Raudsepp T, Murphy WJ. 2018. Evolutionary conservation of Y Chromosome ampliconic gene families despite extensive structural variation. *Genome Res.* [Internet] 28:1826–1840.

Bredemeyer KR, Harris AJ, Li G, Zhao L, Foley NM, Roelke-Parker M, O'Brien SJ, Lyons LA, Warren WC, Murphy WJ. 2021. Ultracontinuous Single Haplotype Genome Assemblies for the Domestic Cat (*Felis catus*) and Asian Leopard Cat (*Prionailurus bengalensis*). *J. Hered.* [Internet] 112:165–173.

Brockdorff N, Turner BM. 2015. Dosage compensation in mammals. *Cold Spring Harb. Perspect. Biol.* 7.

Buckley RM, Davis BW, Brashear WA, Farias FHG, Kuroki K, Graves T, Hillier LDW, Kremitzki M, Li G, Middleton RP, et al. 2020. A new domestic cat genome assembly based on long sequence reads empowers feline genomic medicine and identifies a novel gene for dwarfism. *PLoS Genet.* [Internet] 16.

Burgoyne PS. 1982. Genetic homology and crossing over in the X and Y chromosomes of mammals. *Human Genetics* 61:85-90.

Burgoyne PS, Mahadevaiah SK, Turner JMA. 2009. The consequences of asynapsis for mammalian meiosis. *Nat. Rev. Genet.* 10:207–216.

Campbell P, Good JM, Nachman MW. 2013. Meiotic sex chromosome inactivation is disrupted in sterile hybrid male house mice. *Genetics* 193:819–828.

Chadwick BP. 2008. DXZ4 chromatin adopts an opposing conformation to that of the surrounding chromosome and acquires a novel inactive X-specific role involving CTCF and antisense transcripts. *Genome Res.* 18:1259–1269.

Chadwick BP. 2009. Macrosatellite epigenetics: The two faces of *DXZ4* and *D4Z4*. *Chromosoma* 118:675–681.

Chadwick BP, Willard HF. 2003. Chromatin of the Barr body: Histone and non-histone proteins associated with or excluded from the inactive X chromosome. *Hum. Mol. Genet.* [Internet] 12:2167–2178.

Charlesworth B, Campos JL, Jackson BC. 2018. Faster-X evolution: Theory and evidence from Drosophila. *Mol. Ecol.* 27:3753–3771.

Chen PY, Cokus SJ, Pellegrini M. 2010. BS Seeker: Precise mapping for bisulfite sequencing. BMC Bioinformatics. 11:1-6.

Cho KW, et al. 1997. A proposed nomenclature for the domestic cat karyo- type. Cytogenet. Cell Genet. 79:71–78.

Coyne JA, Orr HA. 1989. Patterns of Speciation in *Drosophila*. *Evolution (N. Y).* [Internet] 51:295.

Coyne JA. 1992. Genetics and speciation. *Nature* [Internet] 355:511–515.

Coyne JA. 2018. "Two Rules of Speciation" revisited. *Mol. Ecol.* 27:3749–3752.

Craven P, Wahba G. 1978. Smoothing noisy data with spline functions. *Numer Math* 31, 377-403.

Daish TJ, Casey AE, Grutzner F. 2015. Lack of sex chromosome specific meiotic silencing in platypus reveals origin of MSCI in therian mammals. *BMC Biol.* [Internet] 13:1–13.

Darrow EM, Huntley MH, Dudchenko O, Stamenova EK, Durand NC, Sun Z, Huang SC, Sanborn AL, Machol I, Shamim M, et al. 2016. Deletion of *DXZ4* on the human inactive X chromosome alters higher-order genome architecture. *Proc. Natl. Acad. Sci. U. S. A.* [Internet] 113:E4504–E4512.

Davis BW, Seabury CM, Brashear WA, Li G, Roelke-Parker M, Murphy WJ. 2015. Mechanisms underlying mammalian hybrid sterility in two feline interspecies models. *Mol. Biol. Evol.* 32:2534–2546.

Delph LF, Demuth JP. 2016. Haldane's Rule: Genetic Bases and Their Empirical Support. *J. Hered.* [Internet]:383–391.

Deng X, Ma W, Ramani V, Hill A, Yang F, Ay F, Berletch JB, Blau CA, Shendure J, Duan Z, et al. 2015. Bipartite structure of the inactive mouse X chromosome. *Genome Biol.* 16:1–21.

Dhanoa JK, Sethi RS, Verma R, Arora JS, Mukhopadhyay CS. 2018. Long non-coding RNA: Its evolutionary relics and biological implications in mammals: A review. *J. Anim. Sci. Technol.* [Internet] 60.

Djureinovic D, Fagerberg L, Hallström B, Danielsson A, Lindskog C, Uhlén M, Pontén F. 2014. The human testis-specific proteome defined by transcriptomics and antibody-based profiling. *Mol. Hum. Reprod.* [Internet] 20:476–488.

Dobin A, Gingeras TR, Spring C, Flores R, Sampson J, Knight R, Chia N, Technologies HS. 2016. Mapping RNA-seq with STAR. *Curr Protoc Bioinforma.* [Internet] 51:586–597.

Dobzhansky, T. 1937. *Genetics and the Origin of Species*. Columbia University Press

DNAnexus. 2020. Dot. https://github.com/dnanexus/dot.

Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, et al. 2017. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science.* [Internet] 356:92–95.

Dumbovic G, Forcales S V., Perucho M. 2017. Emerging roles of macrosatellite repeats in genome organization and disease development. *Epigenetics* 12:515–526.

Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. 2016. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst.* [Internet] 3:99–101.

Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL. 2016. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst.* [Internet] 3:95–98.

Dyer KA, Canlfield TK, Gartler SM. 1989. Molecular cytological differentiation of active from inactive X domains in interphase: Implications for X chromosome inactivation. *Cytogenet. Genome Res.* 50:116–120.

Fang H, Bonora G, Lewandowski JP, Thakur J, Filippova GN, Henikoff S, Shendure J, Duan Z, Rinn JL, Deng X, et al. 2020. Trans- and cis-acting effects of *Firre* on epigenetic features of the inactive X chromosome. *Nat. Commun.* [Internet] 11.

Ferree PM, Barbash DA. 2009. Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in *Drosophila*. *PLoS Biol.* [Internet] 7:1000234.

Ferree PM, Prasad S. 2012. How Can Satellite DNA Divergence Cause Reproductive Isolation? Let Us Count the Chromosomal Ways. *Genet. Res. Int.* [Internet] 2012:1–11.

Figueiró H V., Li G, Trindade FJ, Assis J, Pais F, Fernandes G, Santos SHD, Hughes GM, Komissarov A, Antunes A, et al. 2017. Genome-wide signatures of complex introgression and adaptive evolution in the big cats. *Sci. Adv.* 3:1–13.

Figueroa DM, Darrow EM, Chadwick BP. 2015. Two novel *DXZ4*-associated long noncoding RNAs show developmental changes in expression coincident with heterochromatin formation at the human (Homo sapiens) macrosatellite repeat. *Chromosom. Res.* 23:733–752.

Filippova GN. 2007. Genetics and Epigenetics of the Multifunctional Protein CTCF. *Curr. Top. Dev. Biol.* 80:337–360.

Robert Finestra T, Gribnau J. 2017. X chromosome inactivation: silencing, topology and reactivation. *Curr. Opin. Cell Biol.* [Internet] 46:54–61.

Galupa R, Heard E. 2018. X-Chromosome Inactivation: A Crossroads Between Chromosome Architecture and Gene Regulation. *Annu. Rev. Genet.* 52:535–566.

Garrick D, Fiering S, Martin DIK, Whitelaw E. 1998. Repeat-induced gene silencing in mammals. *Nat. Genet.* [Internet] 18:56–59.

Ghurye J, Pop M, Koren S, Bickhart D, Chin CS. 2017. Scaffolding of long read assemblies using long range contact information. *BMC Genomics* 18:1–11.

Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, Phillippy AM, Koren S. 2019. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput. Biol.* [Internet] 15.

Giacalone J, Friedes J, Francke U. 1992. A novel GC–rich human macrosatellite VNTR in Xq24 is differentially methylated on active and inactive X chromosomes. *Nat. Genet.* 1:137–143.

Giorgetti L, Lajoie BR, Carter AC, Attia M, Zhan Y, Xu J, Chen CJ, Kaplan N, Chang HY, Heard E, et al. 2016. Structural organization of the inactive X chromosome in the mouse. *Nature* [Internet] 535:575–579.

Go VL, Vernon RG, Fritz IB. 1971. Studies on spermatogenesis in rats. I. Application of the sedimentation velocity technique to an investigation of spermatogenesis. *Canadian Journal of Biochemistry* 49:753-760.

Gondo Y, Okada T, Matsuyama N, Saitoh Y, Yanagisawa Y, Ikeda JE. 1998. Human megasatellite DNA RS447: Copy-number polymorphisms and interspecies conservation. *Genomics* [Internet] 54:39–49.

Good JM, Dean MD, Nachman MW. 2008. A complex genetic basis to X-linked hybrid male sterility between two species of house mice. *Genetics* 179:2213–2228.

Good JM, Giger T, Dean MD, Nachman MW. 2010. Widespread Over-expression of the X chromosome in sterile F1 hybrid mice. *PLoS Genet.* 6:30–32.

Graphodatsky A, Perelman P, O'Brien SJ. 2020. An Atlas of Mammalian Chromosomes. 2nd ed. New York (NY): John Wiley & Sons.

Graves JAM. 2006. Sex chromosome specialization and degeneration in mammals. *Cell* 124:901–914.

Gray AP. 1972. Mammalian hybrids, a check-list and bibliography, rev. ed. Commonwealth Agricultural Bureaux, Bucks, England.

Guan D, Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R, Durbin R. 2020. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* [Internet] 36:2896–2898.

Haldane JBS. 1922. Sex Ratio and Unisexual Sterility in Hybrid Animals. *J. Genet* [Internet]:101–109.

Handel MA. 2004. The XY body: A specialized meiotic chromatin domain. *Exp. Cell Res.* 296:57–63.

Harris RS. 2007. Improved pairwise alignment of genomic DNA. Ph.D. Thesis, The Pennsylvania State University.

Heng L. 2018. seqTK, https://github.com/lh3/seqtk.

Henikoff S. 1998. Conspiracy of silence among repeated transgenes. *BioEssays* 20:532–535.

Hewitt JE, Lyle R, Clark LN, Valleley EM, Wright TJ, Wijmenga C, Van Deutekom JC t., Francis F, Sharpe PT, Hofker M, et al. 1994. Analysis of the tandem repeat locus *D4Z4* associated with facioscapulohumeral muscular dystropothhy. *Hum. Mol. Genet.* [Internet] 3:1287–1295.

Homyack JA, Vashon JH, Libby C, Lindquist EL, Loch S, McAlpine DF, Pilgrim KL, Schwartz MK. 2008. Canada lynx-bobcat (Lynx canadensis x L. rufus) hybrids at the southern periphery of lynx range in Maine, Minnesota and New Brunswick. *Am. Midl. Nat.* 159:504–508.

Horakova AH, Calabrese JM, McLaughlin CR, Tremblay DC, Magnuson T, Chadwick BP. 2012. The mouse *DXZ4* homolog retains Ctcf binding and proximity to *Pls3* despite substantial organizational differences compared to the primate macrosatellite. *Genome Biol.* 13:R70.

Horakova AH, Moseley SC, Mclaughlin CR, Tremblay DC, Chadwick BP. 2012. The macrosatellite *DXZ4* mediates CTCF-dependent long-range intrachromosomal interactions on the human inactive X chromosome. *Hum. Mol. Genet.* [Internet] 21:4367–4377.

Howard JG, Brown JL, Bush M, Wildt DE. 1990. Teratospermic and Normospermic Domestic Cats: Ejaculate Traits, Pituitary-Gonadal Hormones, and Improvement of Spermatozoal Motility and Morphology After Swim-Up Processing. *J. Androl.* 11:204–215.

Hoyer-Fender S. 2003. Molecular aspects of XY body formation. *Cytogenet. Genome Res.* 103:245–255.

Hsieh PH, Vollger MR, Dang V, Porubsky D, Baker C, Cantsilieris S, Hoekzema K, Lewis AP, Munson KM, Sorensen M, et al. 2019. Adaptive archaic introgression of copy number variants and the discovery of previously unknown human genes. *Science.* 366.

Hu J, Fan J, Sun Z, Liu S. 2020. NextPolish: A fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* [Internet] 36:2253–2255.

Ishishita S, Tsuboi K, Ohishi N, Tsuchiya K, Matsuda Y. 2015. Abnormal pairing of X and y sex chromosomes during meiosis I in interspecific hybrids of *Phodopus campbelli* and *P. sungorus*. *Sci. Rep.* 5:1–9.

Jablonka E, Lamb MJ. 1991. Sex chromosomes and speciation. *Proc. R. Soc. B Biol. Sci.* [Internet] 243:203–208.

Jégu T, Aeby E, Lee JT. 2017. The X chromosome in space. *Nat. Rev. Genet.* [Internet] 18:377–389.

Johnson WE, Eizirik E, Pecon-Slattery J, Murphy WJ, Antunes A, Teeling E, O'Brien SJ. 2006. The late Miocene radiation of modern felidae: A genetic assessment. *Science.* 311:73–77.

Jung M, Wells D, Rusch J, Ahmad S, Marchini J, Myers SR, Conrad DF. 2019. Unified single-cell analysis of testis gene regulation and pathology in five mouse strains. *Elife* 8.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42:348–354.

Kentepozidou E, Aitken SJ, Feig C, Stefflova K, Ibarra-Soria X, Odom DT, Roller M, Flicek P. 2020. Clustered CTCF binding is an evolutionary mechanism to maintain topologically associating domains. *Genome Biol.* [Internet] 21.

Kierszenbaum AL, Tres LL. 1974. Nucleolar and perichromosomal rna synthesis during meiotic prophase in the mouse testis. *J. Cell Biol.* [Internet] 60:39–53.

Kim JH, Antunes A, Luo SJ, Menninger J, Nash WG, O'Brien SJ, Johnson WE. 2006. Evolutionary analysis of a large mtDNA translocation (numt) into the nuclear genome of the Panthera genus species. *Gene* 366:292–302.

Knibiehler B, Mirre C, Hartung M, Jean P, Stahl A, Delanversin A, Soler M. 1981. Sex vesicle-associated nucleolar organizers in mouse spermatocytes: Localization, structure, and function. *Cytogenet. Genome Res.* 31:47–57.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillipy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27:722–736.

Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, Hiendleder S, Williams JL, Smith TPL, Phillippy AM. 2018. De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* [Internet] 36:1174–1182.

Korneliussen TS, Albrechtsen A, Nielsen R. 2014. ANGSD: Analysis of next generation sequencing data. BMC Bioinf. 15:356.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35:1547–1549.

Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, Ponting CP, Odom DT, Marques AC. 2012. Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genet.* 8.

Lahn BT, Page DC. 1999. Four evolutionary strata on the human X chromosome. *Science.* [Internet] 286:964–967.

Langmead B. 2010. Aligning short sequencing reads with Bowtie. *Curr. Protoc. Bioinforma.* [Internet] 32.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* [Internet] 9:357–359.

Larson EL, Keeble S, Vanderpool D, Dean MD, Good JM. 2016. The Composite Regulatory Basis of the Large X-Effect in Mouse Speciation. *Mol. Biol. Evol.* 34:282–295.

Larson EL, Vanderpool D, Keeble S, Zhou M, Sarver BAJ, Smith AD, Dean MD, Good JM. 2016. Contrasting levels of molecular evolution on the mouse X chromosome. *Genetics* 203:1841–1857.

Lea AJ, Tung J, Zhou X. 2015. A Flexible, Efficient Binomial Mixed Model for Identifying Differential DNA Methylation in Bisulfite Sequencing Data. *PLoS Genet.* [Internet] 11:1005650.

Lifschytz E, Lindsley DL. 1972. The role of X-chromosome inactivation during spermatogenesis (Drosophila-allocycly-chromosome evolution-male sterility-dosage compensation). *Proc. Natl. Acad. Sci. U. S. A.* [Internet] 69:182–186.

Li G, Davis BW, Eizirik E, Murphy WJ. 2016. Phylogenomic evidence for ancient hybridization in the genomes of living cats (Felidae). *Genome Res.* 26:1–11.

Li G, Davis BW, Raudsepp T, Wilkerson AJP, Mason VC, Ferguson-Smith M, O'Brien PC, Waters PD, Murphy WJ. 2013. Comparative analysis of mammalian y chromosomes illuminates ancestral structure and lineage-specific evolution. *Genome Res.* 23:1486–1495.

Li G, Hillier LDW, Grahn RA, Zimin A V., David VA, Menotti-Raymond M, Middleton R, Hannah S, Hendrickson S, Makunin A, et al. 2016. A high-resolution SNP array-based linkage map anchors a new domestic cat draft genome assembly and provides detailed patterns of recombination. *G3 Genes, Genomes, Genet.* [Internet] 6:1607–1616.

Li G, Figueiró H V., Eizirik E, Murphy WJ, Yoder A. 2019. Recombination-Aware Phylogenomics Reveals the Structured Genomic Landscape of Hybridizing Cat Species. *Mol. Biol. Evol.* [Internet] 36:2111–2126.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* 25:1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079.

Liu WS. 2019. Mammalian Sex Chromosome Structure, Gene Content, and Function in Male Fertility. *Annu. Rev. Anim. Biosci.* 7:103–124.

Logsdon GA, Vollger MR, Hsieh P, Mao Y, Liskovykh MA, Koren S, Nurk S, Mercuri L, Dishuck PC, Rhie A, et al. 2020. The structure, function, and evolution of a complete human chromosome 8. *bioRxiv* [Internet]:2020.09.08.285395.

Lopez J V., Cevario S, O'Brien SJ. 1996. Complete nucleotide sequences of the domestic cat (Felis catus) mitochondrial genome and a transposed mtDNA tandem repeat (Numt) in the nuclear genome. *Genomics* 33:229–246.

Low WY, Tearle R, Liu R, Koren S, Rhie A, Bickhart DM, Rosen BD, Kronenberg ZN, Kingan SB, Tseng E, et al. 2020. Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle. *Nat. Commun.* [Internet] 11.

Lucotte EA, Skov L, Jensen JM, Macià MC, Munch K, Schierup MH. 2018. Dynamic copy number evolution of X-and Y-linked ampliconic genes in human populations. *Genetics* [Internet] 209:907–920.

Luo SJ, Johnson WE, Martenson J, Antunes A, Martelli P, Uphyrkina O, Traylor-Holzer K, Smith JL, O'Brien SJ. 2008. Subspecies genetic assignments of worldwide captive tigers increase conservation value of captive populations. Curr Biol. 18:592–596.

Lustyk D, Kinský S, Ullrich KK, Yancoskie M, Kašíková L, Gergelits V, Sedlacek R, Chan YF, Odenthal-Hesse L, Forejt J, et al. 2019. Genomic structure of *Hstx2* modifier of *Prdm9*-dependent hybrid male sterility in mice. *Genetics* [Internet] 213:1047–1063.

Lyons LA, Laughlin TF, Copeland NG, Jenkins NA, Womack JE, O'Brien SJ. 1997. Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian genomes. Nat Genet. 15:47–56.

Lyon MF. 1961. Gene action in the X-chromosome of the mouse (*mus musculus L.*). *Nature* [Internet] 190:372–373.

Mahadevaiah SK, Royo H, VandeBerg JL, McCarrey JR, Mackay S, Turner JMA. 2009. Key Features of the X Inactivation Process Are Conserved between Marsupials and Eutherians. *Curr. Biol.* [Internet] 19:1478–1484.

Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* [Internet] 14:1–14.

Martin M. 2011. Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads. *EMBnet J.* [Internet] 17:10–12.

Marques JP, Seixas FA, Farelo L, Callahan CM, Good JM, Montgomery WI, Reid N, Alves PC, Boursot P, Melo-Ferreira J. 2019. An Annotated Draft Genome of the Mountain Hare (Lepus timidus). *Genome Biol. Evol.* 12:3656–3662.

Masly JP, Presgraves DC. 2007. High-Resolution Genome-Wide Dissection of the Two Rules of Speciation in *Drosophila*. *PLoS Biol.* [Internet] 5:1890–1898.

Mayr E. 1942. *Systematics and the origin of species, from the viewpoint of a zoologist*. New York: Columbia University Press.

McCarrey JR, Berg WM, Paragioudakis SJ, Zhang PL, Dilworth DD, Arnold BL, Rossi JJ. 1992. Differential transcription of Pgk genes during spermatogenesis in the mouse. *Dev. Biol.* 154:160–168.

McCarrey JR, Dilworth DD. 1992. Expression of Xist in mouse germ cells correlates with X–chromosome inactivation. *Nat. Genet.* [Internet] 2:200–203.

McCarthy DJ, Chen Y, Smyth GK. 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* [Internet] 40:4288–4297.

McKee BD, Handel MA. 1993. Sex chromosomes, recombination, and chromatin conformation. *Chromosoma* 102:71–80.

McLaughlin CR, Chadwick BP. 2011. Characterization of *DXZ4* conservation in primates implies important functional roles for CTCF binding, array expression and tandem repeat organization on the X chromosome. *Genome Biol.* [Internet] 12.

Meisel RP, Connallon T. 2013. The faster-X effect: Integrating theory and data. *Trends Genet.* 29:537–544.

Menotti-Raymond M, David VA, Lyons LA, Schäffer AA, Tomlin JF, Hutton MK, O'Brien SJ. 1999. A genetic linkage map of microsatellites in the domestic cat (Felis catus). *Genomics* [Internet] 57:9–23.

Menotti-Raymond M, David VA, Chen ZQ, Menotti KA, Sun S, Schaffer AA, Agarwala R, Tomlin JF, O'Brien SJ, Murphy WJ. 2003. Second-Generation Integrated Genetic Linkage/Radiation Hybrid Maps of the Domestic Cat (Felis catus). In: Journal of Heredity. Vol. 94. Oxford University Press. p. 95–106.

Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. 2019. PANTHER version 14: More genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.* [Internet] 47:D419–D426.

Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA, et al. 2020. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* [Internet] 585:79–84.

Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. 2018. Versatile genome assembly evaluation with QUAST-LG. In: Bioinformatics. Vol. 34. Oxford University Press. p. i142–i150.

Miller SA, Dykes DD, Polesky HF. 1988. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res.* 16:1215.

Minajigi A, Froberg JE, Wei C, Sunwoo H, Kesner B, Colognori D, Lessing D, Payer B, Boukhali M, Haas W, et al. 2015. A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation. *Science.* [Internet] 349.

Modi WS, O'Brien SJ. 1988. Quantitative cladistic analyses of chromosomal banding data among species in three orders of mammals: hominoid primates, felids and arvicolid rodents. In: Gustafson JP, Appels R., editors Chromosome Structure and Function. New York (NY): Plenum Publishing Corporation. p. 215–242.

Montague MJ, Li G, Golfi B, Khan R, Aken BL, Searle SMJ, Minx P, Hillier LDW, Koboldt DC, Davis BW, et al. 2014. Comparative analysis of the domestic cat genome reveals genetic signatures underlying feline biology and domestication. *Proc. Natl. Acad. Sci. U. S. A.* 111:17230–17235.

Moore CM, Janish C, Eddy CA, Hubbard GB, Leland MM, Rogers J. 1999. Cytogenetic and fertility studies of a rheboon, rhesus macaque (*Macaca mulatta*) x baboon (*Papio hamadryas*) cross: Further support for a single karyotype nomenclature. *Am. J. Phys. Anthropol.* 110:119–127.

Morán T, Fontdevila A. 2014. Genome-wide dissection of hybrid sterility in drosophila confirms a polygenic threshold architecture. *J. Hered.* [Internet] 105:381–396.

Moretti C, Vaiman D, Tores F, Cocquet J. 2016. Expression and epigenomic landscape of the sex chromosomes in mouse post-meiotic male germ cells. *Epigenetics Chromatin* 9:47.

Murphy WJ, O'Brien SJ. 2007. Designing and optimizing comparative anchor primers for comparative gene mapping and phylogenetic inference. *Nat. Protoc.* 2:3022–3030.

Murphy WJ, Pearks Wilkerson AJ, Raudsepp T, Agarwala R, Schaffer AA, Stanyon R, Chowdhary BP. 2006. Novel Gene Acquisition on Carnivore Y Chromosomes. *PLoS Genet.* [Internet] preprint:e43.

Nagano T, Lubling Y, Yaffe E, Wingett SW, Dean W, Tanay A, Fraser P. 2015. Single-cell Hi-C for genome-wide detection of chromatin interactions that occur simultaneously in a single cell. *Nat. Protoc.* [Internet] 10:1986–2003.

Namekawa SH, Park PJ, Zhang LF, Shima JE, McCarrey JR, Griswold MD, Lee JT. 2006. Postmeiotic Sex Chromatin in the Male Germline of Mice. *Curr. Biol.* 16:660–667.

Nattestad M, Schatz MC. 2016. Assemblytics: A web analytics tool for the detection of variants from an assembly. *Bioinformatics* 32:3021–3023.

Nextomics. 2020. NextDenovo, https://github.com/Nextomics/NextDenovo/.

Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Phillippy AM, Koren S. 2020. HiCanu: Accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* 30:1291–1305.

O'Brien SJ, Johnson WE, Driscoll CA, Dobrynin P, Marker L. 2017. Conservation genetics of the cheetah: lessons learned and new opportunities. J Hered. 108:671–677.

O'Brien SJ, Menotti-Raymond M, Murphy WJ, Yuhki N. 2002. The feline genome project. Annu Rev Genet. 36:657–686.

O'Brien SJ, Nash WG, Winkler CA, Reeves RH. 1982. Genetic analysis in the domestic cat as an animal model for inborn errors, cancer and evolution. Prog Clin Biol Res. 94:67–90.

O'Brien SJ, Troyer JL, Roelke M, Marker L, Pecon-Slattery J. 2006. Plagues and adaptation: Lessons from the Felidae models for SARS and AIDS. Biol Conserv. 131:255–267.

Ogaki Y, Fukuma M, Shimizu N. 2020. Repeat induces not only gene silencing, but also gene activation in mammalian cells. *PLoS One* 15.

Ohno, S. 1967. *Sex Chromosomes and Sex Linked Genes*. Springer-Verlag; Berlin.

Ong CT, Corces VG. 2014. CTCF: An architectural protein bridging genome topology and function. *Nat. Rev. Genet.* [Internet] 15:234–246.

Patel L, Kang R, Rosenberg SC, Qiu Y, Raviram R, Chee S, Hu R, Ren B, Cole F, Corbett KD. 2019. Dynamic reorganization of the genome shapes the recombination landscape in meiotic prophase. *Nat. Struct. Mol. Biol.* 26:164–174.

Pearks Wilkerson AJ, Raudsepp T, Graves T, Albracht D, Warren W, Chowdhary BP, Skow LC, Murphy WJ. 2008. Gene discovery and comparative analysis of X-degenerate genes from the domestic cat Y chromosome. *Genomics* 92:329–338.

Pohlers M, Mauro Calabrese J, Magnuson T. 2014. Small RNA expression from the human macrosatellite *DXZ4*. *G3 Genes, Genomes, Genet.* [Internet] 4:1981–1989.

Pollard MO, Gurdasani D, Mentzer AJ, Porter T, Sandhu MS. 2018. Long reads: their purpose and place. *Hum. Mol. Genet.* [Internet] 27:R234–R241.

Presgraves DC. 2008. Sex chromosomes and speciation in *Drosophila*. *Trends Genet* [Internet] 24:336–343.

Presgraves DC. 2010. The molecular evolutionary basis of species formation. *Nat. Rev. Genet.* [Internet] 11:175–180.

Presgraves DC. 2018. Evaluating genomic signatures of "the large X-effect" during complex speciation. *Mol. Ecol.*:1–9.

Presgraves DC, Orr HA. 1995. Haldane's Rule in Taxa Lacking a Hemizygous X. *Proc. Natl. Acad. Sci. U.S.A.* 93:632.

Proskuryakova AA, Kulemzina AI, Perelman PL, Makunin AI, Larkin DM, Farré M, Kukekova A V., Lynn Johnson J, Lemskaya NA, Beklemisheva VR, et al. 2017. X chromosome evolution in cetartiodactyla. *Genes (Basel).* 8.

Pukazhenthi BS, Neubauer K, Jewgenow K, Howard JG, Wildt DE. 2006. The impact and potential etiology of teratospermia in the domestic cat and its wild relatives. *Theriogenology* 66:112–121.

Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* [Internet] 26:841–842.

R Core Team. 2019. A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.

Ramani V, Cusanovich DA, Hause RJ, Ma W, Qiu R, Deng X, Blau CA, Disteche CM, Noble WS, Shendure J, et al. 2016. Mapping 3D genome architecture through in situ DNase Hi-C. *Nat. Protoc.* [Internet] 11:2104–2121.

Ramírez-Colmenero A, Oktaba K, Fernandez-Valverde SL. 2020. Evolution of Genome-Organizing Long Non-coding RNAs in Metazoans. *Front. Genet.* [Internet] 11:589697.

Rao SSP, Huang SC, Glenn St Hilaire B, Engreitz JM, Perez EM, Kieffer-Kwon KR, Sanborn AL, Johnstone SE, Bascom GD, Bochkov ID, et al. 2017. Cohesin Loss Eliminates All Loop Domains. *Cell* [Internet] 171:305-320.e24.

Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* [Internet] 159:1665–1680.

Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Gedman GL, et al. 2020. Towards complete and error-free genome assemblies of all vertebrate species. *bioRxiv* [Internet] 52:2020.05.22.110833.

Rice ES, Koren S, Rhie A, Heaton MP, Kalbfleisch TS, Hardy T, Hackett PH, Bickhart DM, Rosen BD, Ley B Vander, et al. 2020. Continuous chromosome-scale haplotypes assembled from a single interspecies F1 hybrid of yak and cattle. *Gigascience* [Internet] 9:1–9.

Robinson MD, McCarthy DJ, Smyth GK. 2009. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* [Internet] 26:139–140.

Rodríguez Delgado CL, Waters PD, Gilbert C, Robinson TJ, Graves JAM. 2009. Physical mapping of the elephant X chromosome: Conservation of gene order over 105 million years. *Chromosom. Res.* 17:917–926.

Royo H, Polikiewicz G, Mahadevaiah SK, Prosser H, Mitchell M, Bradley A, De Rooij DG, Burgoyne PS, Turner JMA. 2010. Evidence that meiotic sex chromosome inactivation is essential for male fertility. *Curr. Biol.* 20:2117–2123.

Schaap M, Lemmers RJLF, Maassen R, van der Vliet PJ, Hoogerheide LF, van Dijk HK, Baştürk N, de Knijff P, van der Maarel SM. 2013. Genome-wide analysis of macrosatellite repeat copy number variation in worldwide populations: Evidence for differences and commonalities in size distributions and size restrictions. *BMC Genomics* 14.

Schimenti J. 2005. Synapsis or silence. *Nat. Genet.* [Internet] 37:11–13.

Schwahn DJ, Wang RJ, White MA, Payseur BA. 2018. Genetic dissection of hybrid male sterility across stages of spermatogenesis. *Genetics* [Internet] 210:1453–1465. Available from: https://doi.org/10.1534/genetics.118.301658

Seabury CM, Oldeschulte DL, Saatchi M, Beever JE, Decker JE, Halley YA, Bhattarai EK, Molaei M, Freetly HC, Hansen SL, et al. 2017. Genome-wide association study for feed efficiency and growth traits in U.S. beef cattle. *BMC Genomics* 18.

Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q, Nordborg M. 2012. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* 44:825–830.

Shumate A, Salzberg SL. 2020. Liftoff: an accurate gene annotation mapping tool. *bioRxiv* [Internet]:2020.06.24.169680.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31:3210–3212.

Smit AFA, Hubley R, Green P. 2013-2015. *RepeatMasker Open-4.0*. <http://www.repeatmasker.org>.

Solari AJ. 1974. The behavior of the XY pair in mammals, *Int. Rev. Cytol* 38**:**273–317.

Sosa E, Flores L, Yan W, McCarrey JR. 2015. Escape of X-linked miRNA genes from meiotic sex chromosome inactivation. *Dev.* [Internet] 142:3791–3800.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* [Internet] 30:1312–1313.

Tange O. 2011. GNU Parallel: The Command-Line Power Tool. *USENIX Mag.*:42–47.

Thomsen PD, Schauser K, Bertelsen MF, Vejlsted M, Grøndahl C, Christensen K. 2011. Meiotic studies in infertile domestic pig-babirusa hybrids. *Cytogenet. Genome Res.* [Internet] 132:124–128.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* [Internet] 7:562–578.

Tremblay DC, Moseley S, Chadwick BP. 2011. Variation in array size, monomer composition and expression of the macrosatellite *DXZ4*. *PLoS One* [Internet] 6:18969.

Trigo TC, Freitas TR, Kunzler G, Cardoso L, Silva JC, Johnson WE, O'Brien SJ, Bonatto SL, Eizirik E. 2008. Inter-species hybridization among Neotropical cats of the genus Leopardus, and evidence for an introgressive hybrid zone between L. geoffroyi and L. tigrinus in southern Brazil. Mol Ecol. 17:4317–4333.

Trigo TC, Schneider A, de Oliveira TG, Lehugeur LM, Silveira L, Freitas TR, Eizirik E. 2013. Molecular data reveal complex hybridization and a cryptic species of neotropical wild cat. Curr Biol. 23:2528–2533.

Turelli M, Orr HA. 1995. The Dominance Theory of Haldane's Rule. *Genetics* 140:389–402.

Turner JMA. 2007. Meiotic sex chromosome inactivation. *Development* [Internet] 134:1823–1831.

Turner JMA, Mahadevaiah SK, Elliott DJ, Garchon HJ, Pehrson JR, Jaenisch R, Burgoyne PS. 2002. Meiotic sex chromosome inactivation in male mice with targeted disruptions of Xist. *J. Cell Sci.* [Internet] 115:4097–4105.

Turner JMA, Mahadevaiah SK, Ellis PJI, Mitchell MJ, Burgoyne PS. 2006. Pachytene asynapsis drives meiotic sex chromosome inactivation and leads to substantial postmeiotic repression in spermatids. *Dev. Cell* 10:521–529.

Turner JMA, Mahadevaiah SK, Fernandez-Capetillo O, Nussenzweig A, Xu X, Deng CX, Burgoyne PS. 2005. Silencing of unsynapsed meiotic chromosomes in the mouse. *Nat. Genet.* [Internet] 37:41–47.

Turner LM, Harr B. 2014. Genome-wide mapping in a house mouse hybrid zone reveals hybrid sterility loci and Dobzhansky-Muller interactions. *Elife* 3:2504.

Turner LM, White MA, Tautz D, Payseur BA. 2014. Genomic Networks of Hybrid Sterility. *PLoS Genet.* [Internet] 10:1004162.

Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, et al. 2015. Tissue-based map of the human proteome. *Science.* [Internet] 347.

Van Overveld PGM, Lemmers RJFL, Sandkuijl LA, Enthoven L, Winokur ST, Bakels F, Padberg GW, Van Ommen GJB, Frants RR, Van Der Maarel SM. 2003.

Hypomethylation of *D4Z4* in 4q-linked and non-4q-linked facioscapulohumeral muscular dystrophy. *Nat. Genet.* [Internet] 35:315–317.

Vernet N, Mahadevaiah SK, de Rooij DG, Burgoyne PS, Ellis PJI. 2016. *Zfy* genes are required for efficient meiotic sex chromosome inactivation (MSCI) in spermatocytes. *Hum. Mol. Genet.* 25:5300–5310.

Vollger MR, Dishuck PC, Sorensen M, Welch AME, Dang V, Dougherty ML, Graves-Lindsay TA, Wilson RK, Chaisson MJP, Eichler EE. 2019. Long-read sequence and assembly of segmental duplications. *Nat. Methods* 16:88–94.

Wang H, Höög C. 2006. Structural damage to meiotic chromosomes impairs DNA recombination and checkpoint control in mammalian oocytes. *J. Cell Biol.* [Internet] 173:485–495.

Wang PJ, McCarrey JR, Yang F, Page DC. 2001. An abundance of X-linked genes expressed in spermatogonia. *Nat. Genet.* 27:422–426.

Wurster-Hill DH, Centerwall WR. 1982. The interrelationships of chromo- some banding patterns in canids, mustelids, hyena, and felids. Cytogenet Cell Genet. 34:178–192

Yan W, McCarrey JR. 2009. Sex chromosome inactivation in the male. *Epigenetics* 4:452–456.

Yang F, Deng X, Ma W, Berletch JB, Rabaia N, Wei G, Moore JM, Filippova GN, Xu J, Liu Y, et al. 2015. The lncRNA *Firre* anchors the inactive X chromosome to the nucleolus by binding CTCF and maintains H3K27me3 methylation. *Genome Biol.* [Internet] 16.

Yunis JJ, Yasmineh WG. 1971. Heterochromatin, satellite DNA, and cell function. Structural DNA of eukaryotes may support and protect genes and aid in speciation. *Science* [Internet] 174:1200–1209.

Zhang LF, Huynh KD, Lee JT. 2007. Perinucleolar Targeting of the Inactive X during S Phase: Evidence for a Role in the Maintenance of Silencing. *Cell* 129:693–706.

Zhang X, Wu R, Wang Y, Yu J, Tang H. 2020. Unzipping haplotypes in diploid and polyploid genomes. Comput Struct Biotechnol J. 18:66–72.

APPENDIX A SUPPLEMENTAL TEXT

Text A3.1. Details of Jungle Cat *de novo* Assembly

The initial assembly using NextDenovo yielded 174 raw contigs with a single chimeric contig joining chromosomes B3 and E1. 68 haplotigs, duplicate contigs representative of the alternative haplotype, were identified and purged resulting in 106 final contigs. Prior to scaffolding, we identified and isolated 13 Chr. Y contigs totaling 3.48 Mb in length. This was done to prevent incorporation of repetitive Y chromosome contigs into highly similar, but paralogous autosomal regions (Brashear et al., 2018) during scaffolding. The homologous X and Y pseudoautosomal region (PAR) was collapsed into a single contig with ~206 kb of male specific single copy Y sequence incorporated at the end. Male-specific single copy sequence was removed and manually joined to two additional Y contigs to generate a single Chr. Y scaffold representing the single copy region (SCR) (**Table C.A.1**). Using BLAST and LASTZ Alignments we were able to identify a complete Jungle cat mitochondrial genome (MT) sequence 17,251 bp in length. The new Jungle cat MT was 598 bp longer than the previous short-read based sequence (16,653 bp; Li et al., 2016) and more similar in length to the reference domestic cat mitochondrial genome (felCat9: 17,009 bp) (Buckley et al., 2020). Self-self dotplot comparisons demonstrated this gain was due to assembly of the highly repetitive control region present in domestic cat but lacking in the published Jungle cat MT sequence (**Figure B.A.1**). Jungle cat contigs were aligned to the single haplotype domestic cat assembly (Fca-508: GCA_016509815.1) using Nucmer and revealed a majority of chromosome arms were captured in single contigs (70%) with only a single chimeric contig observed prior to scaffolding (**Figure B.A.2**). Scaffolding of the contigs using Hi-C data yielded chromosome length-scaffolds (N50=148.6 Mb). Alignments between the Jungle cat and domestic cat chromosomes revealed large-scale collinearity between the two species, as previously suggested by karyotypic analysis (O'Brien 2020, Atlas of Mammalian Chromosomes, 2ndEdition) (**Figure B.A.3**). A total of 54 gaps remained in the 19 chromosome-length scaffolds (**Table C.A.2**). 32 contigs remained unplaced representing 0.3% of the un-gapped assembly length (11 of these were orthologous to ampliconic regions of the domestic cat ChrY, Brashear et al. 2018). BUSCO analysis revealed that 95% of the 9226 mammalian BUSCOs were represented in the final Jungle cat assembly with most (98%) being complete single-copy. 33.67% of the genome was identified as repetitive (**Table C.A.3**). Gene liftover from the felCat9 reference assembly resulted in the annotation of 19,611 protein coding genes with 203 of these identified as an extra copy relative to the reference (**Tables C.A.4 & C.A.5**). Structural variant analysis comparing the Jungle cat assembly to Fca-508 revealed an increase of 5.7 Mb due to repeat expansions as well as insertions of various size (**Table C.A.6**), which accounts for a majority of the 6 Mb un-gapped assembly length difference between the two species.

Figure B2.1. Phase Haplotype Analysis read comparison for reference and replacement crosses. a) Assuming *TrioCanu* sorted the reads correctly to the reference cross, green arrows indicate correctly sorted reads, while the dashed red lines indicate incorrect sorted reads in the replacement cross. b) Organization of incorrectly sorted reads into subtypes that describe how the read was sorted differently compared to the reference cross.

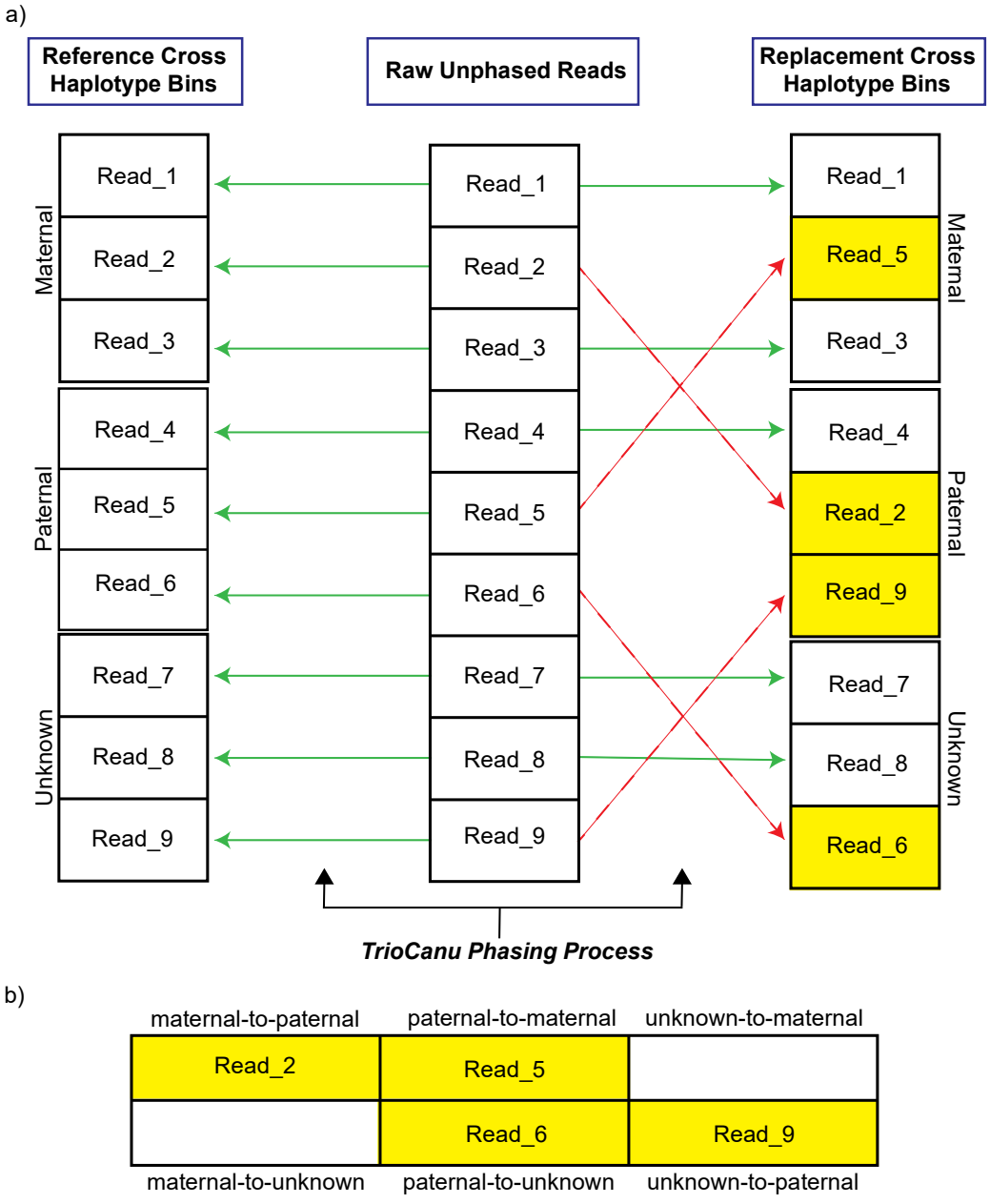


Figure B2.2. Nucmer alignments of assembly contigs to the felCat9 reference genome. **a)** Domestic cat polished contigs. Chimeric contig ctg000007 is an interchromosomal join between Chr D4 and Chr E2. **b)** Asian leopard cat polished contigs. Chimeric contig ctg000062 is an interchromosomal join between Chr B1 and Chr E4 (Chr F1 in felCat9) and chimeric contig ctg000126 is an interchromosomal join between Chr A2 and Chr E1.

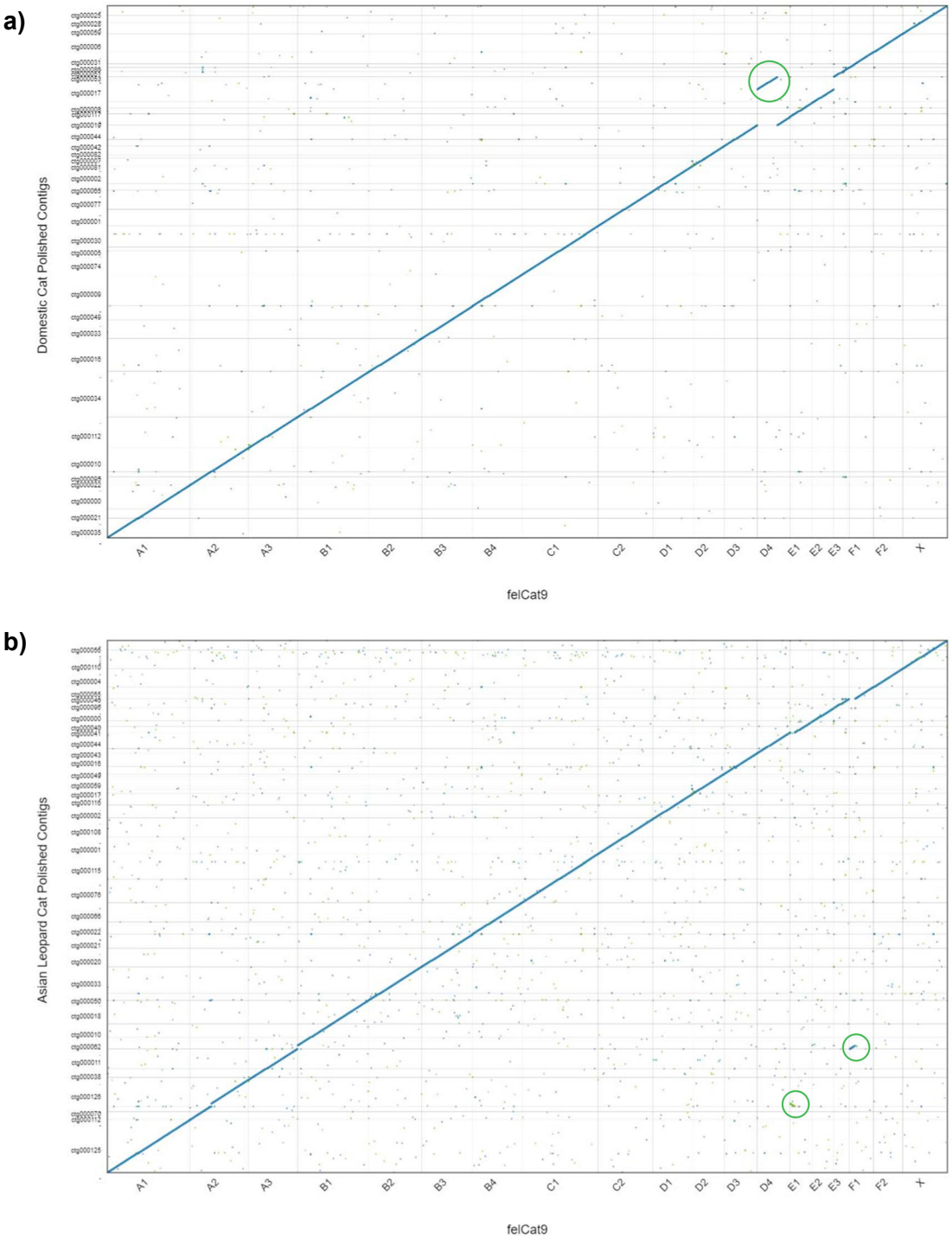
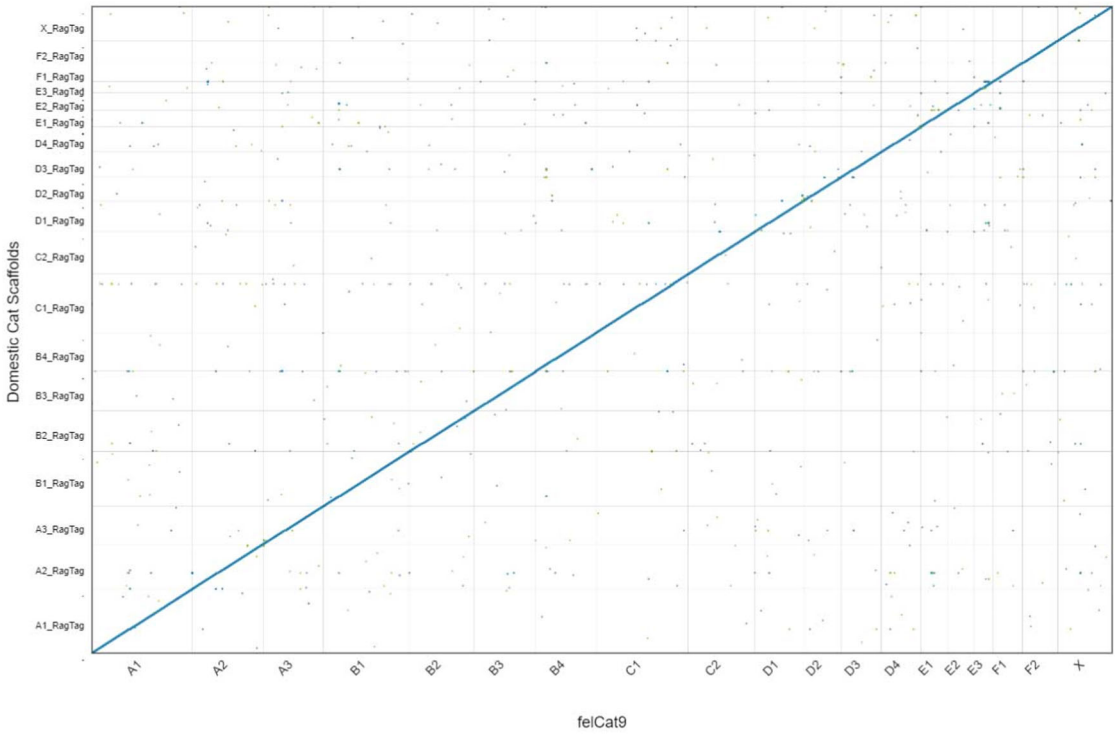


Figure B2.3. Nucmer alignments of assembly scaffolds to felCat9 reference genome. **a)** Domestic cat scaffolds. **b)** Asian leopard cat scaffolds.

a)



b)

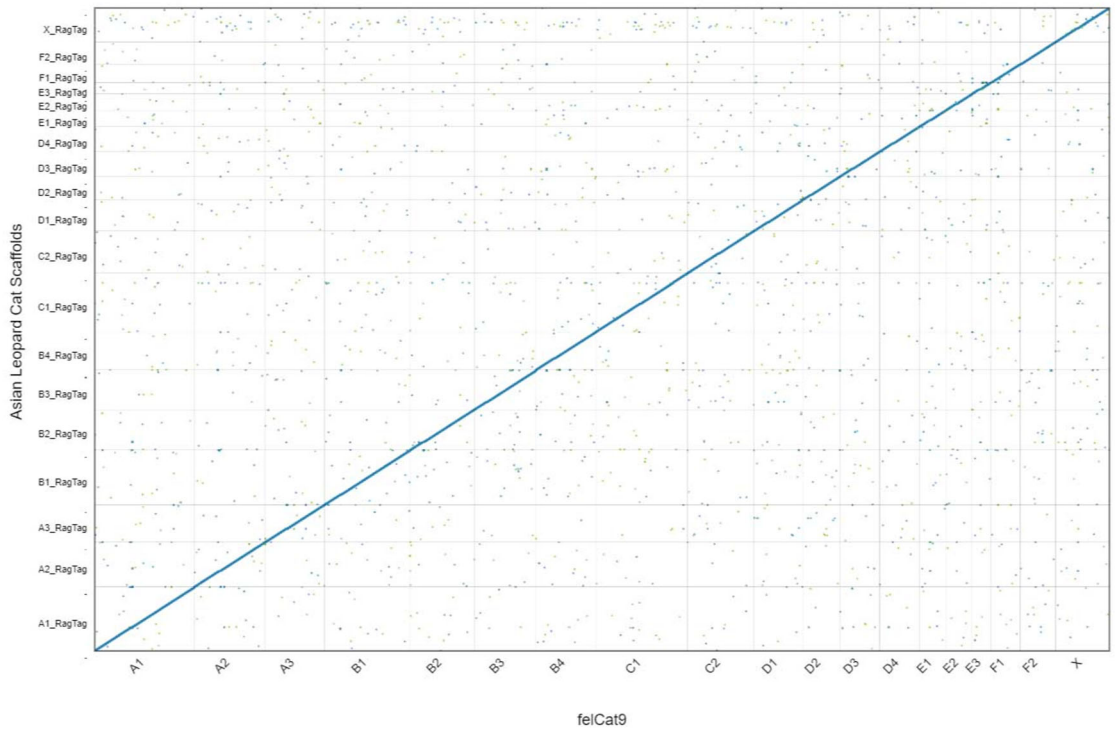


Figure B2.4. Hi-C contact map for the domestic cat haploid assembly.

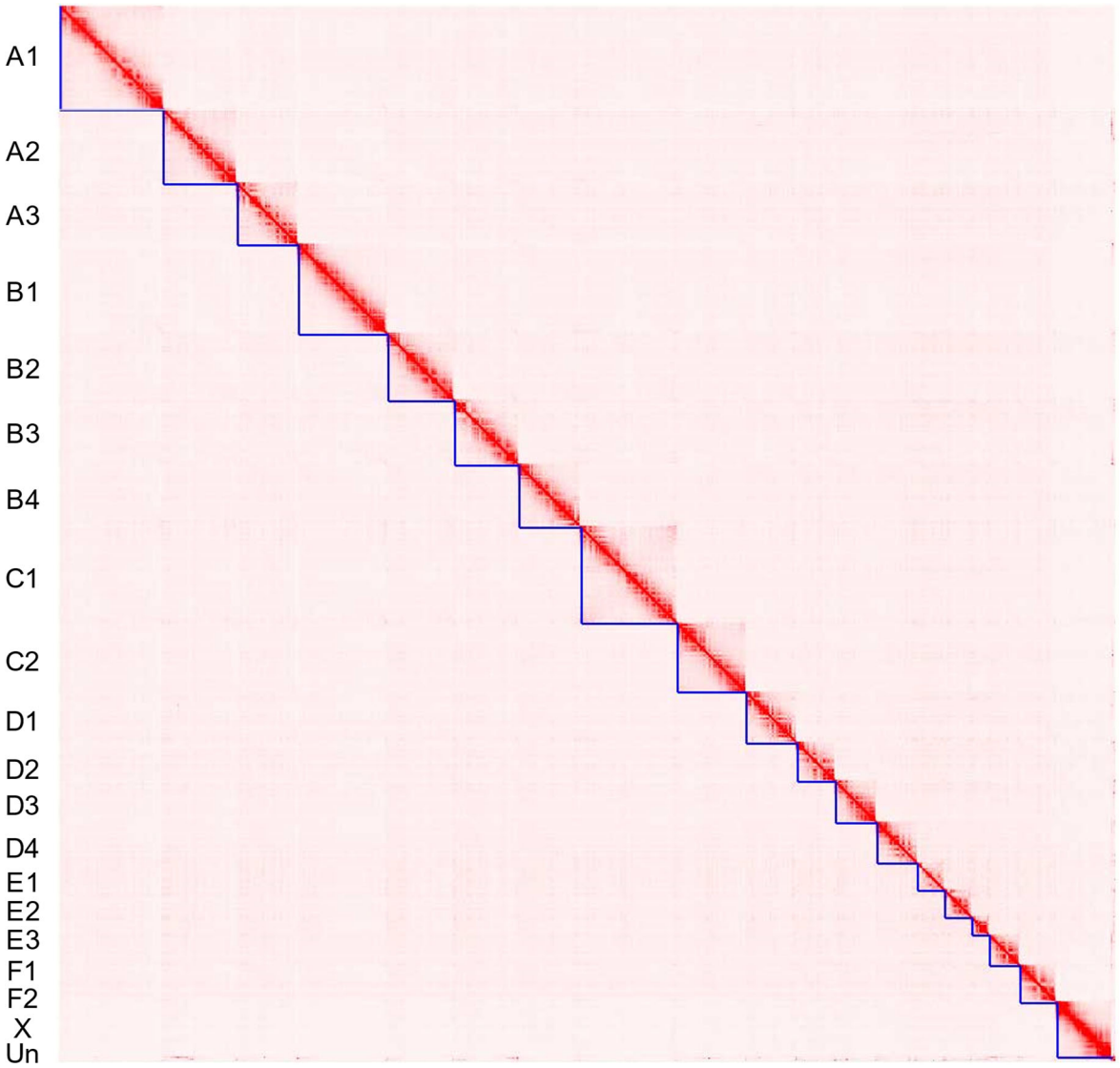


Figure B2.5. Hi-C contact map for the leopard cat haploid assembly.

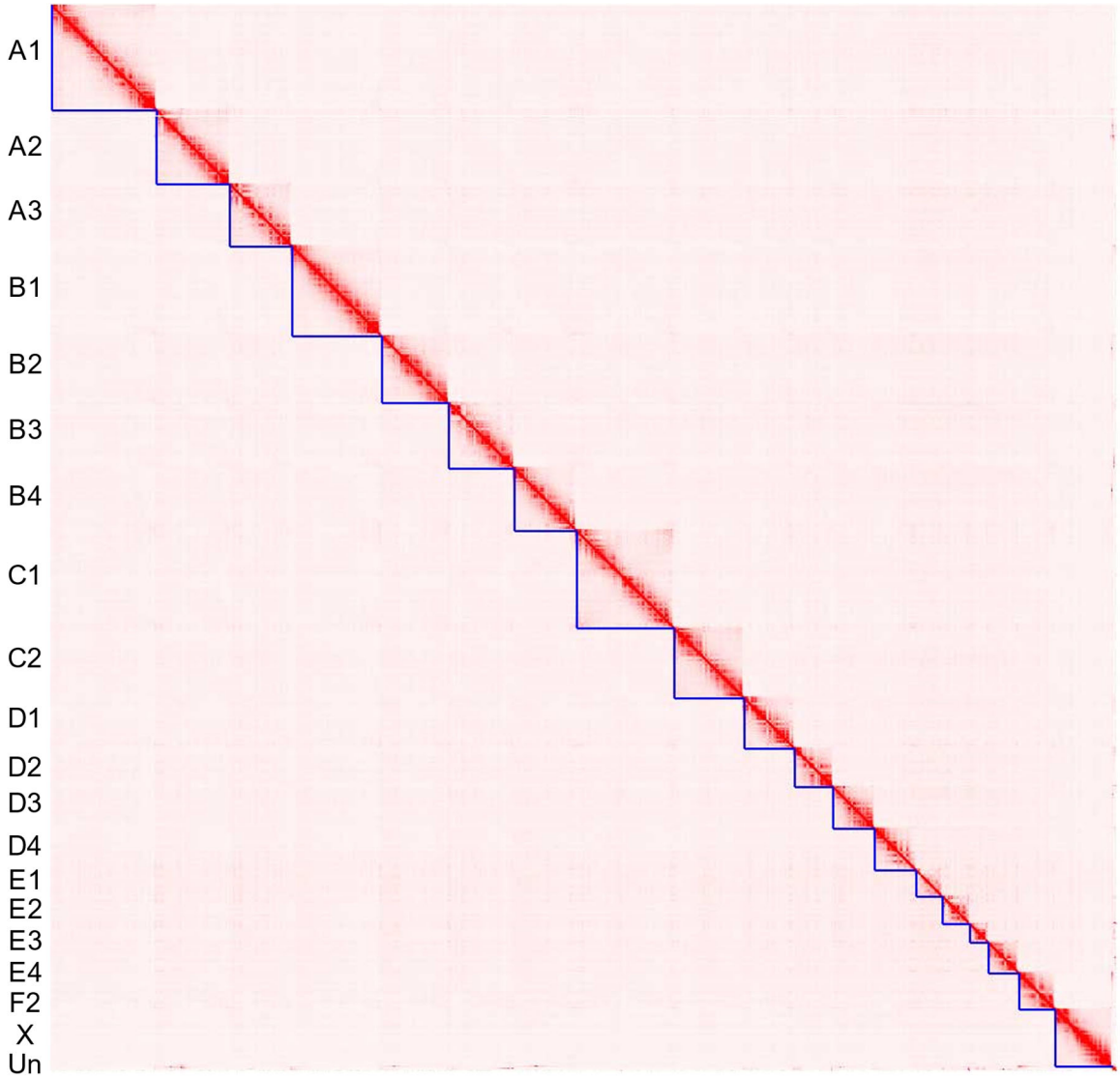


Figure B2.6. Nucmer alignments of single haplotype Asian leopard cat scaffolds to single haplotype domestic cat scaffolds.

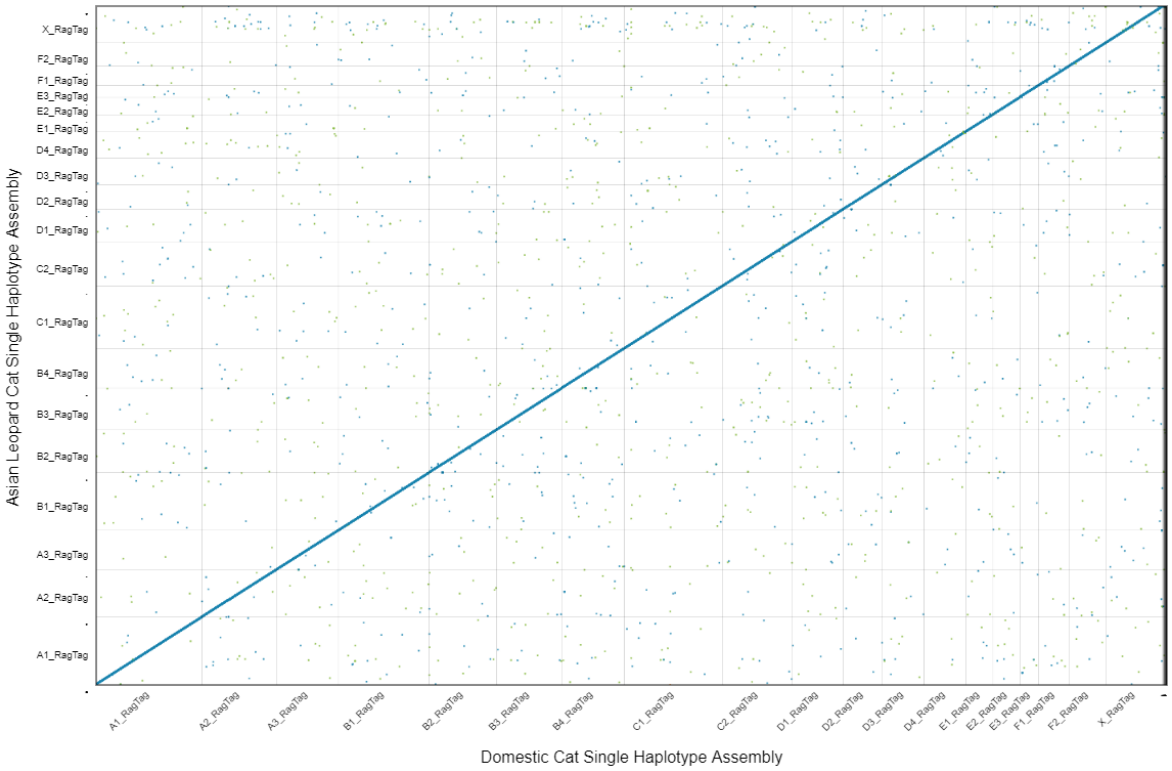


Figure B2.7. P-distance traces of the test samples from both species mapped to the Fca508 single haplotype reference assembly.

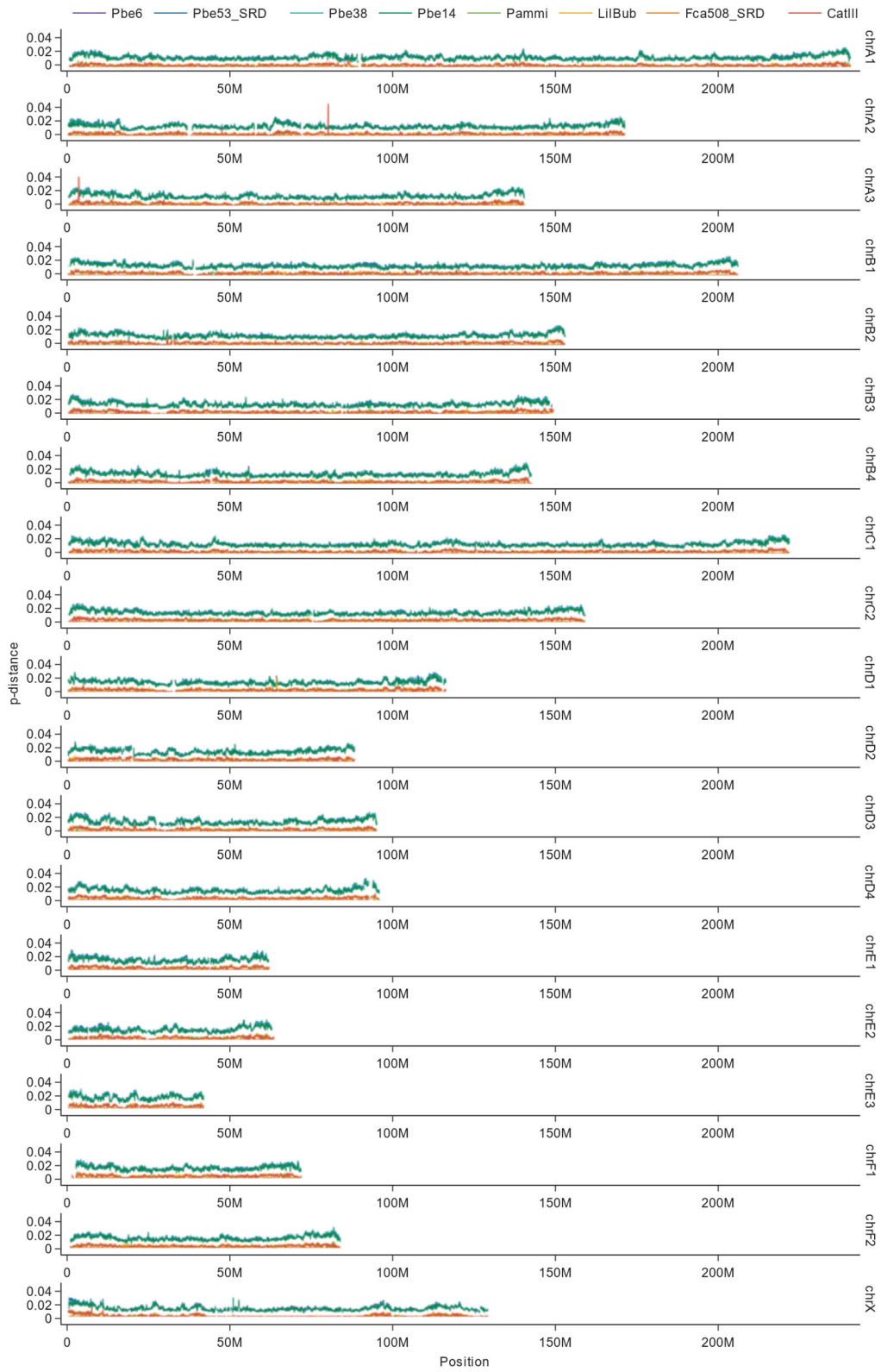


Figure B2.8. P-distance traces of the test samples from both species mapped to the Pbe53 single haplotype reference assembly.

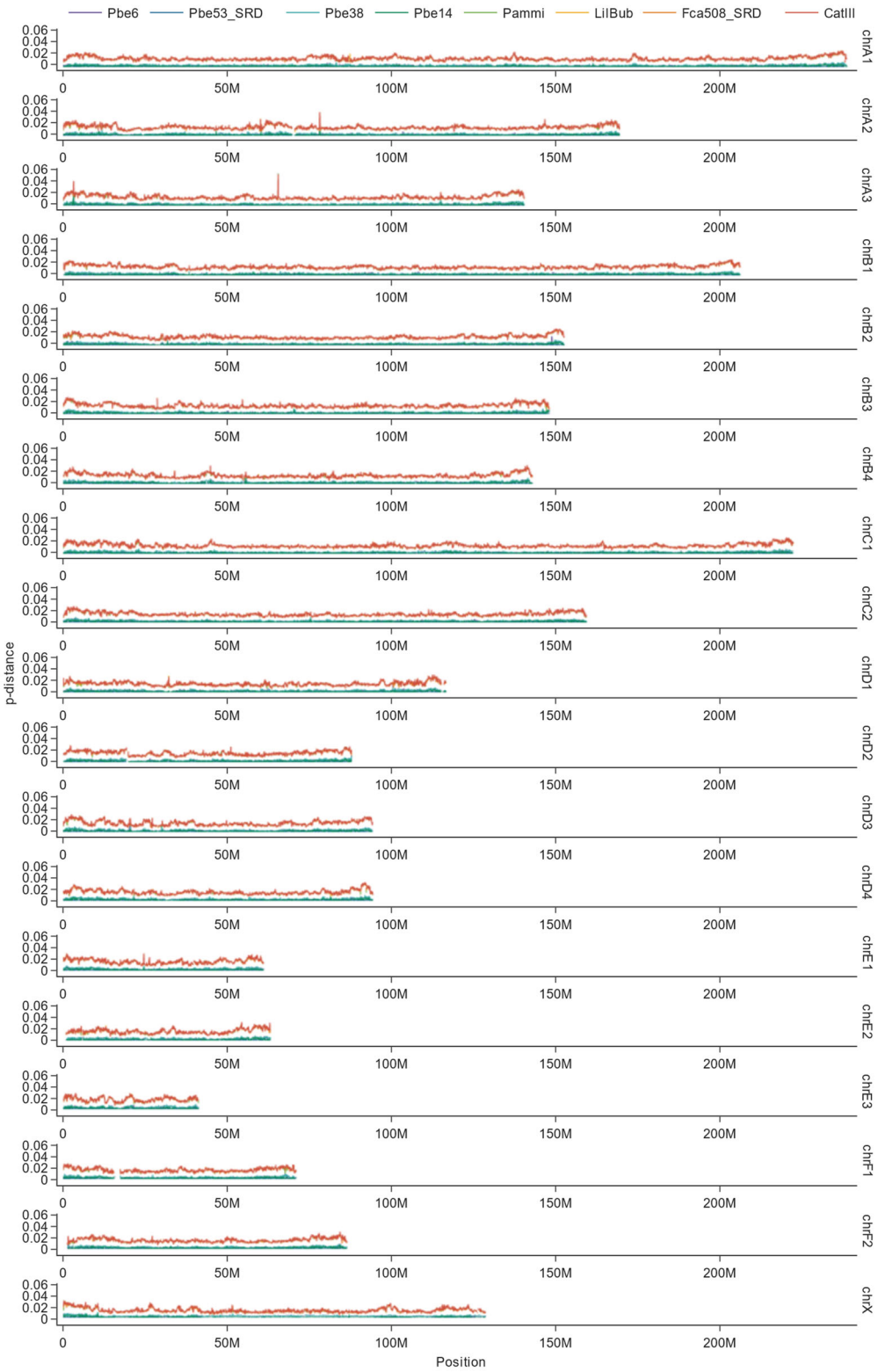


Figure B2.9. Read length distribution for two replacement crosses separated into correctly sorted reads and six incorrectly sorted subtypes. 79.83% (LilBubxPbe53) and 80.29% (Fca508xPbe14) of the incorrectly sorted reads are less than 10-kb in length, and 56.99% (LilBubxPbe53) and 51.83% (Fca508xPbe14) of the incorrectly sorted reads are less than 5-kb in length.

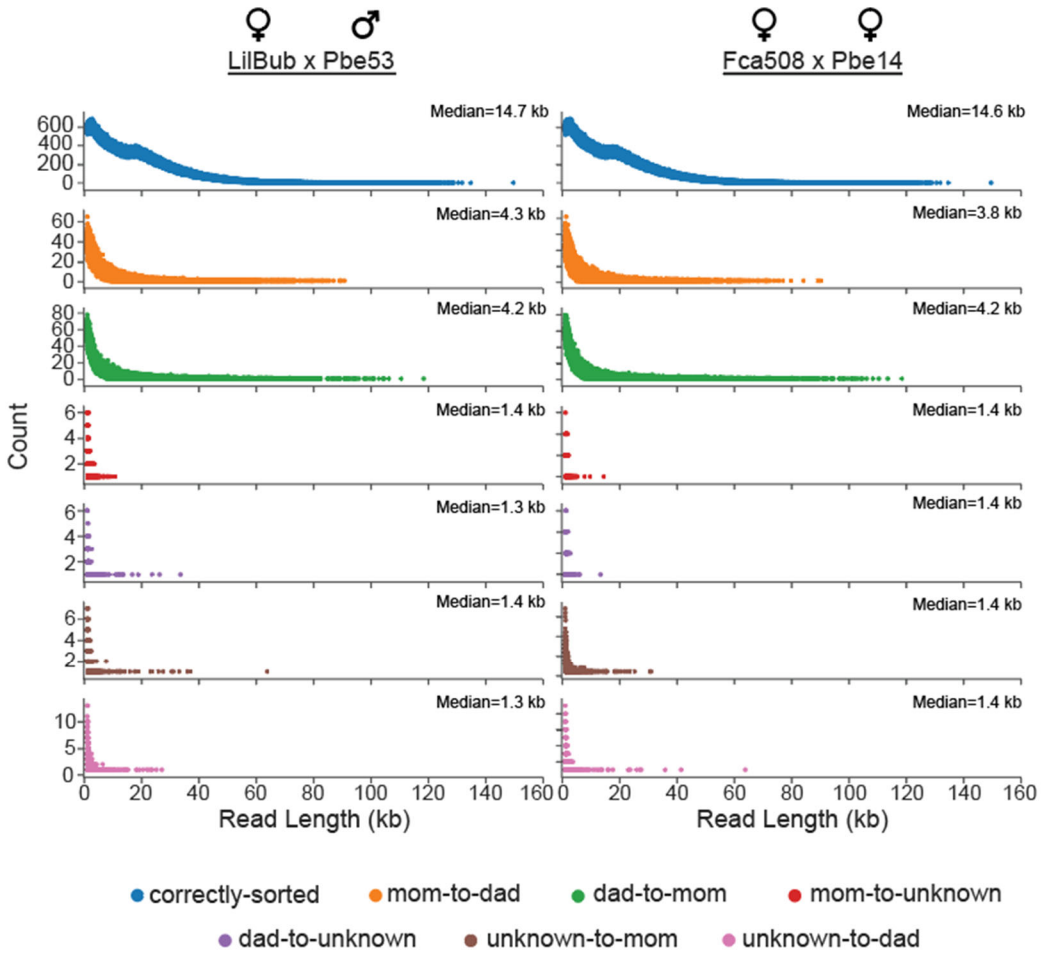
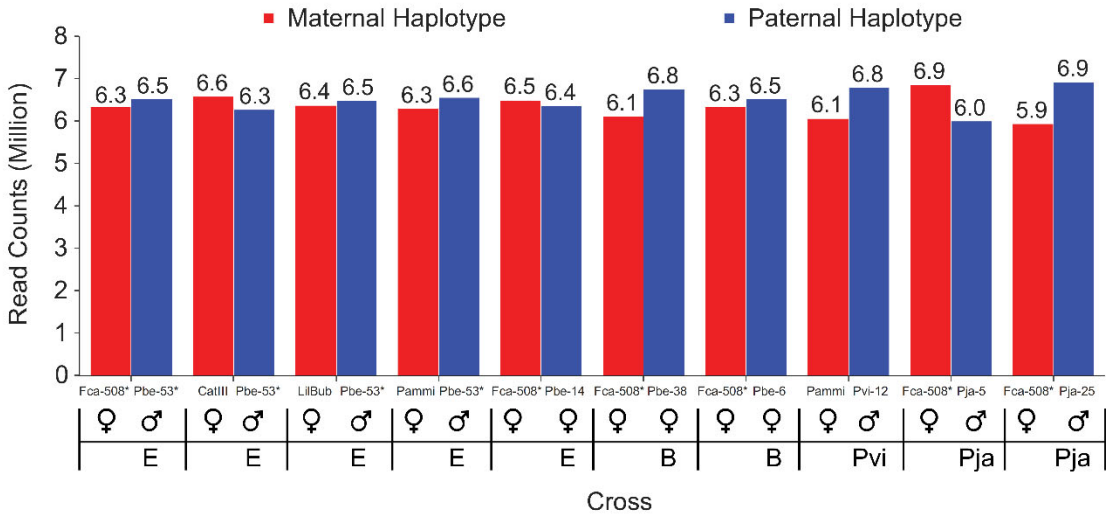


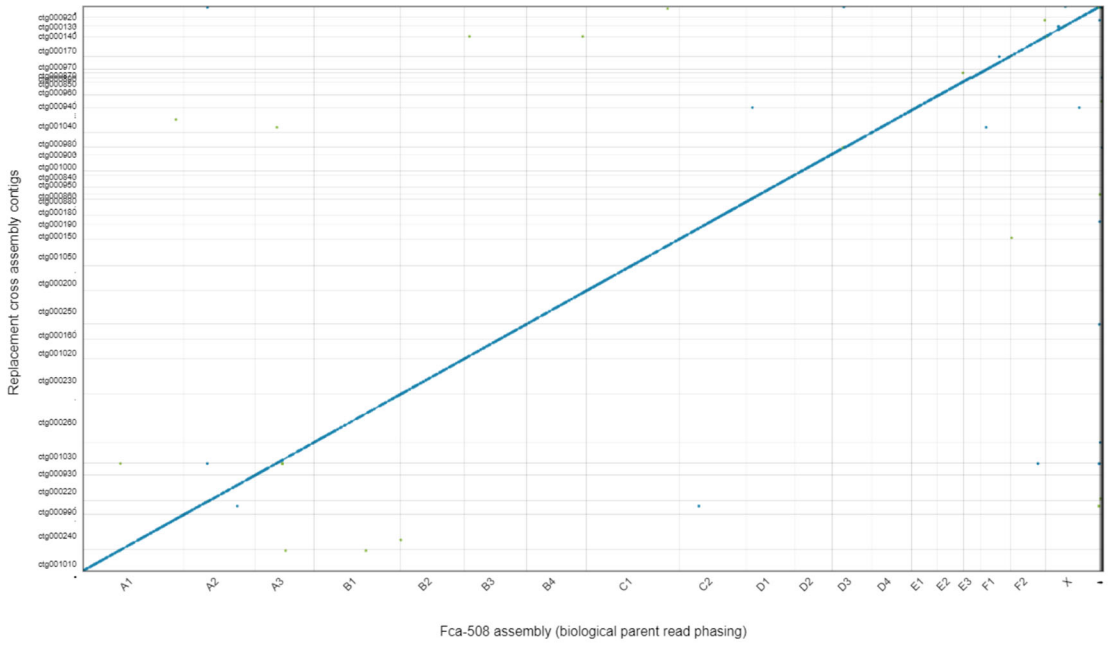
Figure B2.10. Maternal and Paternal haplotype read counts for single biological replacement crosses. We observe a tendency to skew read phasing toward one parental haplotype or the other when using increasingly divergent non-biological replacement parent samples (Pja5, Pja25, Pvi12). (*) indicates biological samples, (E)= *Prionailurus bengalensis euptilurus*, (B)= *Prionailurus bengalensis bengalensis*, (Pvi)= *Prionailurus viverrinus*, (Pja)= *Prionailurus javanenesis.*



Sample	Species/Subspecies
Pja-25	*Prionailurus javanensis*
Pja-5	*Prionailurus javanensis*
Pvi-12	*Prionailurus viverrinus*
Pbe-14	*Prionailurus bengalensis euptilurus*
Pbe-38	*Prionailurus bengalensis bengalensis*
Pbe-6	*Prionailurus bengalensis bengalensis*

Figure B2.11. Nucmer alignments comparing assemblies generated from reads sorted using data from biological parents and non-biological parent trios. In both alignments polished contigs generated from reads sorted using non-parental short read data are located along the Y axis, and the final chromosome length assembly generated from reads sorted using the biological parent's short read data is located along the X axis. **a)** Alignment of domestic cat non-biological contigs to biological assembly. **b)** Alignment of Asian leopard cat non-biological contigs to biological assembly.

a)



b)

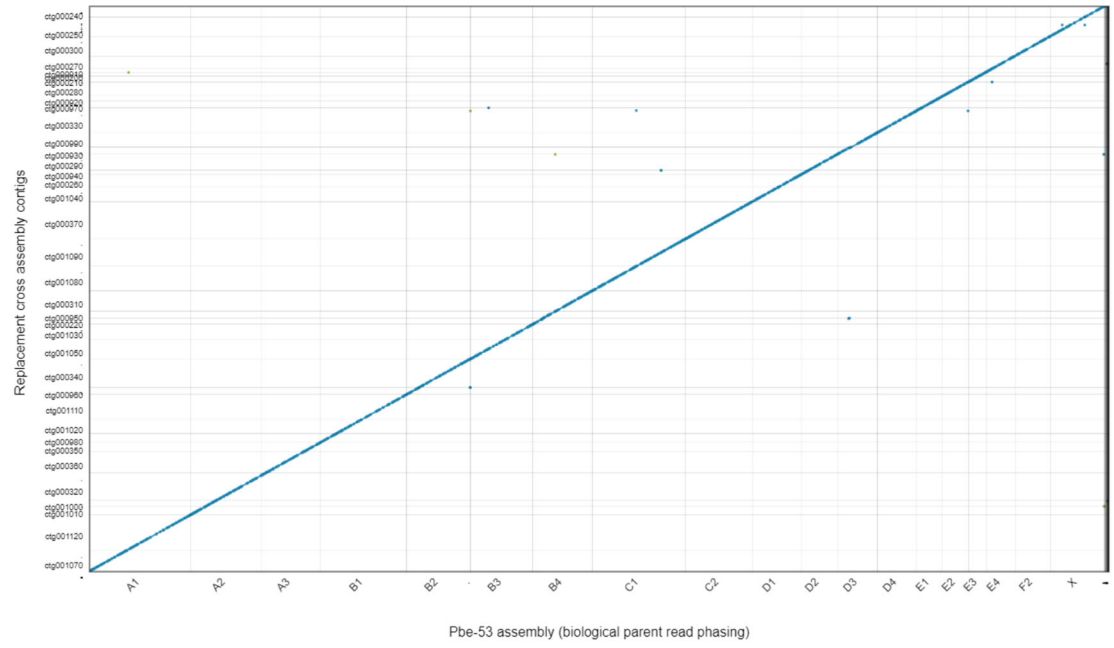


Figure B3.1. Per chromosome log-fold change (logFC) of genes differentially expressed in the testes of sterile backcross Chausie hybrids relative to the testes of fertile Chausie hybrids. All autosomes show a similar logFC profile while the X chromosome exhibits increased upregulation (logFC= +2.45).

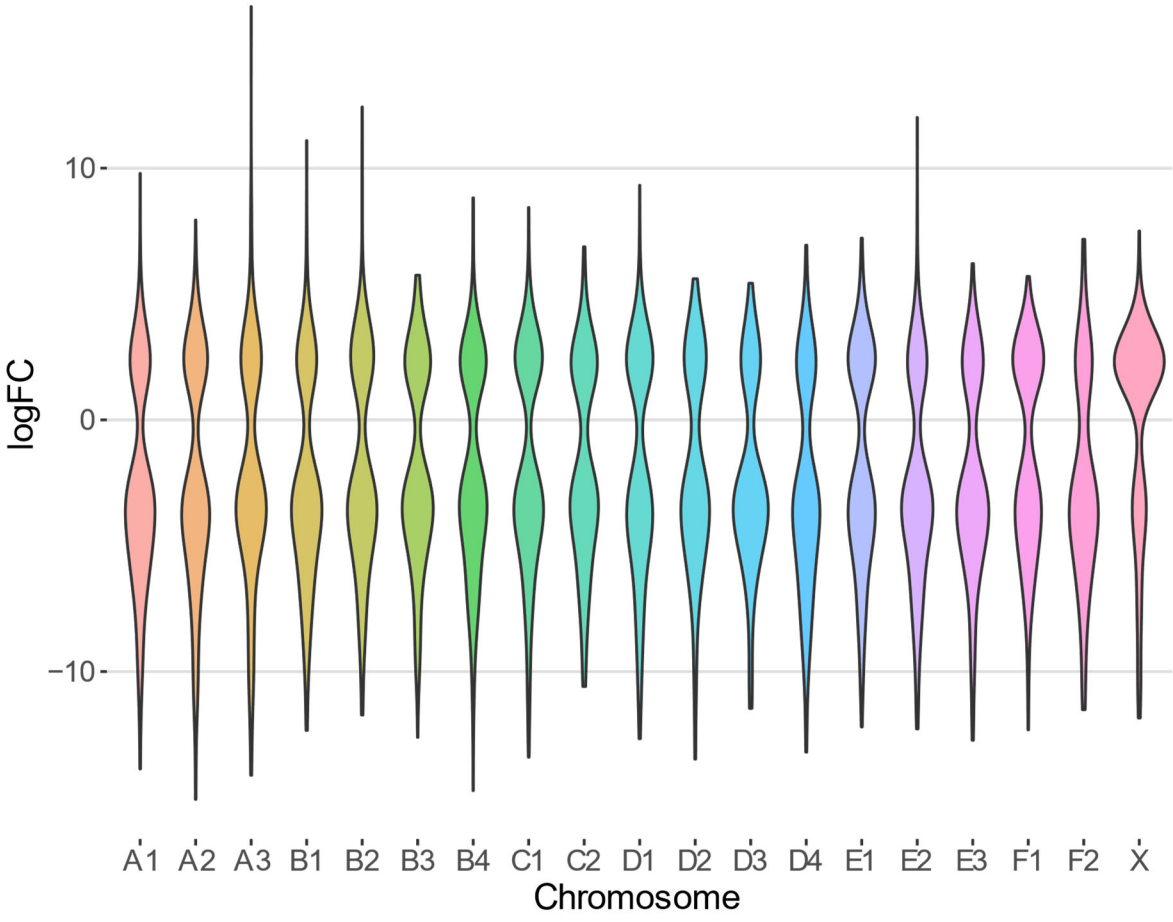


Figure B3.2. RNA-seq data tracks for *PLS3* in the NCBI Genome Browser for domestic cat reference assembly felCat9. The x-axis reflects coordinates across the X Chromosome. Expression tracks for various tissue types are arrayed across the y-axis and are log-scaled. *PLS3* is ubiquitously expressed across tissue types in the domestic cat.

Figure B3.3. Mauve alignment of the *DXZ4* region from three previous domestic cat reference genome assemblies. Assemblies are shown in chronological order with 6.2 being the earliest iteration. Loss of *DXZ4* region in subsequent assemblies are indicated by shortening of the light green arrows. Colored blocks represent stretches of contiguous sequence between the three assemblies.

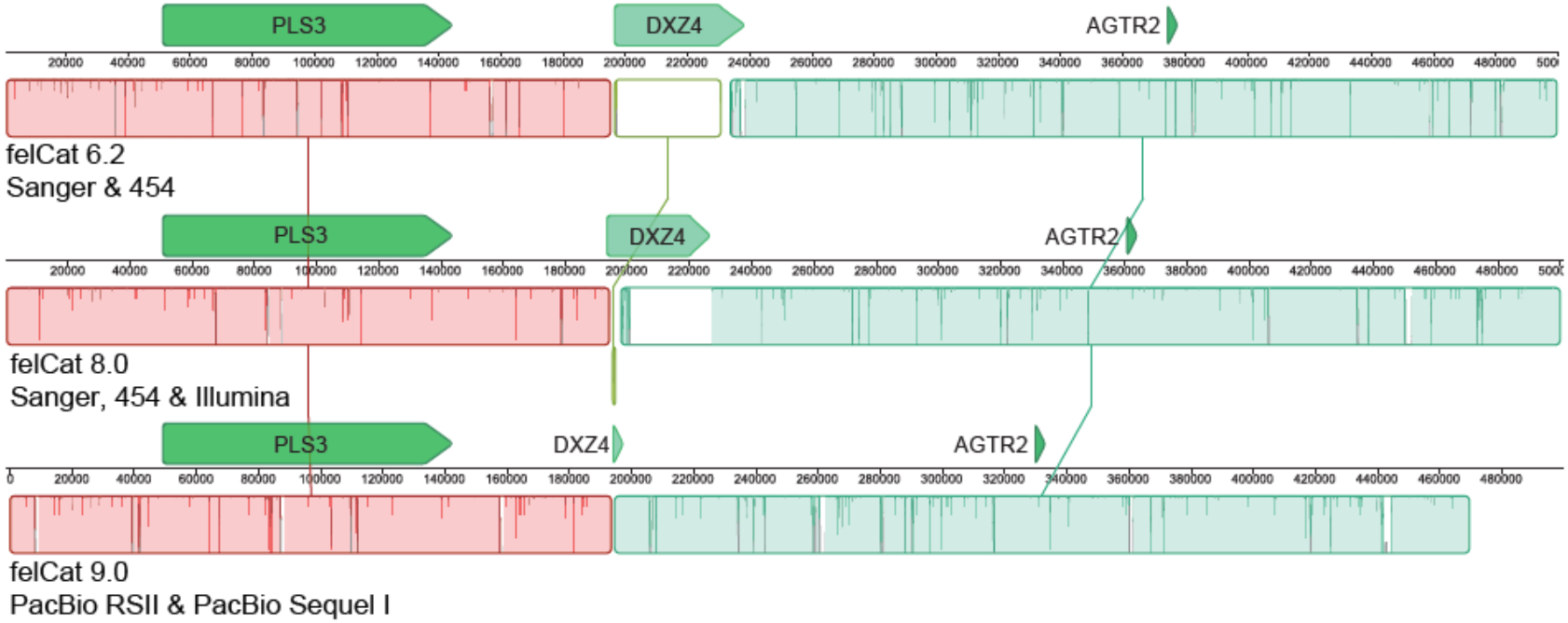


Figure B3.4. Dotplots between three cat single haplotype X chromosome assemblies (y-axis) and the diploid reference, felCat9 (x-axis) reveal *DXZ4* is captured within a single contig. Sequence gain was estimated from the alignment shift within each contig (y-axis), relative to the reference.

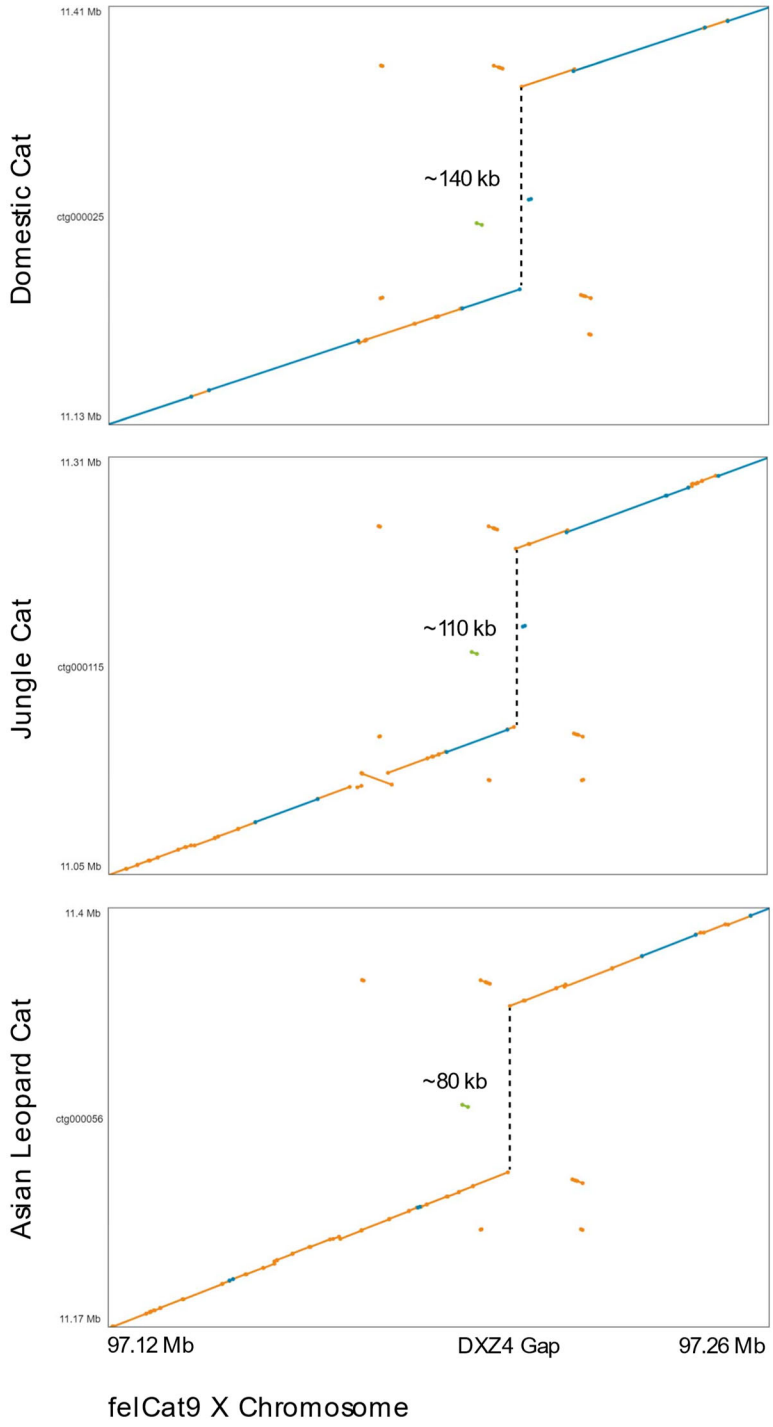


Figure B3.5. Annotated DXZ4 regions from domestic, Jungle, and Asian leopard cat assemblies.

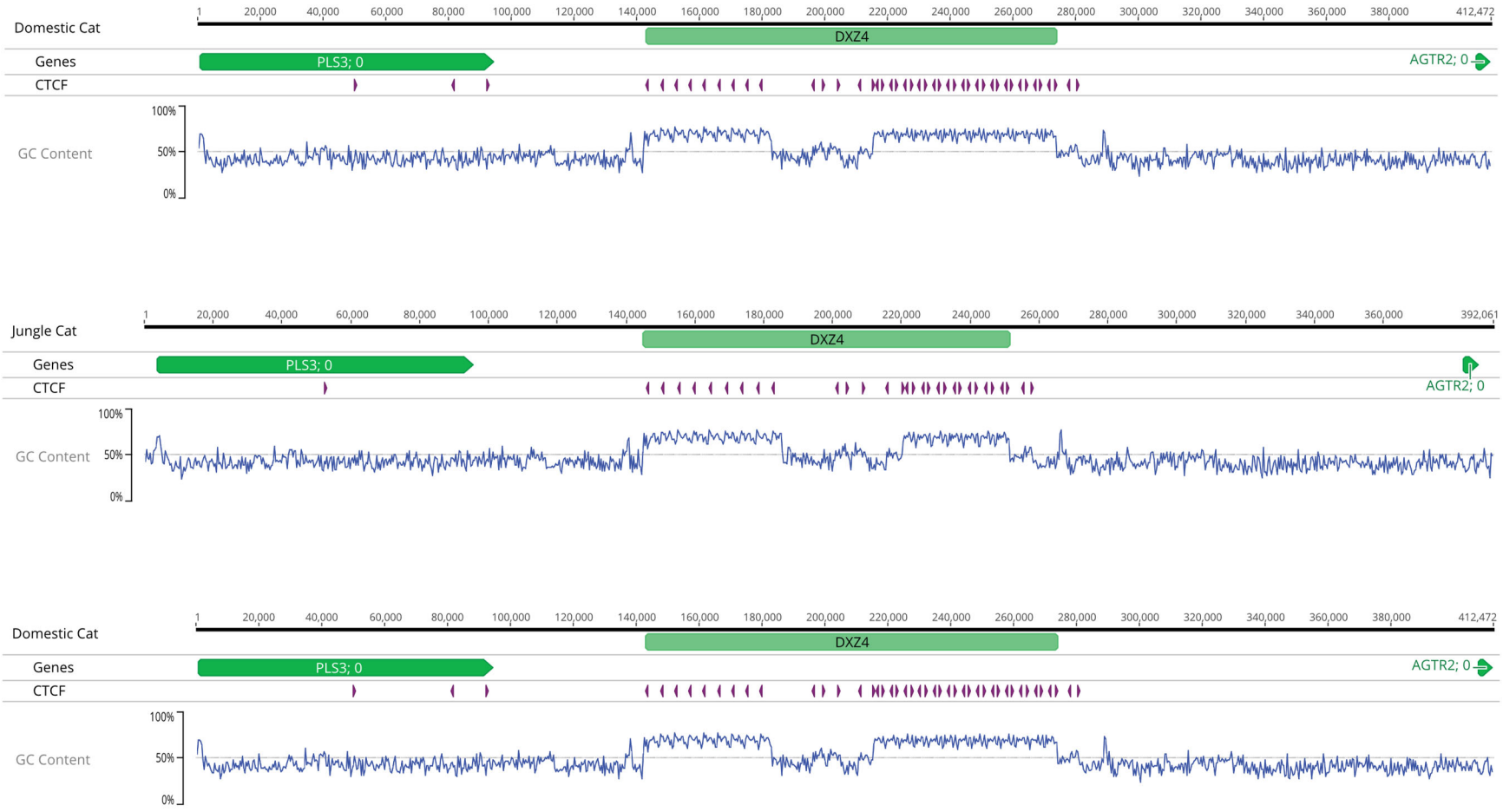


Figure B3.6. Felid spacer (sequence between DXZ4 RA and RB arrays) sequence alignment. Alignment gaps distinguished by grey annotations.

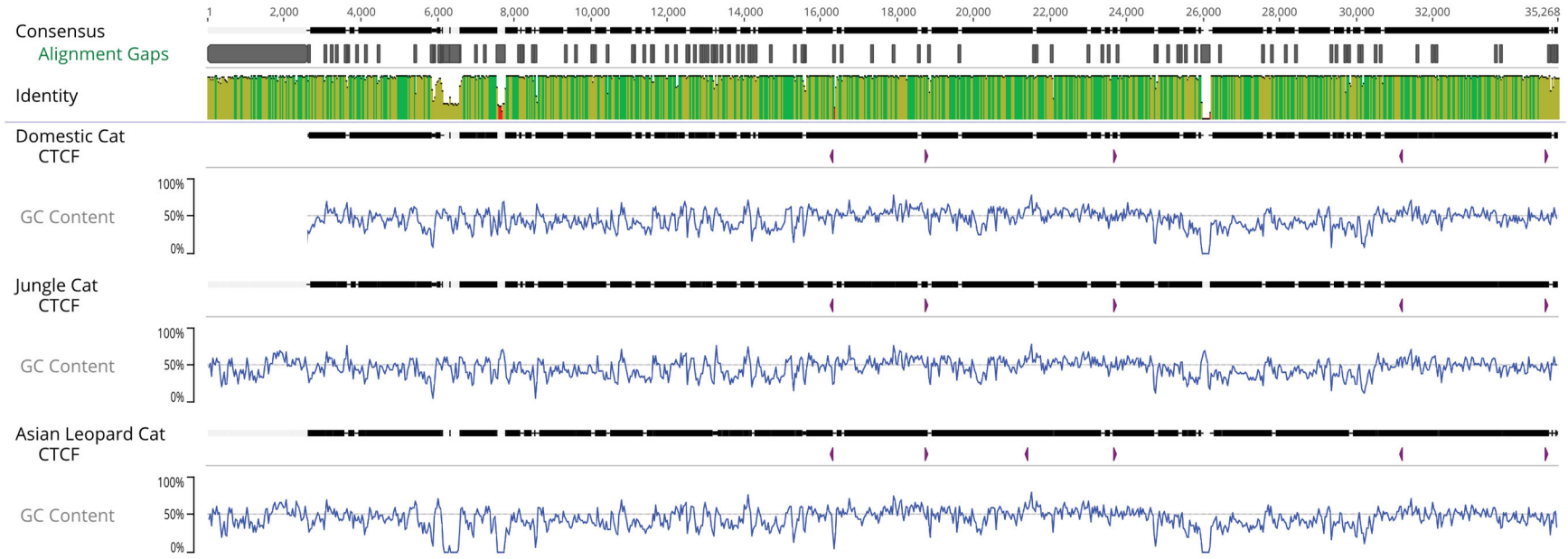


Figure B3.7. *DXZ4* repeat modified for in silico copy number estimation using short read mapping. Unmodified *DXZ4* repeat array (Top). *DXZ4* repeat array modified for *in silico* read mapping (Bottom).

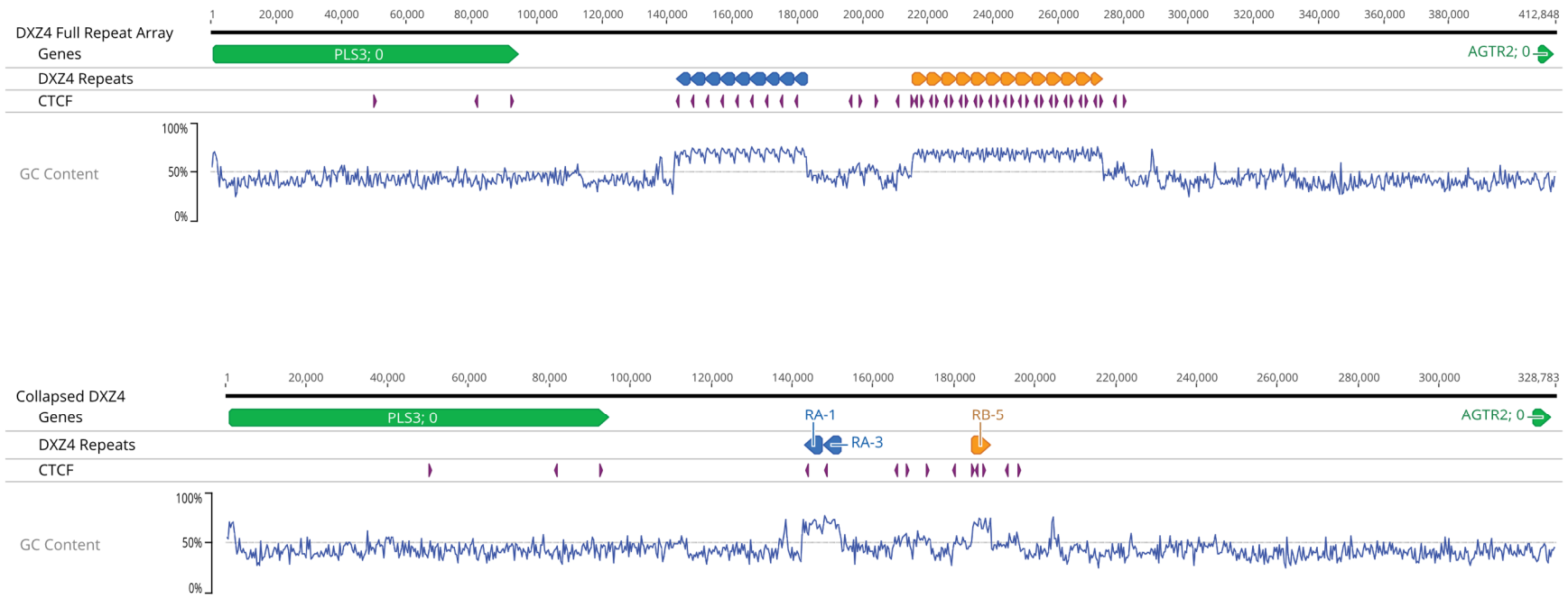


Figure B3.8. Comparison between the domestic cat and human *DXZ4* annotations. Human assembly GRCh38.p13 is shown because the human T2T assembly (Miga et al, 2020) lacks *de novo* annotation.

Figure B3.9. Human *DANT1/2* pairwise alignments to domestic cat *CUFF1.1/5.1* (RA/RB spanning transcripts). Green bars represent regions of shared sequence identity with colored bars below reflecting different nucleotides.

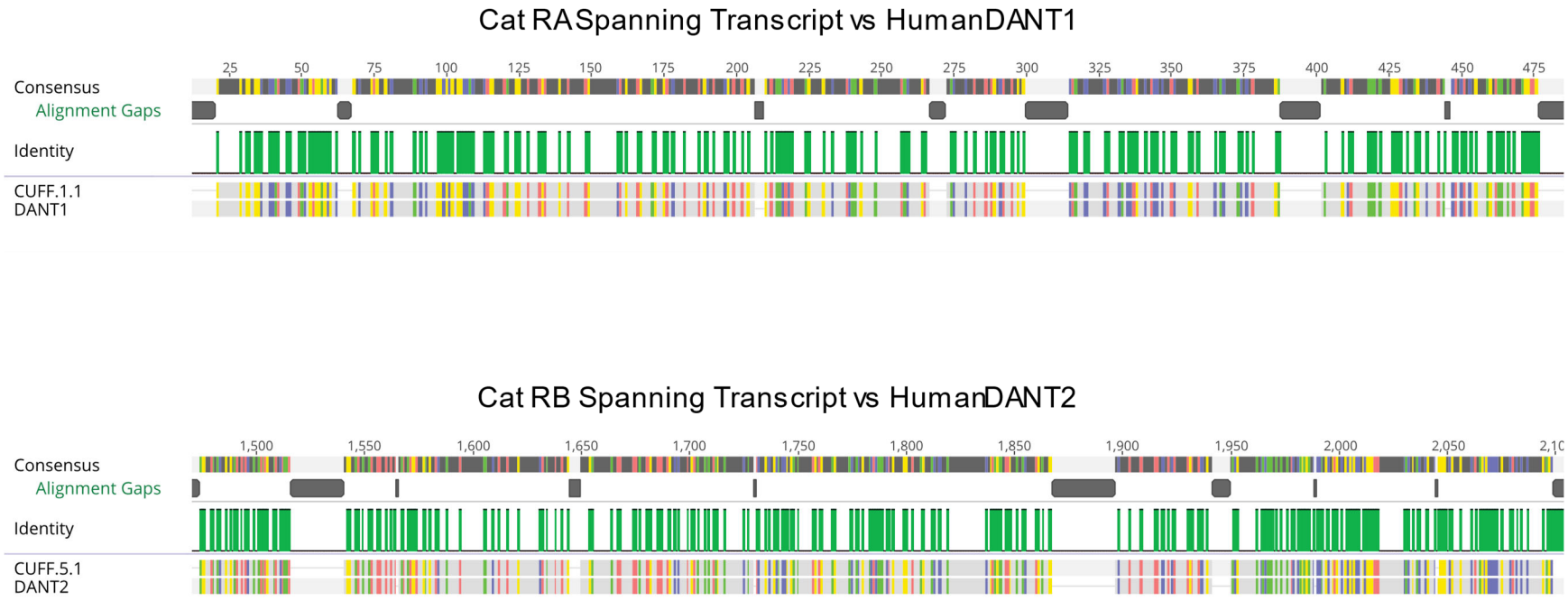


Figure B3.10. Comparison between annotated *DXZ4* transcripts in domestic and Jungle cat from whole testes RNA-seq data.

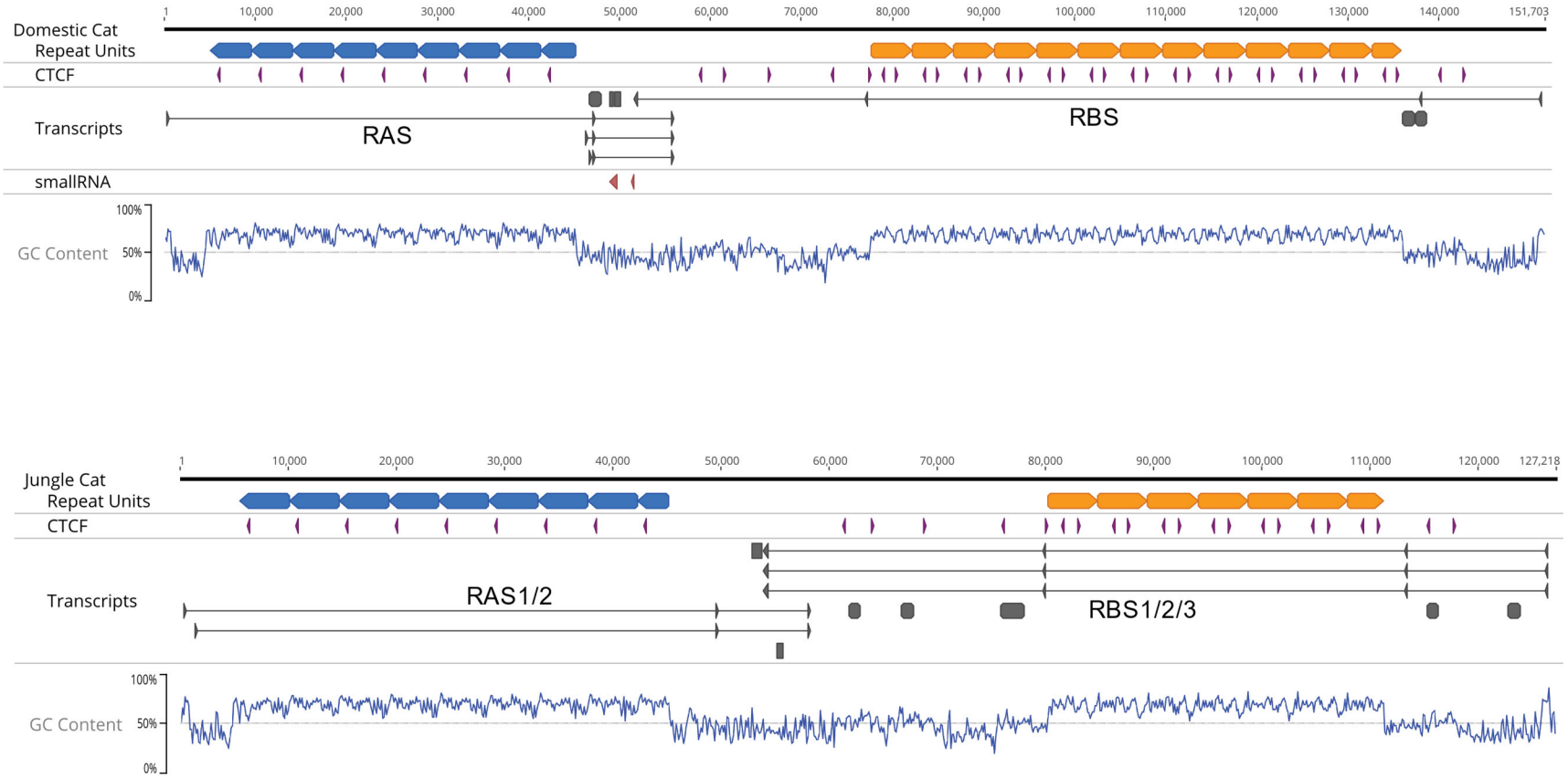


Figure B3.11. Alignment between domestic and Jungle cat *DXZ4* regions. Alignment gaps are indicated by light-gray annotations.

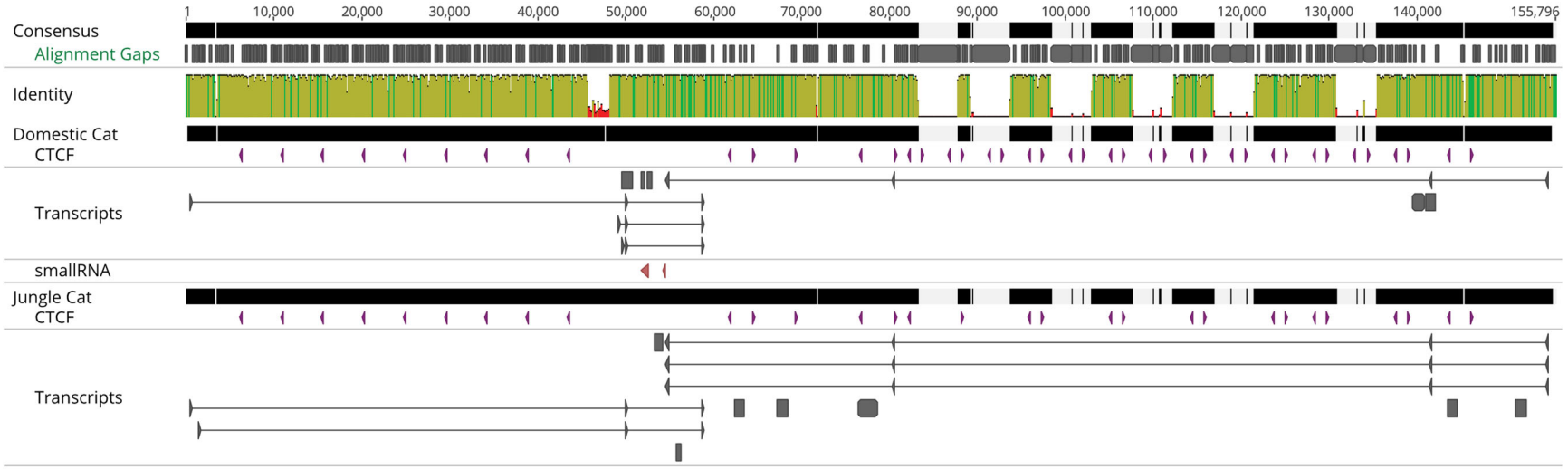


Figure B3.12. allRNA and smallRNA expression across the *DXZ4* locus in domestic and Jungle cat. Y-axis scaled to 80x coverage for allRNA and 10x coverage for smallRNA in both species.

Figure B3.13. Expression across first repeat unit of *DXZ4* Repeat B in both domestic and Jungle cats.

Figure B3.14. RNA-seq read coverage across the *DXZ4* region for fertile and sterile Chausies. Y-axis represents raw expression scaled by highest coverage across all tracks.

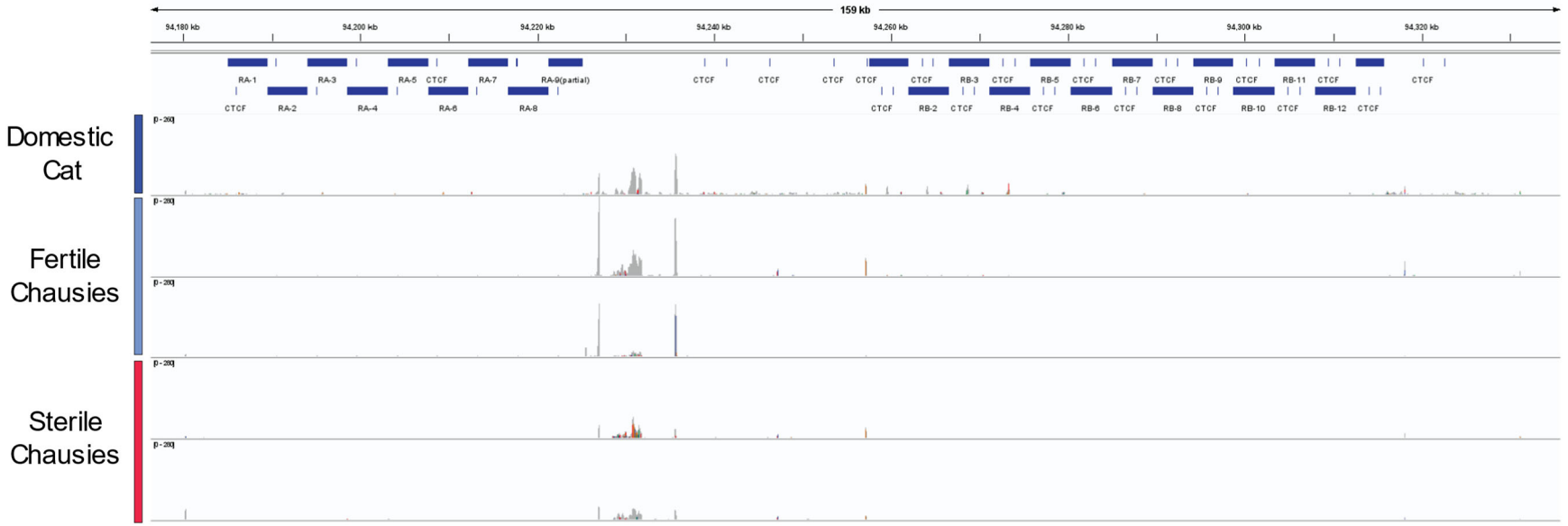


Figure B3.15. PCA of methylation frequency between fertility phenotypes.

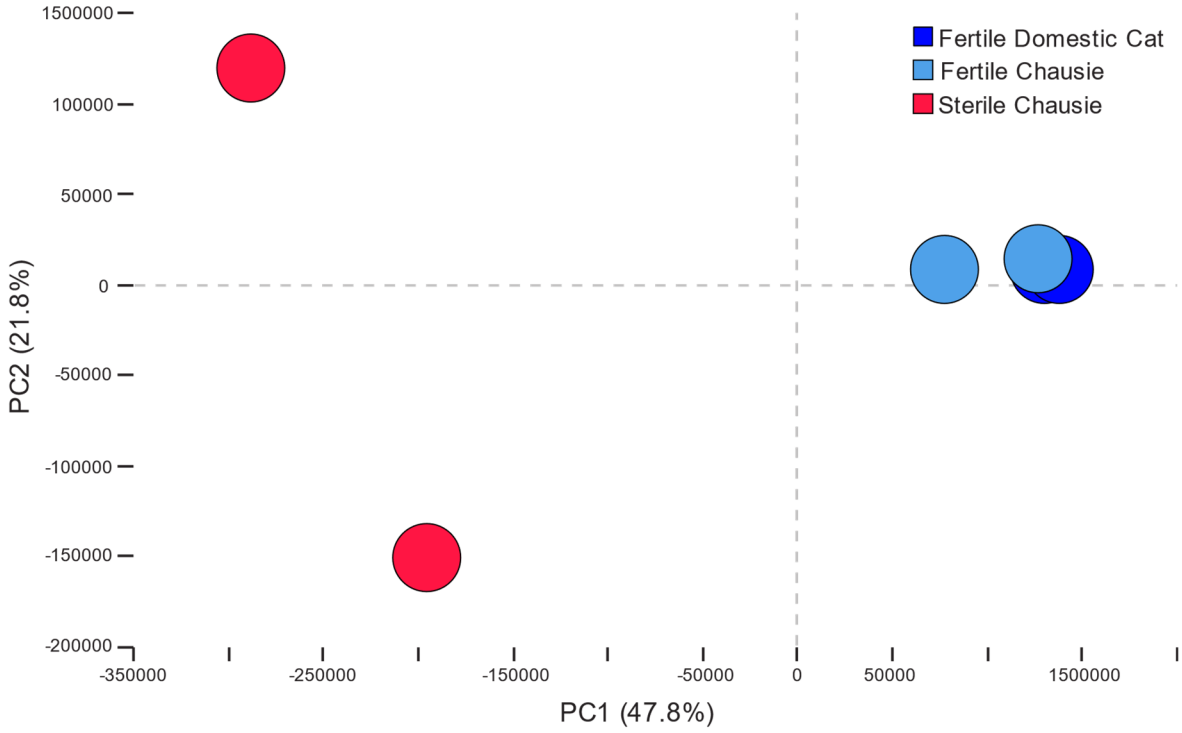


Figure B3.16. Methylation frequency averages for each chromosome per each fertile or sterile individual felid. The X chromosome is indicated by an "X".

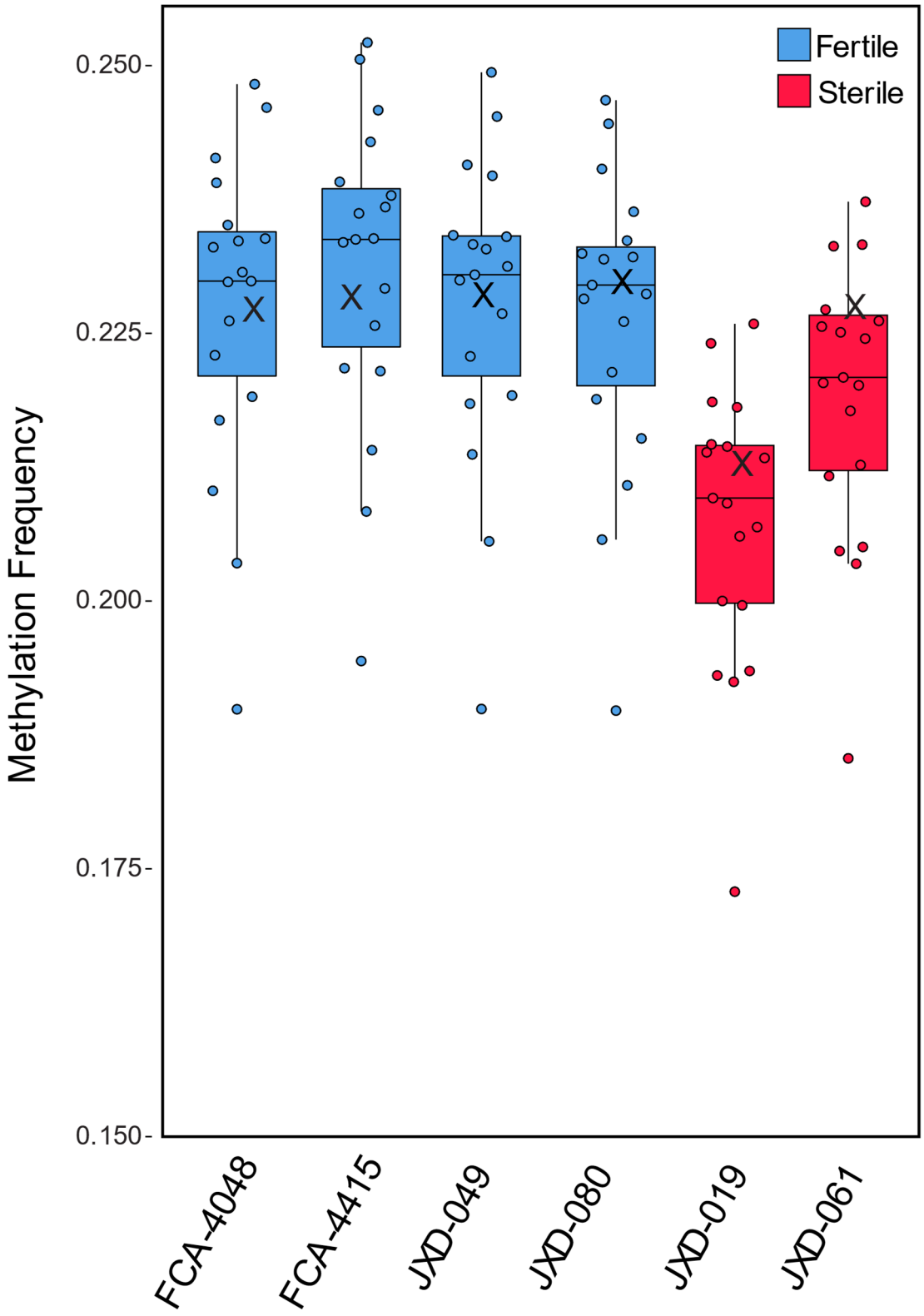


Figure B3.17. Bipartite structural conformation formed by *DXZ4* is conserved on the inactive X of females in cat, human and mouse. Cat Hi-C data from domestic haplotype of F1 Bengal (Bredemeyer et al., 2020). Human Hi-C data from GM12878 cell line (Rao et al., 2014). Mouse Hi-C data from patski cell line (Darrow et al., 2016). Resolution 250kb with "balanced" normalization for all maps.

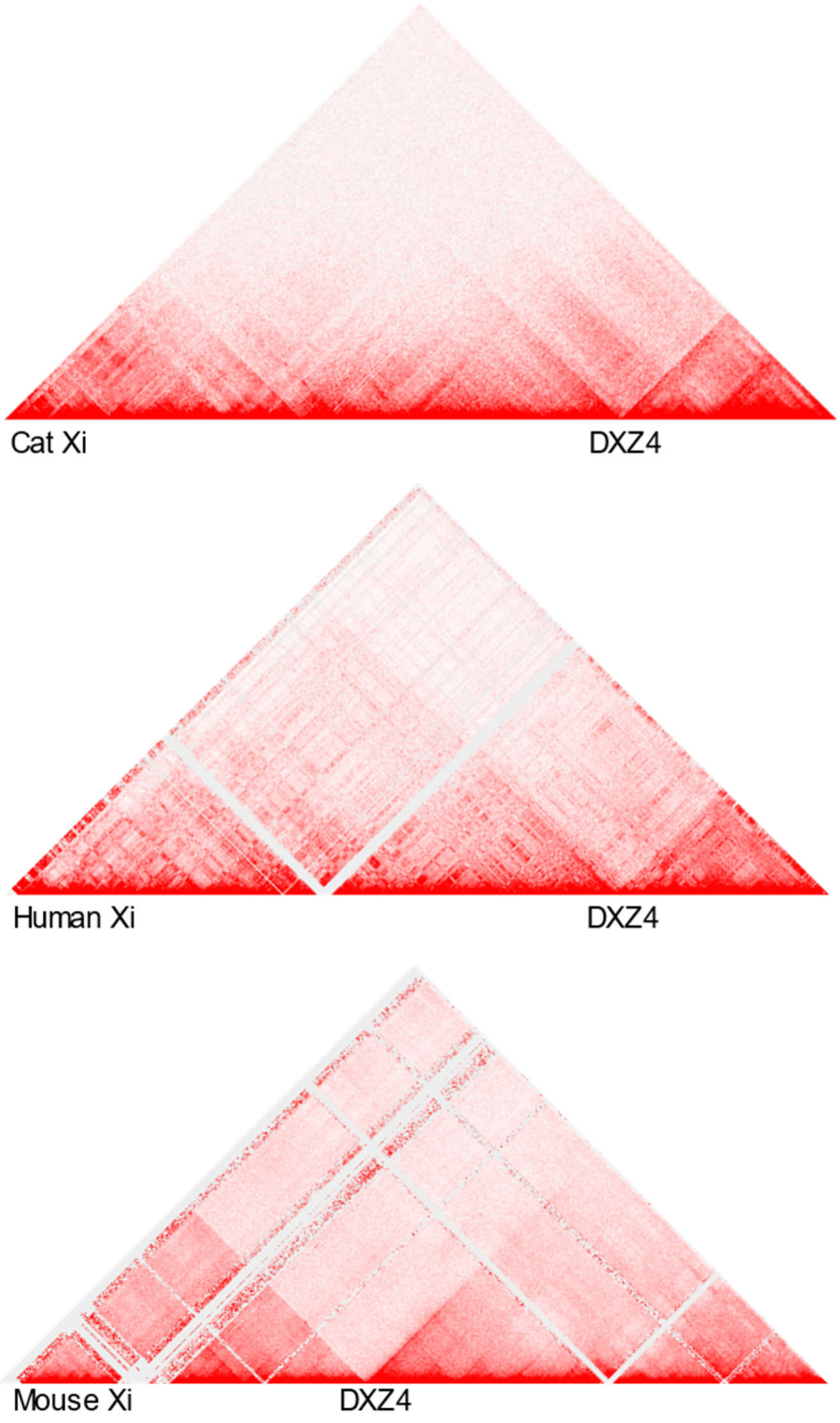


Figure B3.18. Raw contact maps and Pearson's correlation maps of the X chromosome from Hi-C data generated and phased from female fibroblasts of an F1 Bengal. The domestic cat haplotype exhibits distinct features of the Xi while the Asian leopard cat haplotype resembles the Xa, suggesting potential skewing of XCI in the domestic cat haplotype.

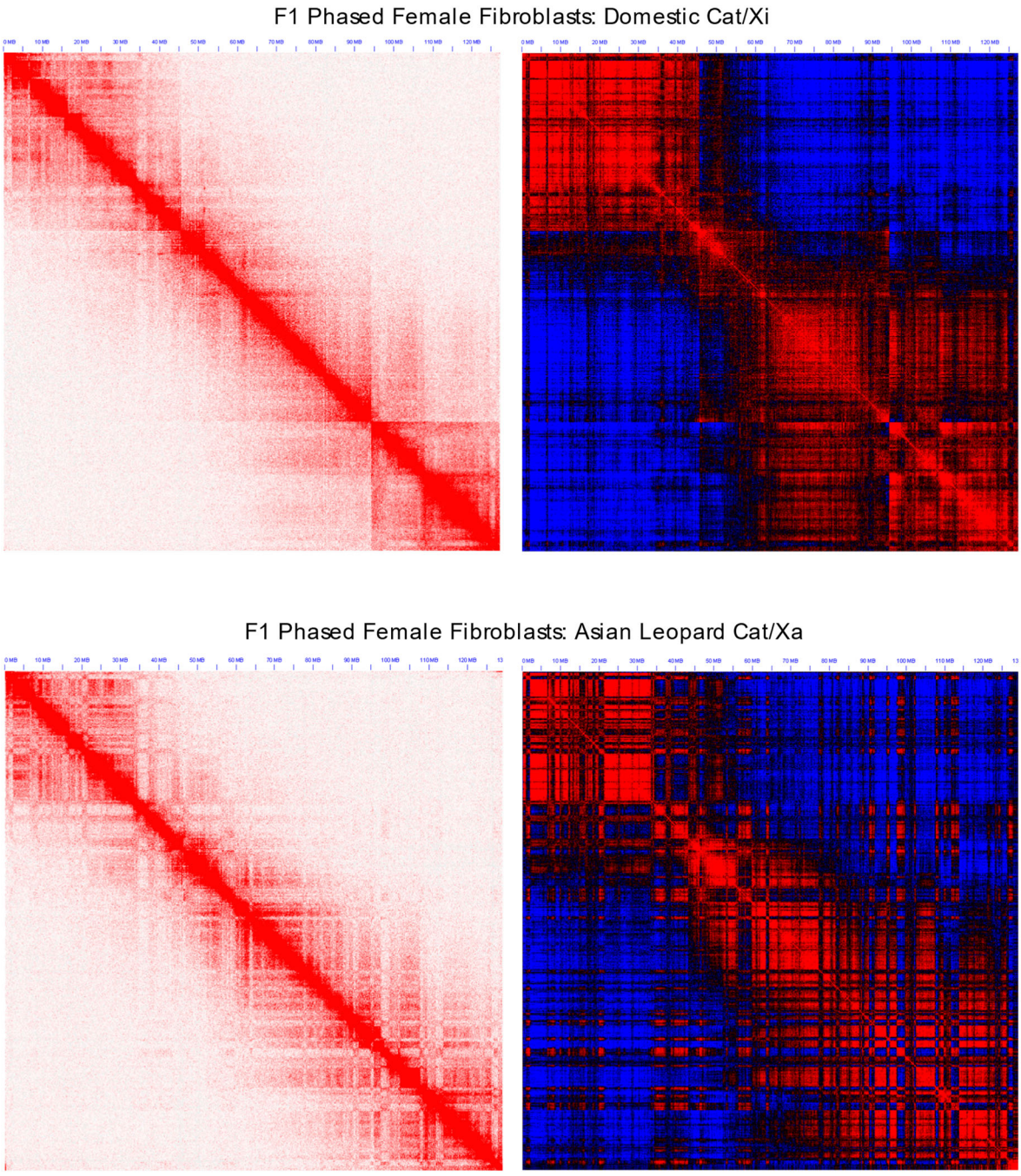


Figure B3.19. Pearson's maps showing differences in compartmentalization between Xa and Xi states of cat, human (Rao et al., 2014) and mouse (Darrow et al., 2016). Cat Xa and Xi are X chromosomes from Asian leopard cat and domestic cat phased from an F1 Bengal. Resolution is 250 kb for cat and 500 kb for human and mouse.

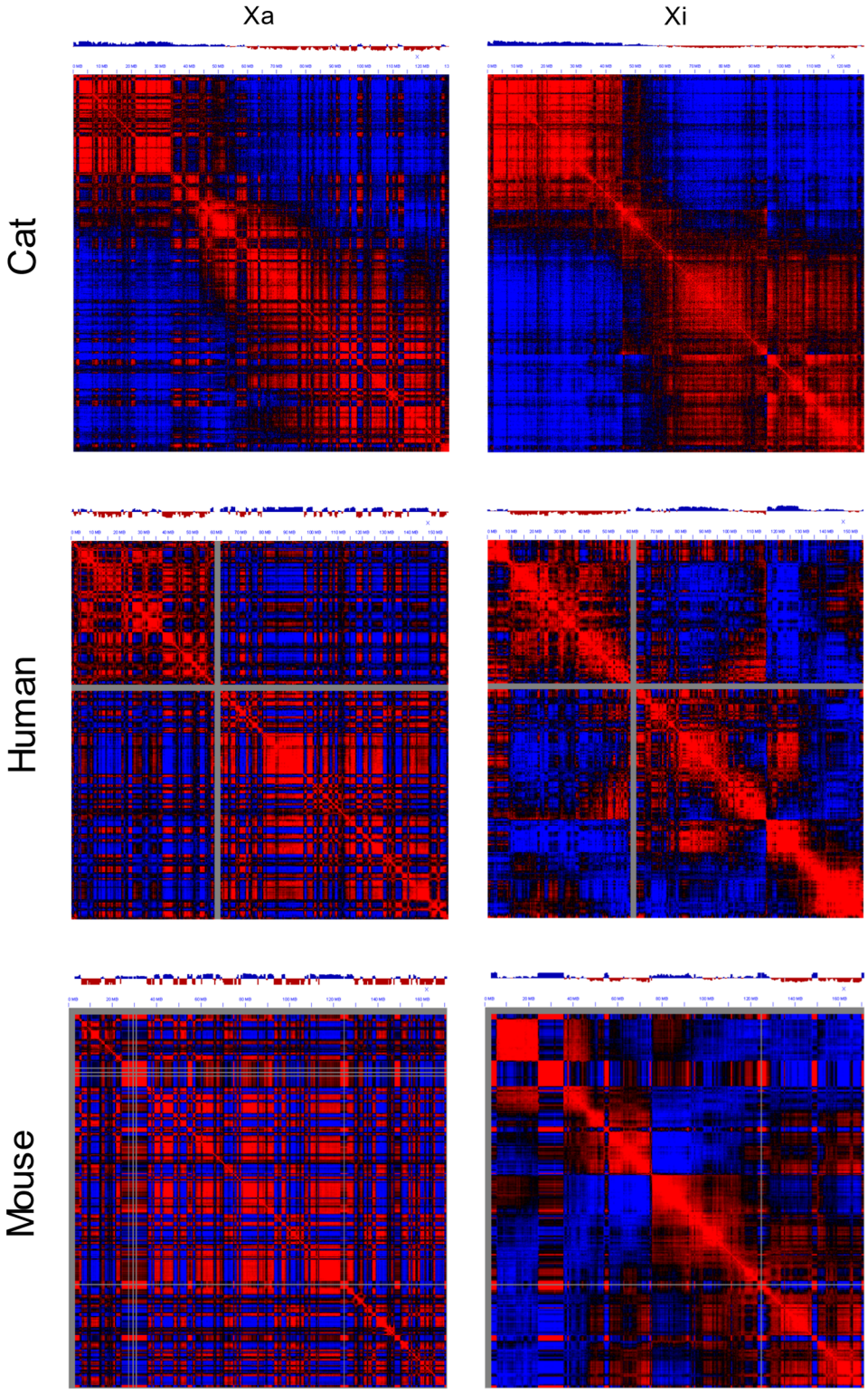


Figure B3.20. *DXZ4* structural comparison between human, mouse and cat.

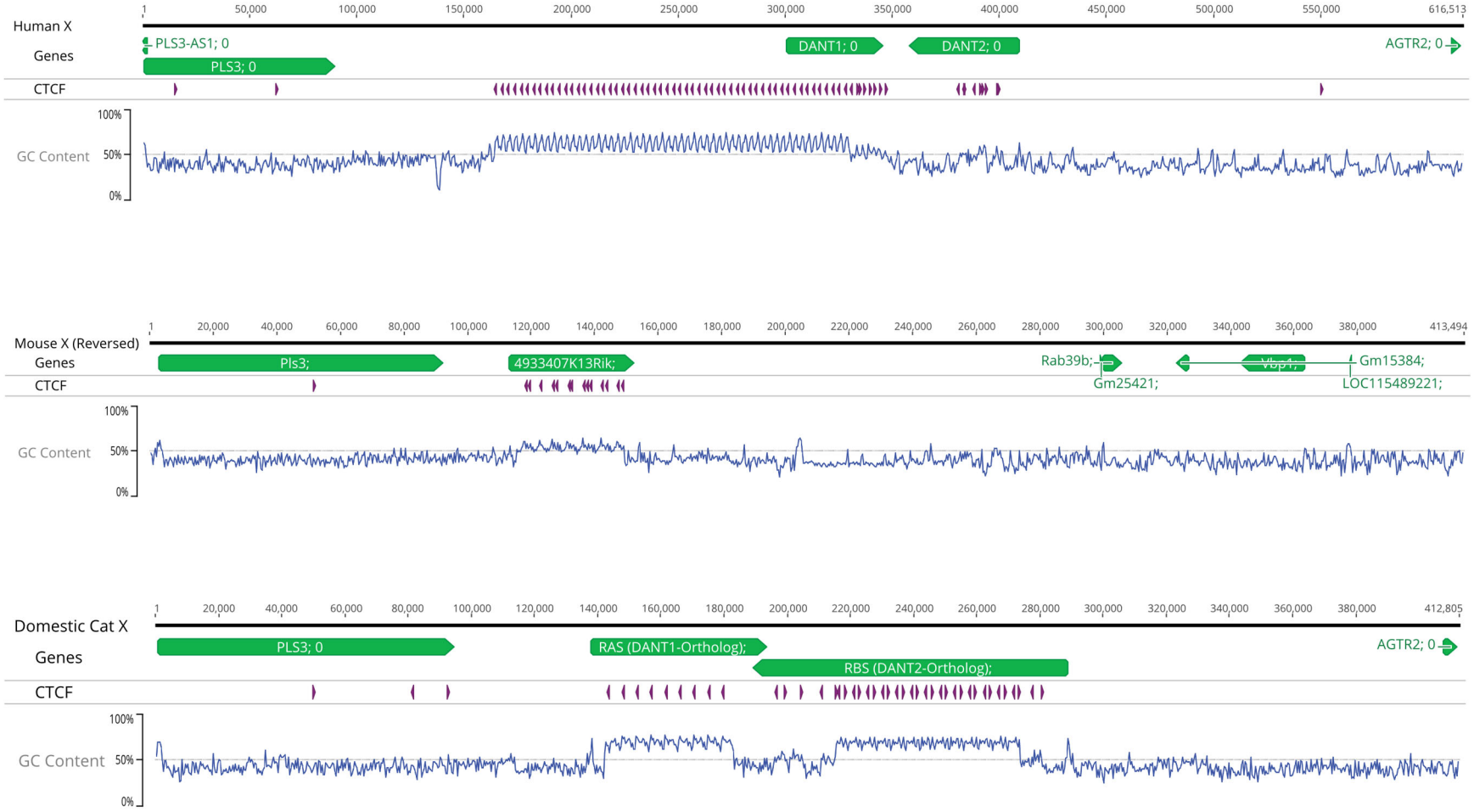
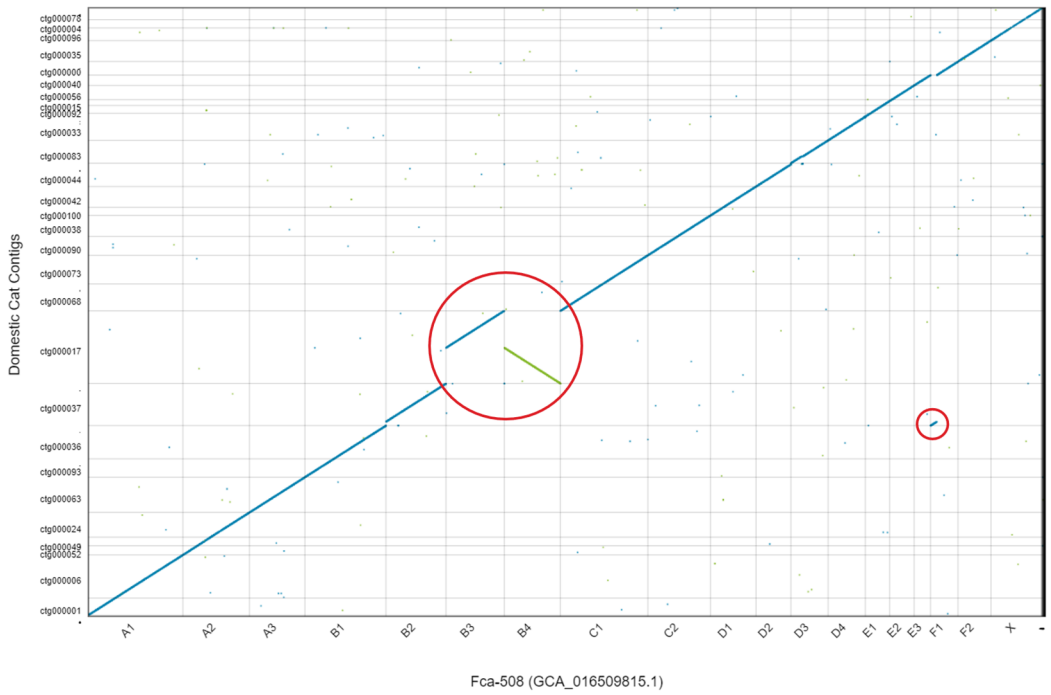


Figure B4.1. Contig alignments to domestic cat single haplotype assembly (GCA_016509815.1). **A)** Domestic Cat. Chimeric contigs ctg000017 and ctg000037 are representative of misjoins between chromosomes B3-B4 and chromosomes B2-F1, respectively. **B)** Geoffroy's Cat. Chimeric contig ctg000039 and ctg000000 are representative of misjoins between chromosomes B4-D2 and E1-F1, respectively. Chimeric contig ctg000031 representative of the F1-F2 chromosome fusion forming chromosome C3.

A)



B)

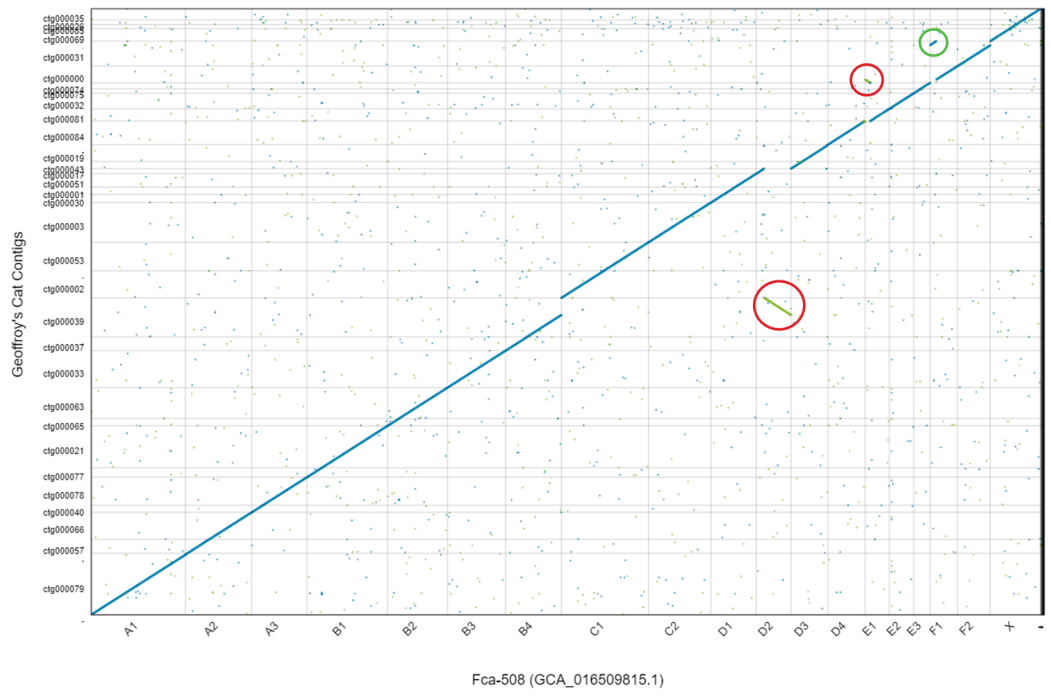


Figure B4.2. Contig alignments to domestic cat single haplotype assembly (GCA_016509815.1). **A)** Tiger. Chimeric contigs ctg000047 and ctg000023 are representative of misjoins between chromosomes B1-C2-E2 and chromosomes C1-E1, respectively. **B)** Lion. Chimeric contig ctg000079, ctg000062, ctg000043 and ctg000053 are representative of misjoins between chromosomes A3-B3, B1-B4, B4-D3 and F1-D4, respectively.

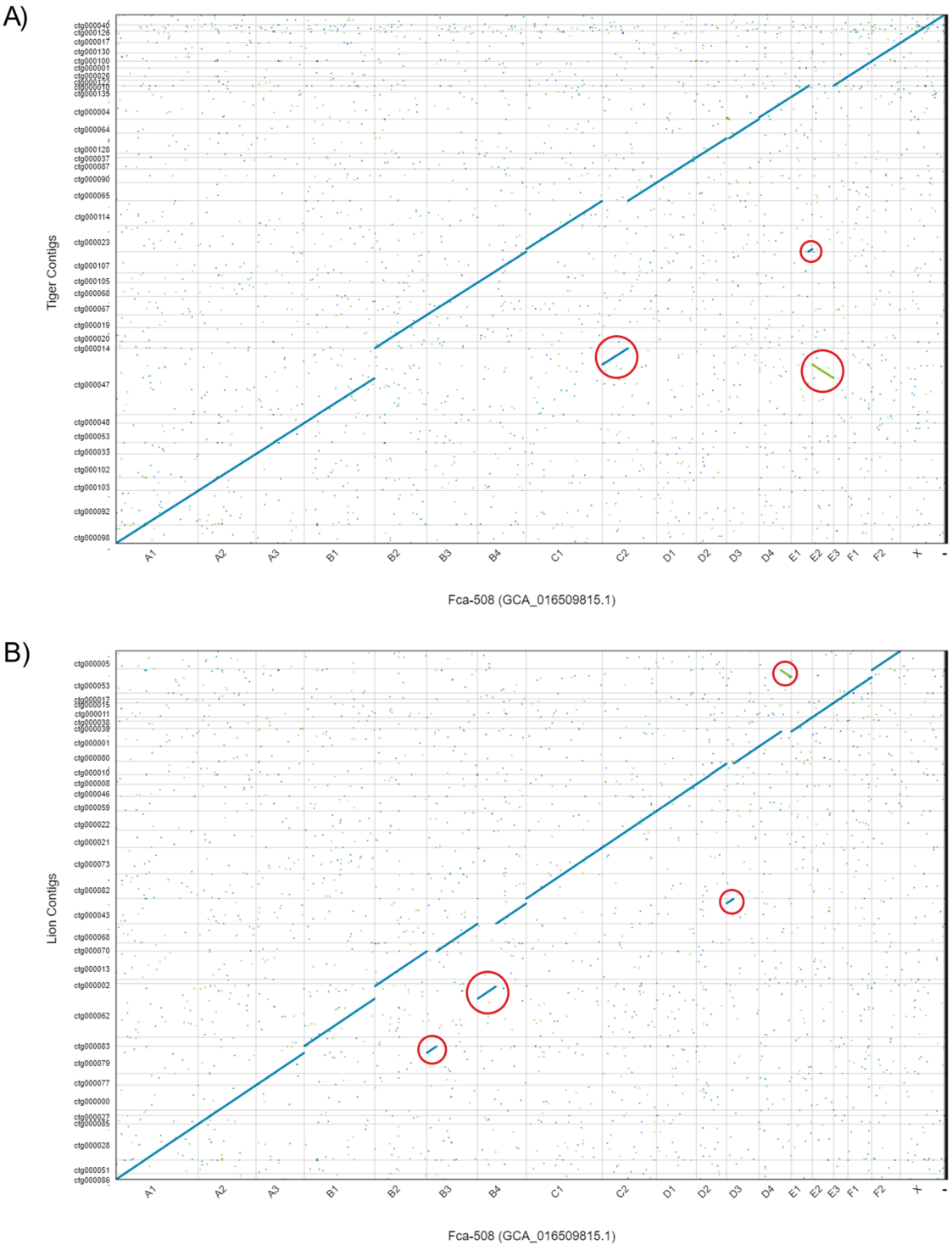


Figure B4.3. Cytogenetic analysis of the F1 Safari Cat cell line used for generation of the Geoffroy's and domestic cat (Fca-126) assemblies. The F1 and F2 chromosomes originate from the domestic haplotype while the F1-F2 fused C3 originates from the Geoffroy's cat haplotype. Performed by Terje Raudsepp, Unpublished.

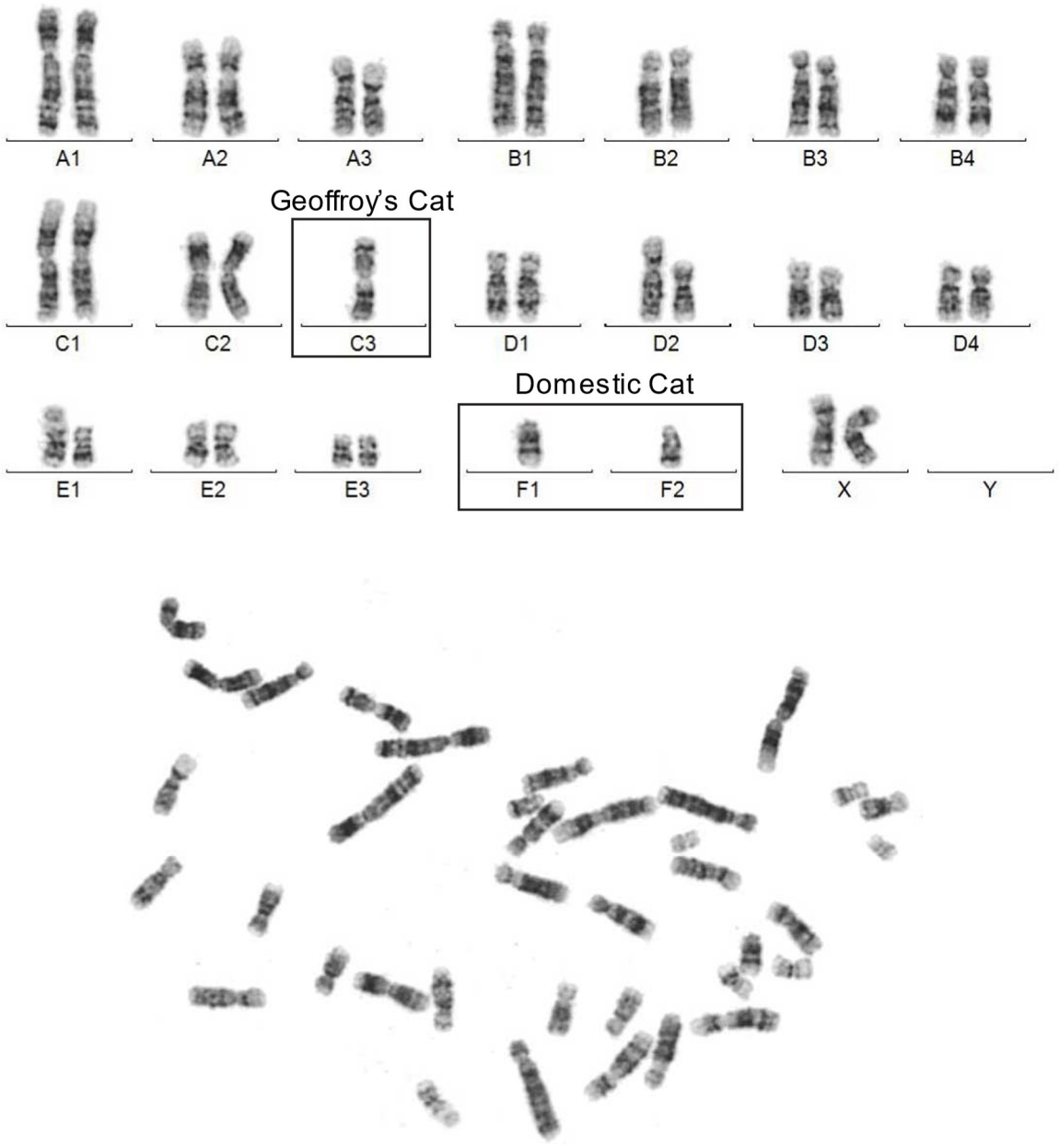


Figure B4.4. Dotplots between three cat single haplotype X chromosome assemblies (y-axis) and the diploid reference, felCat9 (x-axis) reveal *DXZ4* is captured within a single contig. Sequence gain was estimated from the alignment shift within each contig (y-axis), relative to the reference.

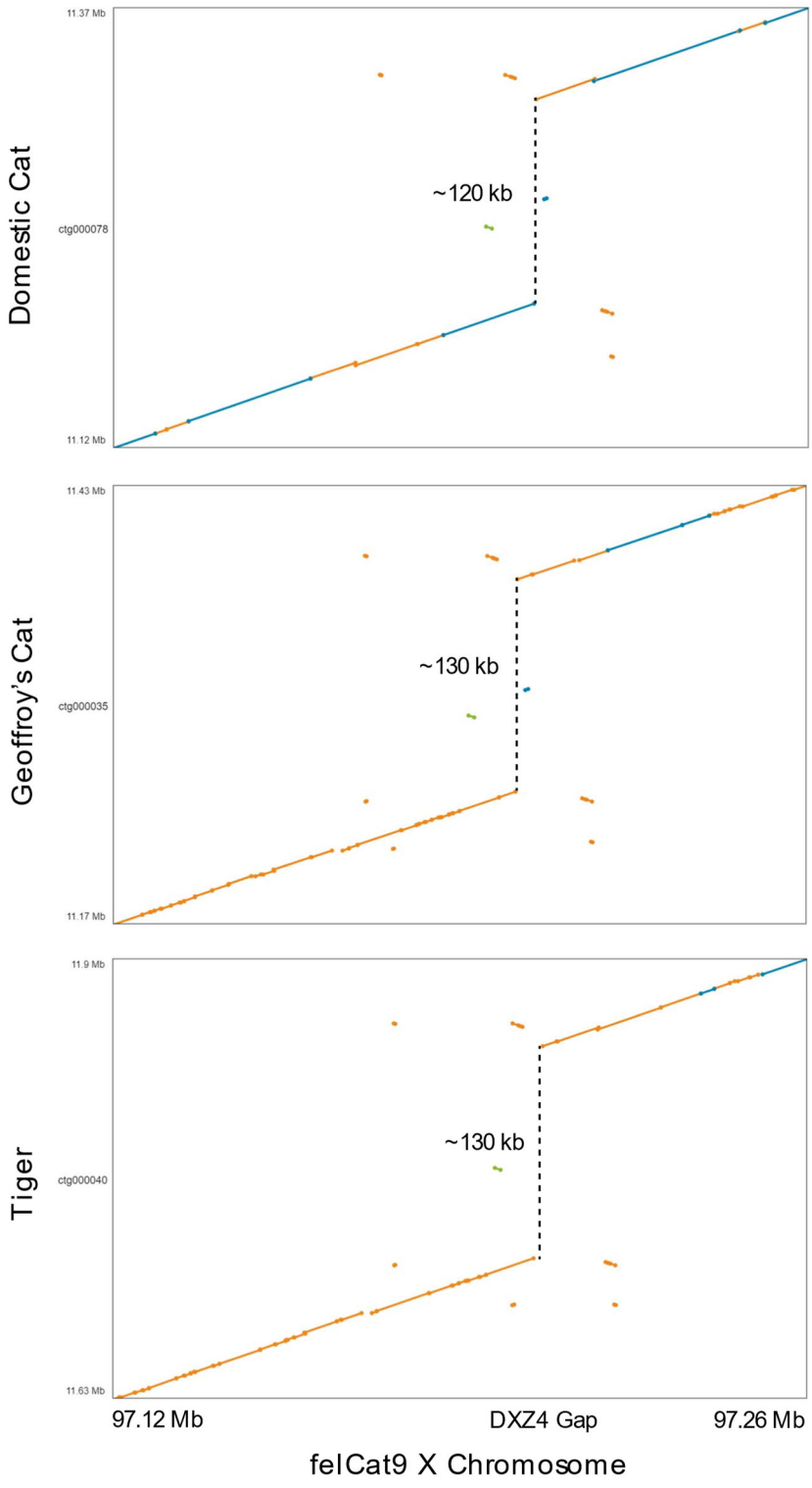


Figure B4.5. Tiger repeat A monomer alignment reveals a majority of variation in length is due to expansion/contraction of the GGGAA microsatellite (purple annotations labeled "microsat"). Grey bars represent regions masked due to alignment gaps.

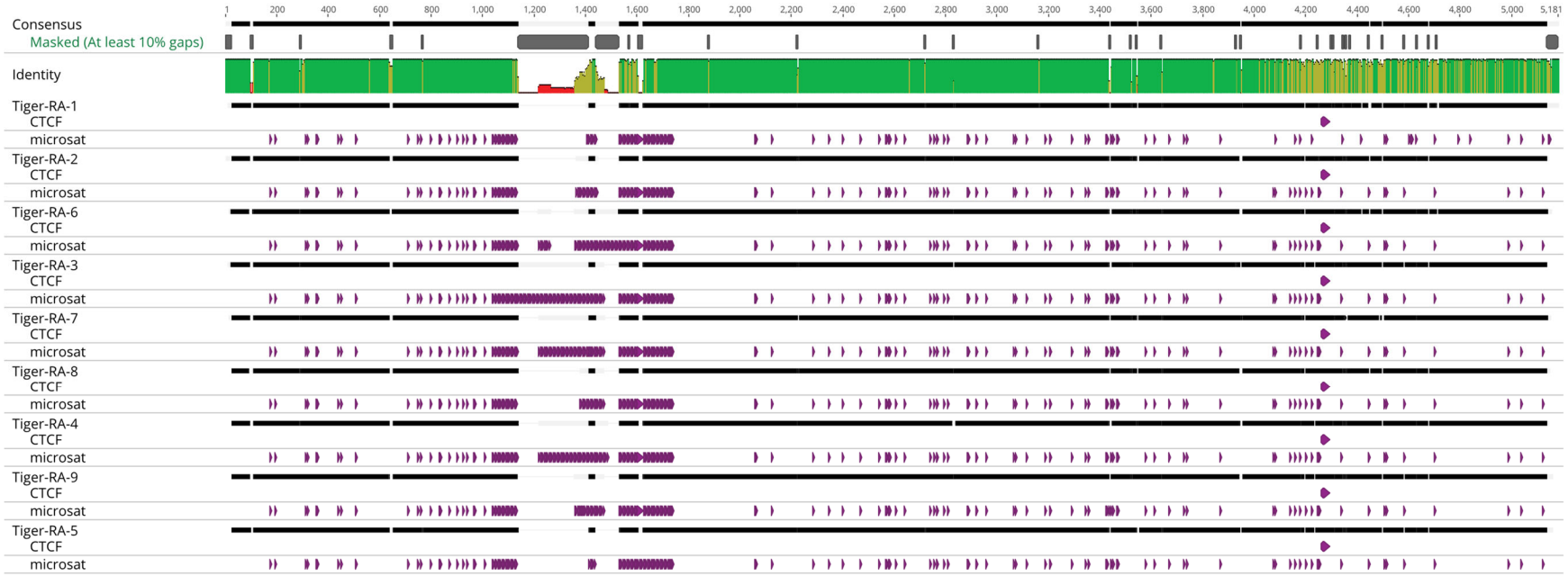


Figure B4.6. *DXZ4* spacer sequence alignment. Alignment gaps distinguished by light-grey annotations.

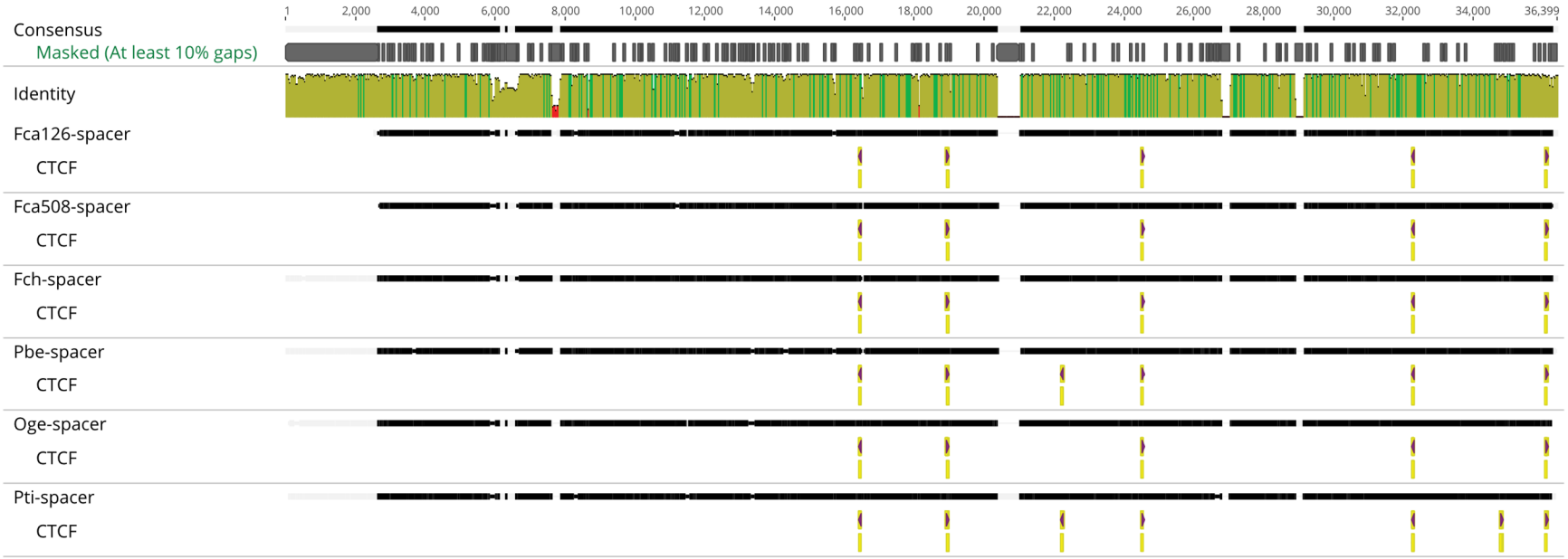


Figure B4.7. The canonical *DXZ4* macrosatellite is absent from all Bovid X-chromosome assemblies. **A)** Cow and yak region upstream of *AGTR2* exhibits CTCF binding sites identified as *DXZ4*-derived but lacks repeat structure. **B)** Sheep and goat canonical *DXZ4* regions feature an embedded *KLHL4* gene and *DXZ4*-derived CTCF clusters lacking repeat structure, similar to cow and yak. CTCF sites indicated by purple arrows. Blue line at bottom representative of changes in GC content. Sheep sequence represented in reverse orientation to align with canonical *PLS3* directionality.

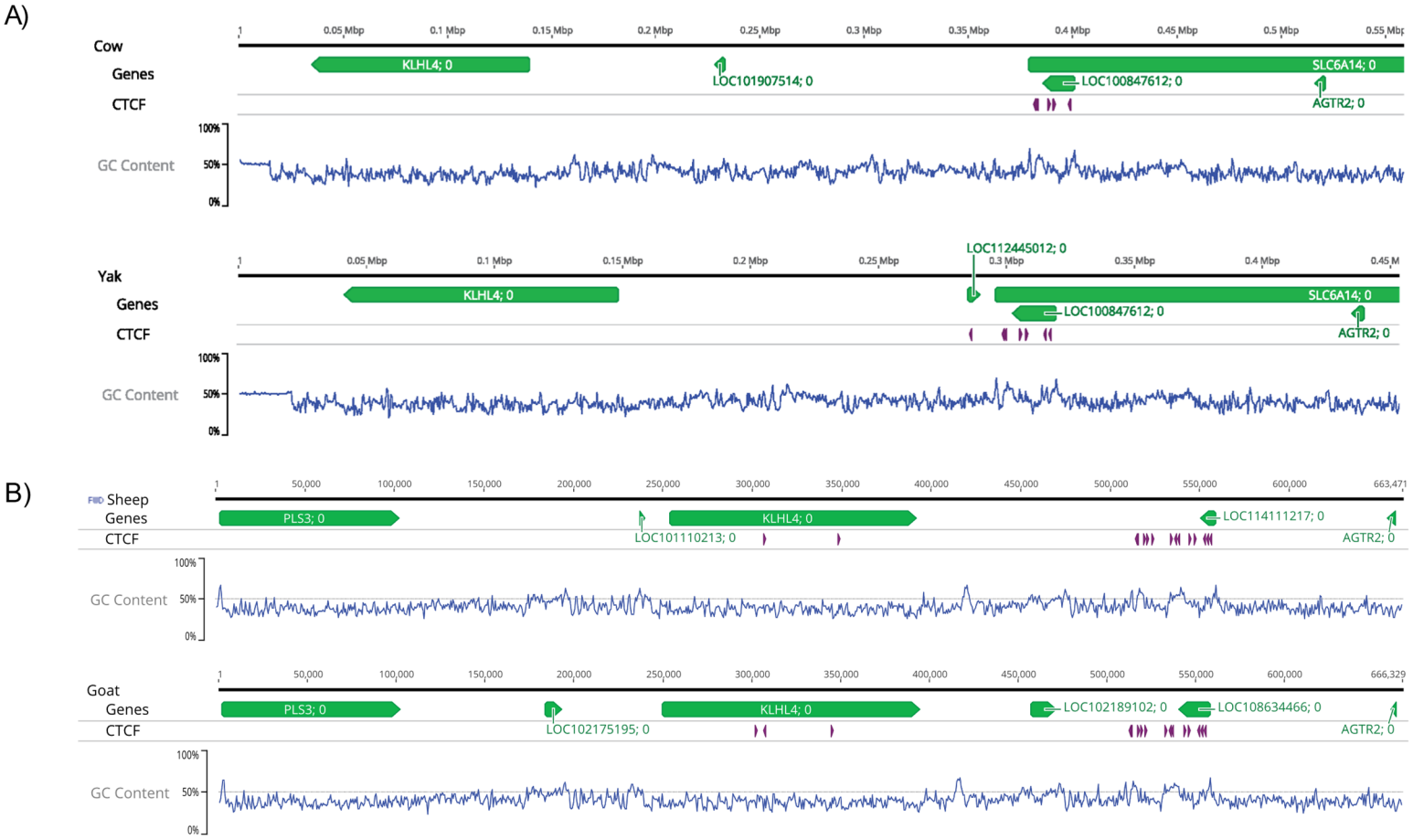


Figure B4.8. Self-alignment of *DXZ4* regions, word size=15.

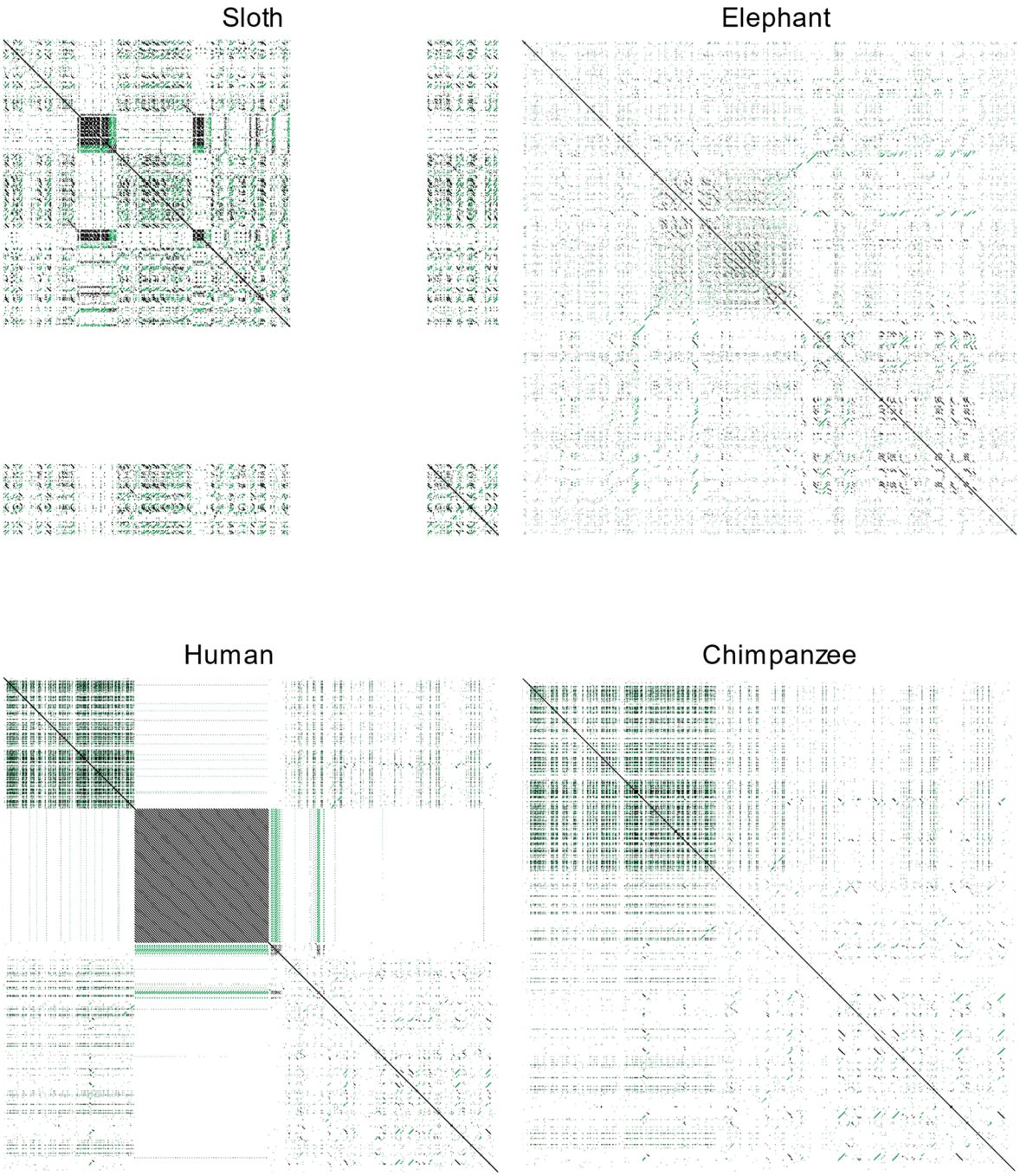


Figure B4.9. Self-alignment of *DXZ4* regions, word size=15. Rabbit and mouse sequences represented in reverse orientation to correspond with canonical *PLS3* directionality.

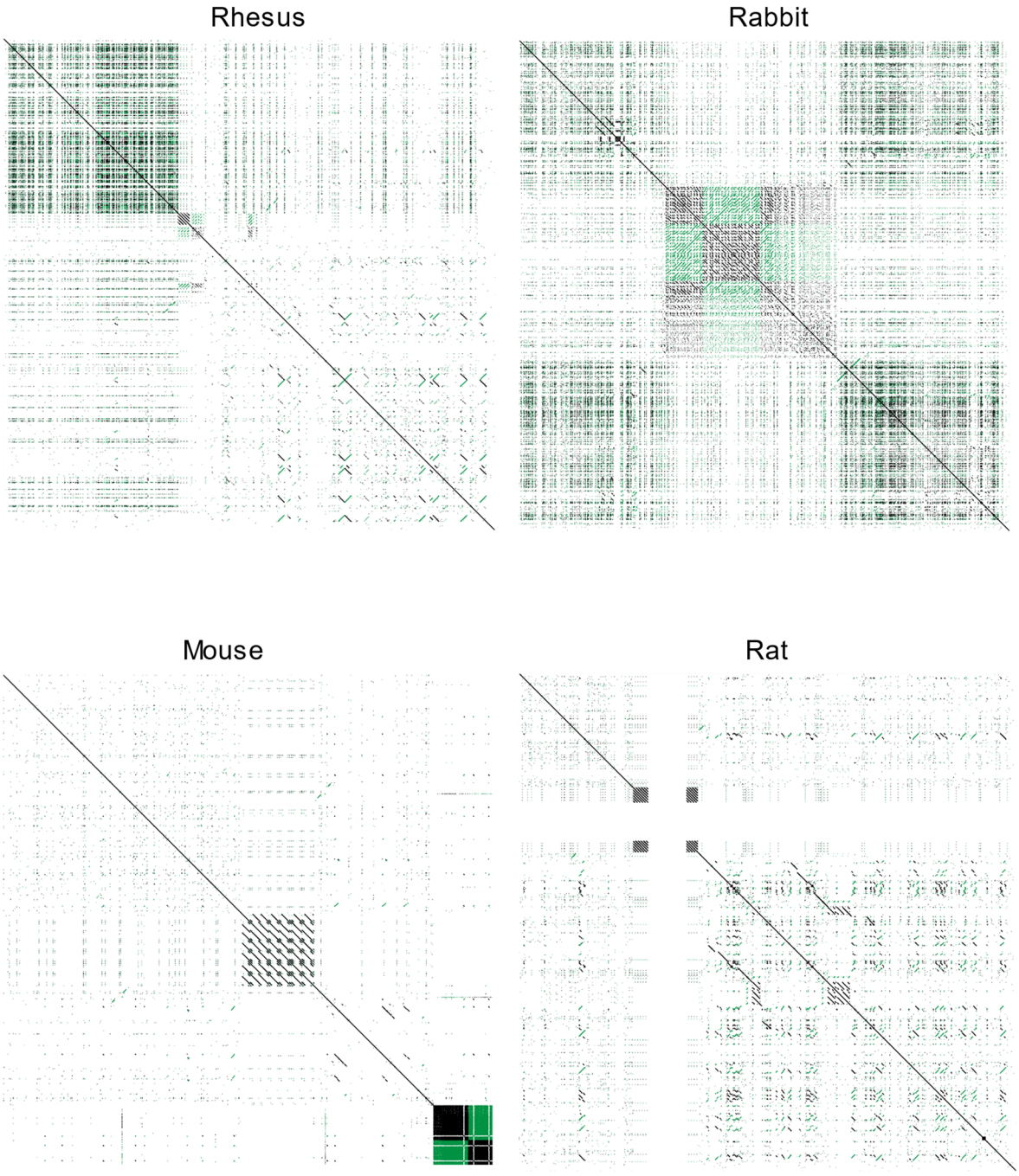


Figure B4.10. Self-alignment of *DXZ4* regions, word size=15.

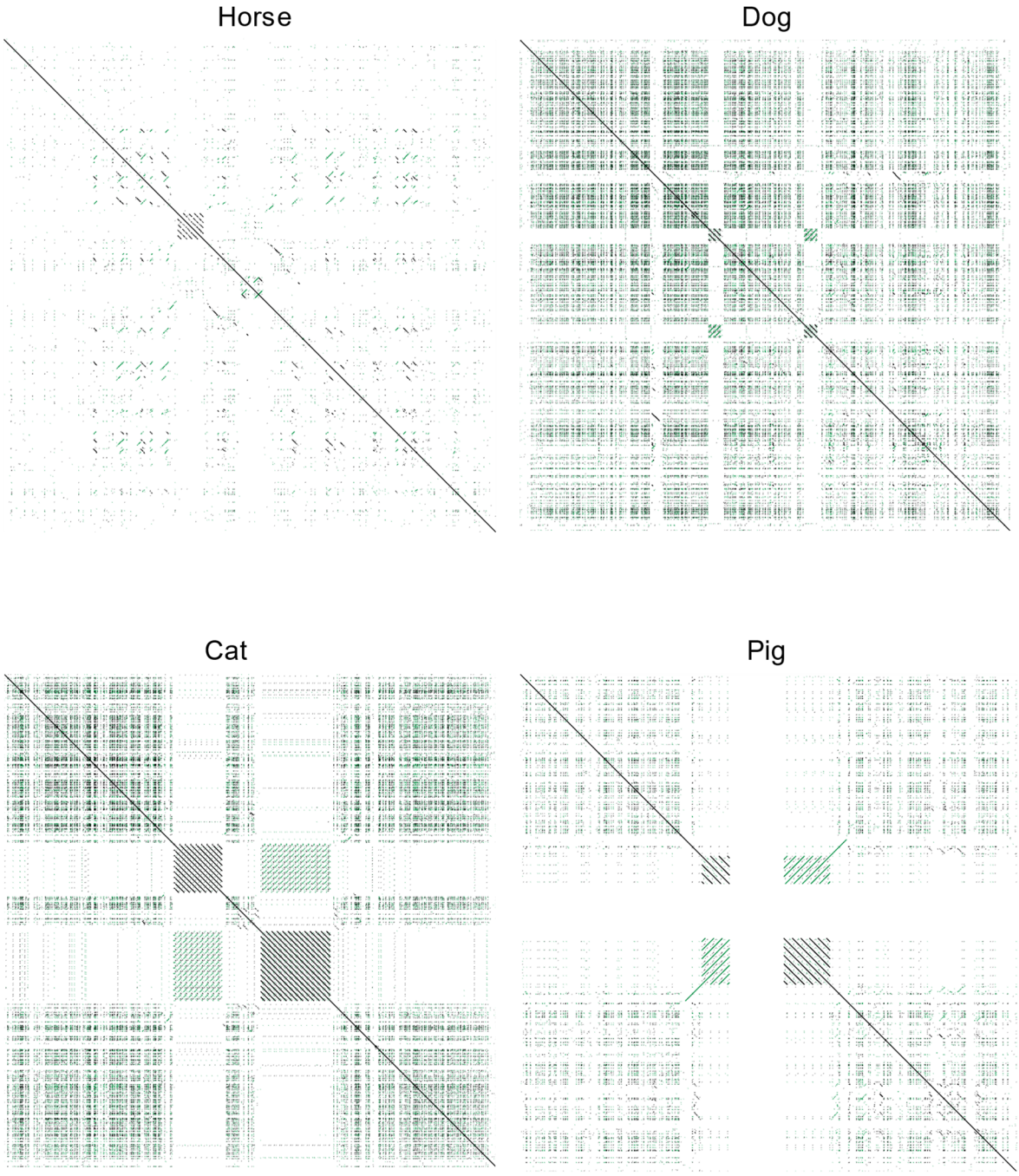


Figure B4.11. Self-alignment of *DXZ4* regions, word size=15.

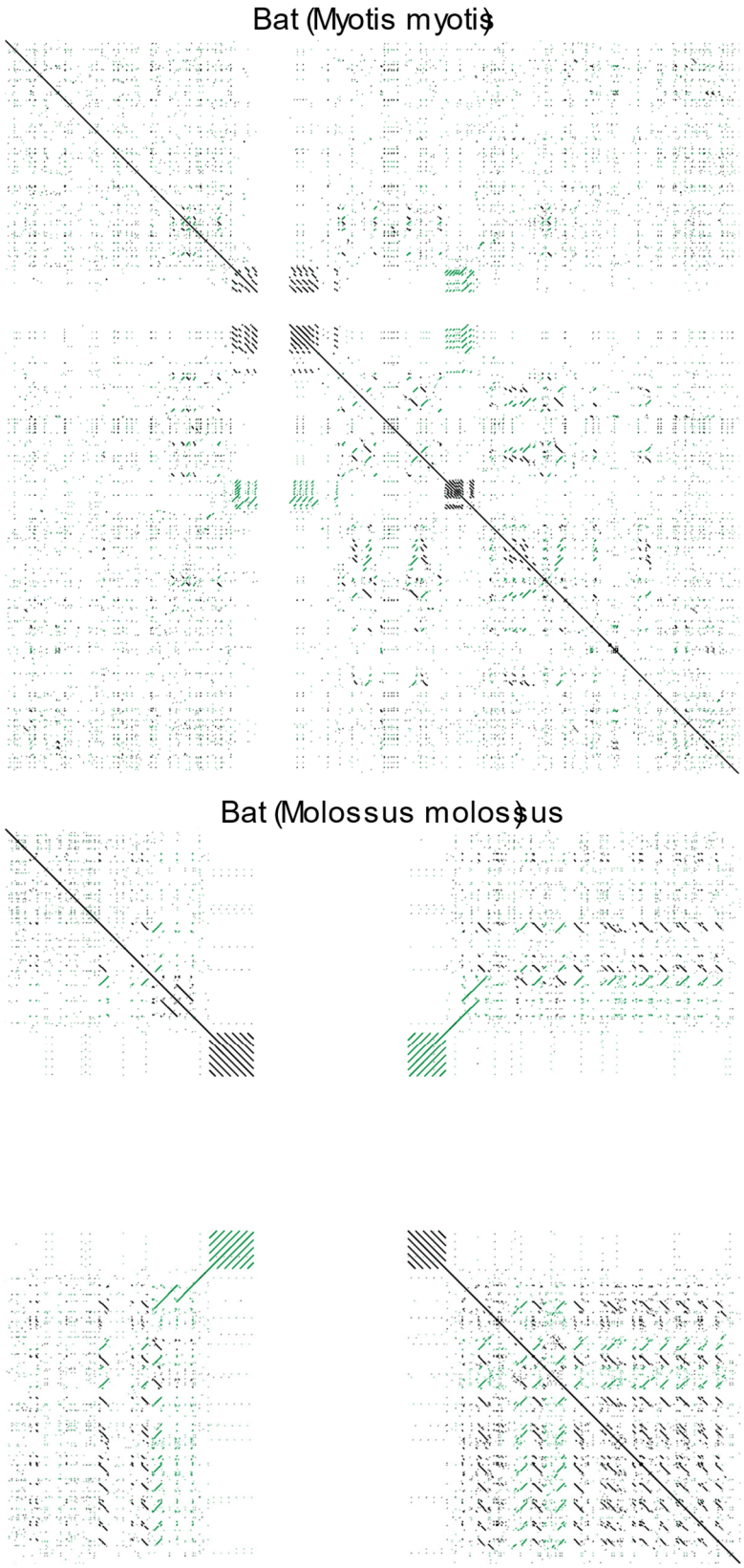


Figure B4.12. Alignment between human and chimpanzee *DXZ4* regions. The entirety of the human *DXZ4* macrosatellite resides within the chimpanzee gap, suggesting the region is incomplete but not lacking in chimpanzees.

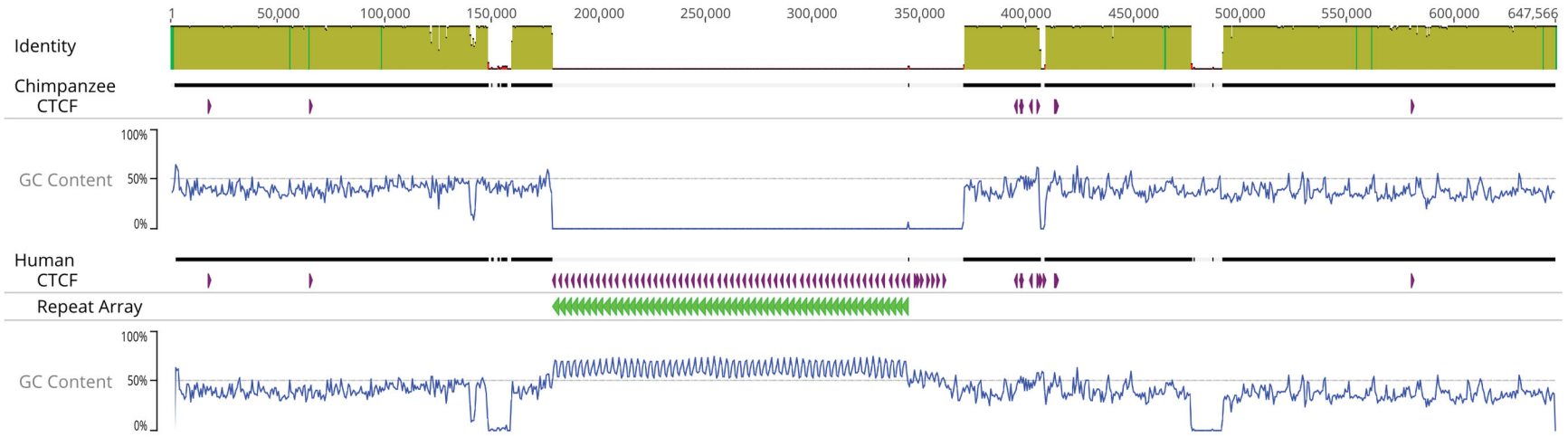


Figure B4.13. *DXZ4* regions between *PLS3* and *AGTR2*. Genes represented by green arrows, gaps represented by black rectangles and CTCF binding motifs represented by purple arrows. Blue GC content traces used to visualize high GC content and satellite sequence.

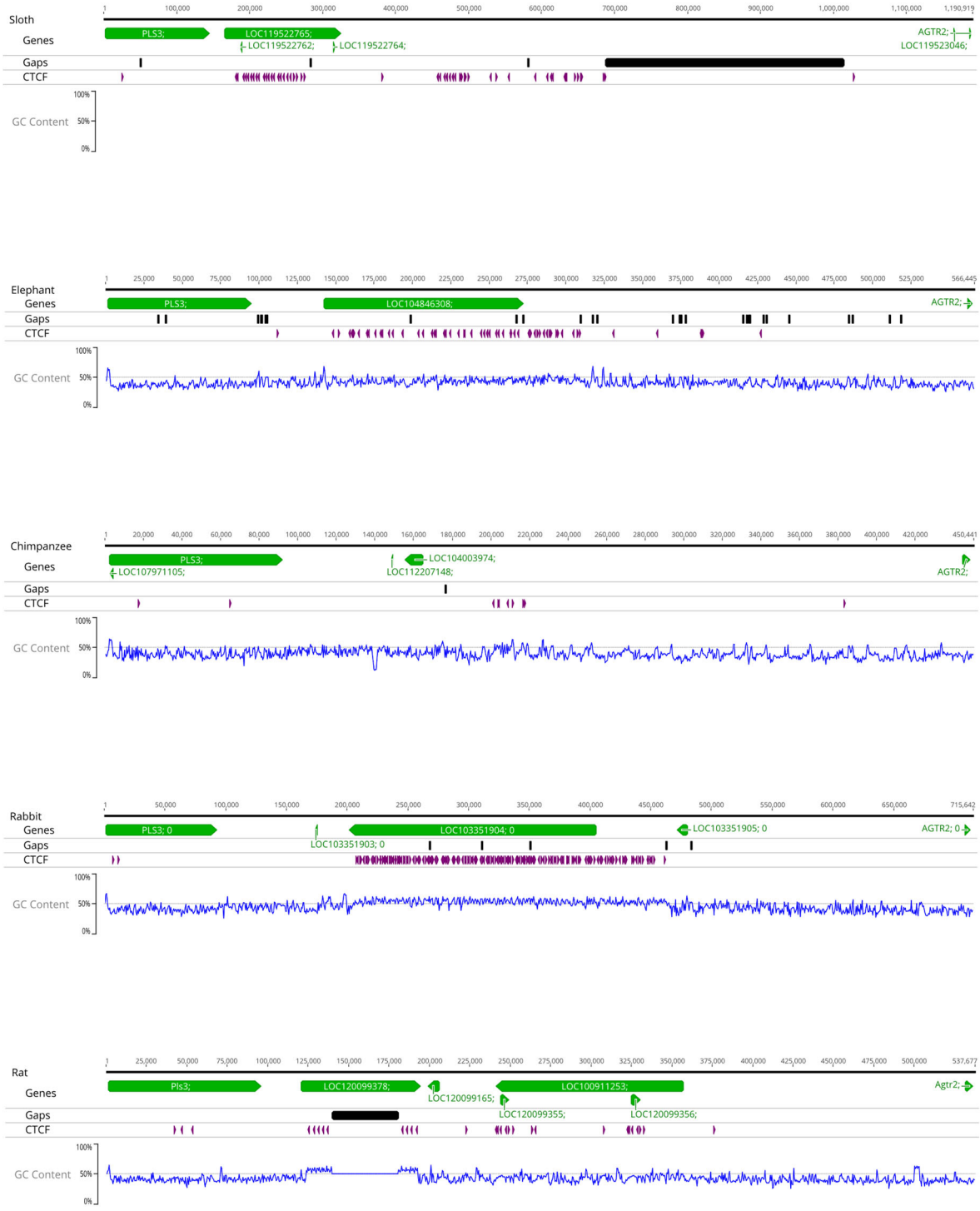


Figure B4.14. *DXZ4* regions between *PLS3* and *AGTR2*. Genes represented by green arrows, gaps represented by black rectangles and CTCF binding motifs represented by purple arrows. Blue GC content traces used to visualize high GC content and satellite sequence.

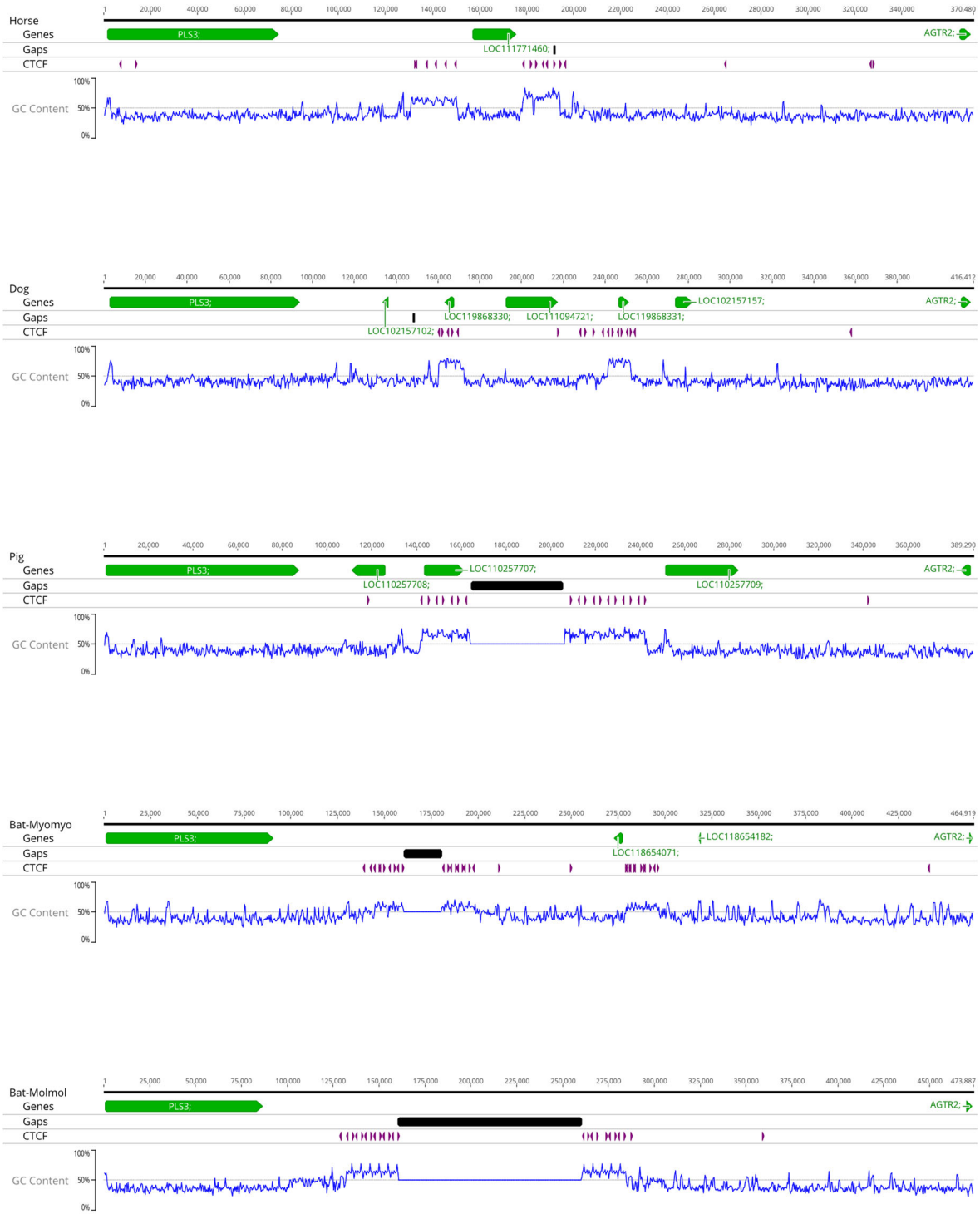


Figure B4.15. Resolved *DXZ4* macrosatellites. Green arrows are genes, purple arrows are CTCF binding motifs and blue trace reflects GC content and aids in visualization of repeat monomers. Mouse orientation is reversed to reflect canonical *PLS3* direction.

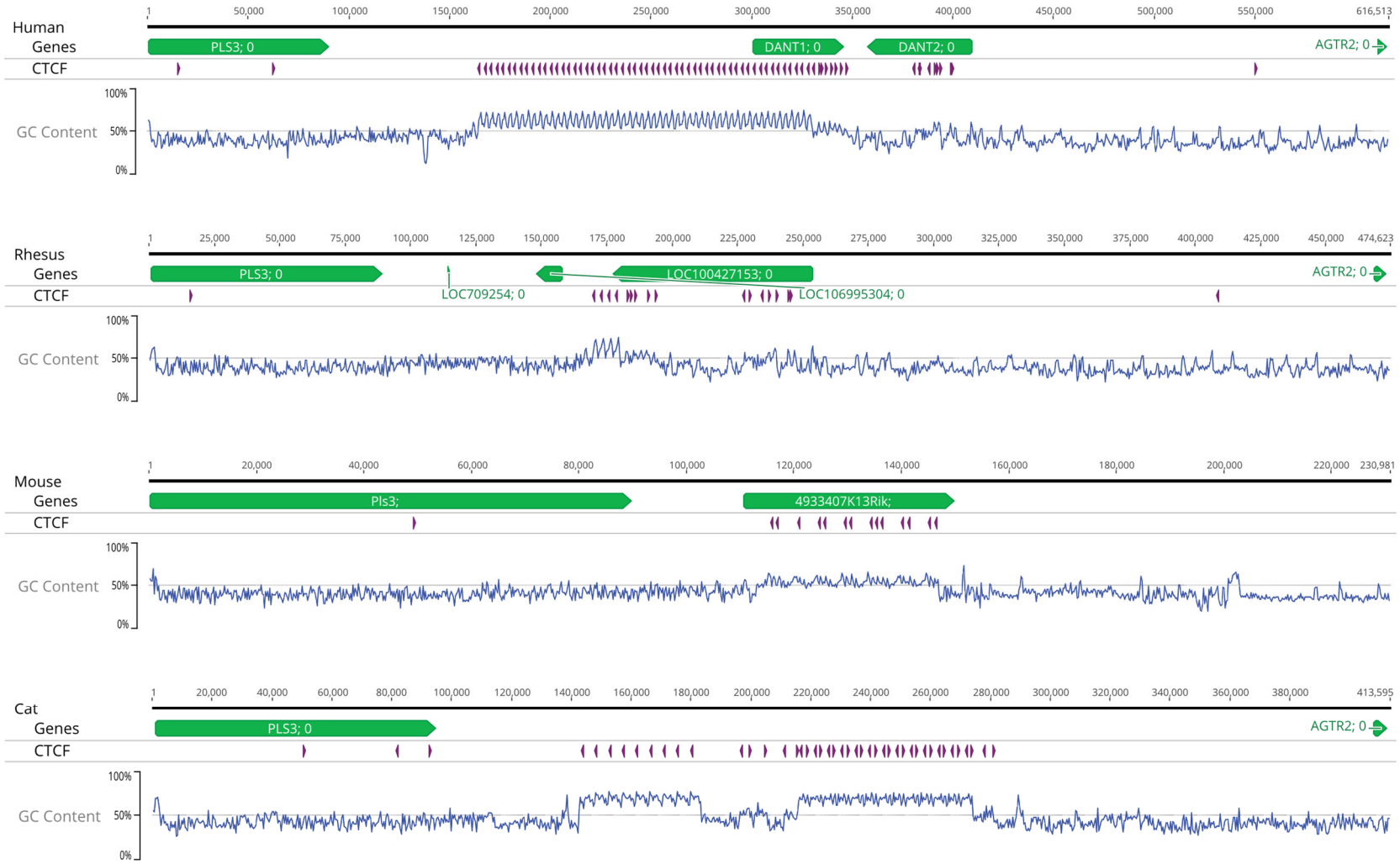


Figure B4.16. Inversion within rabbit *DXZ4* macrosatellite is bordered by assembly gaps and likely due to misassembly. Green line in dotplot representative of reverse complemented sequence. Black annotations identify gaps in assembly.

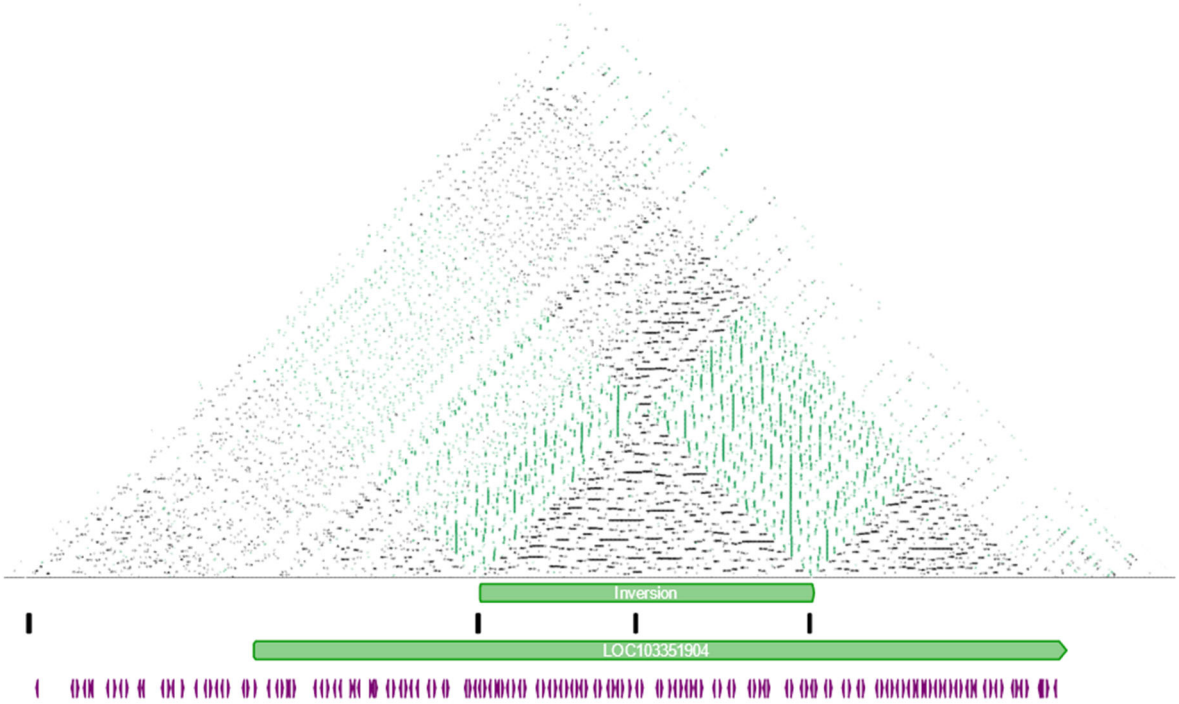
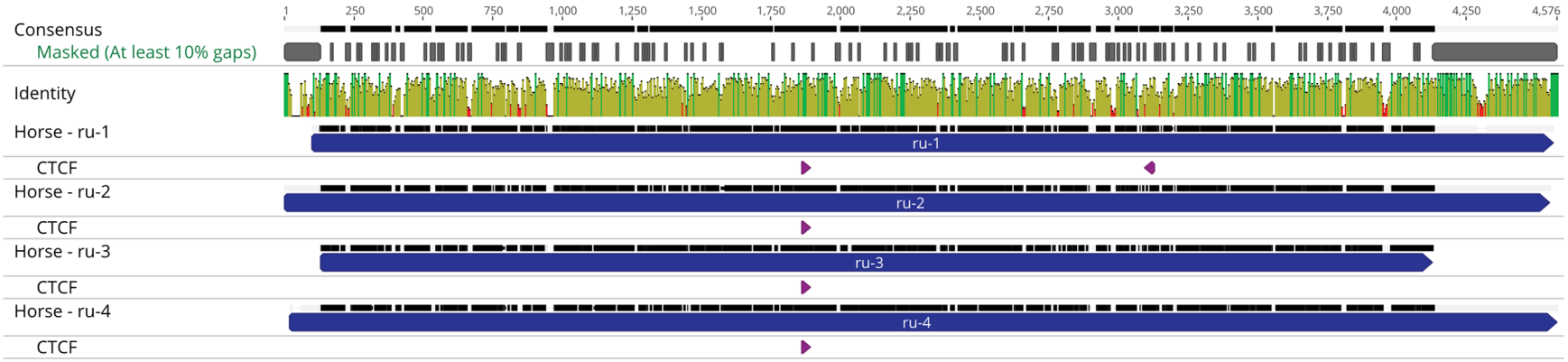


Figure B4.17. Top. Mafft alignment between horse repeat monomers identifies high sequence divergence across the entire length. Grey annotations reflect gapped regions, Purple annotations represent CTCF binding motifs. **Bottom.** Pairwise distance matrix reveals low sequence identity between all horse monomers after masking gaps.



	Horse - ru-1	Horse - ru-2	Horse - ru-3	Horse - ru-4
Horse - ru-1		77%	77%	77%
Horse - ru-2	77%		78%	76%
Horse - ru-3	77%	78%		76%
Horse - ru-4	77%	76%	76%	

Figure B4.18. Maximum likelihood phylogeny generated from 10% gapped alignment of all annotated *DXZ4* repeat monomers.

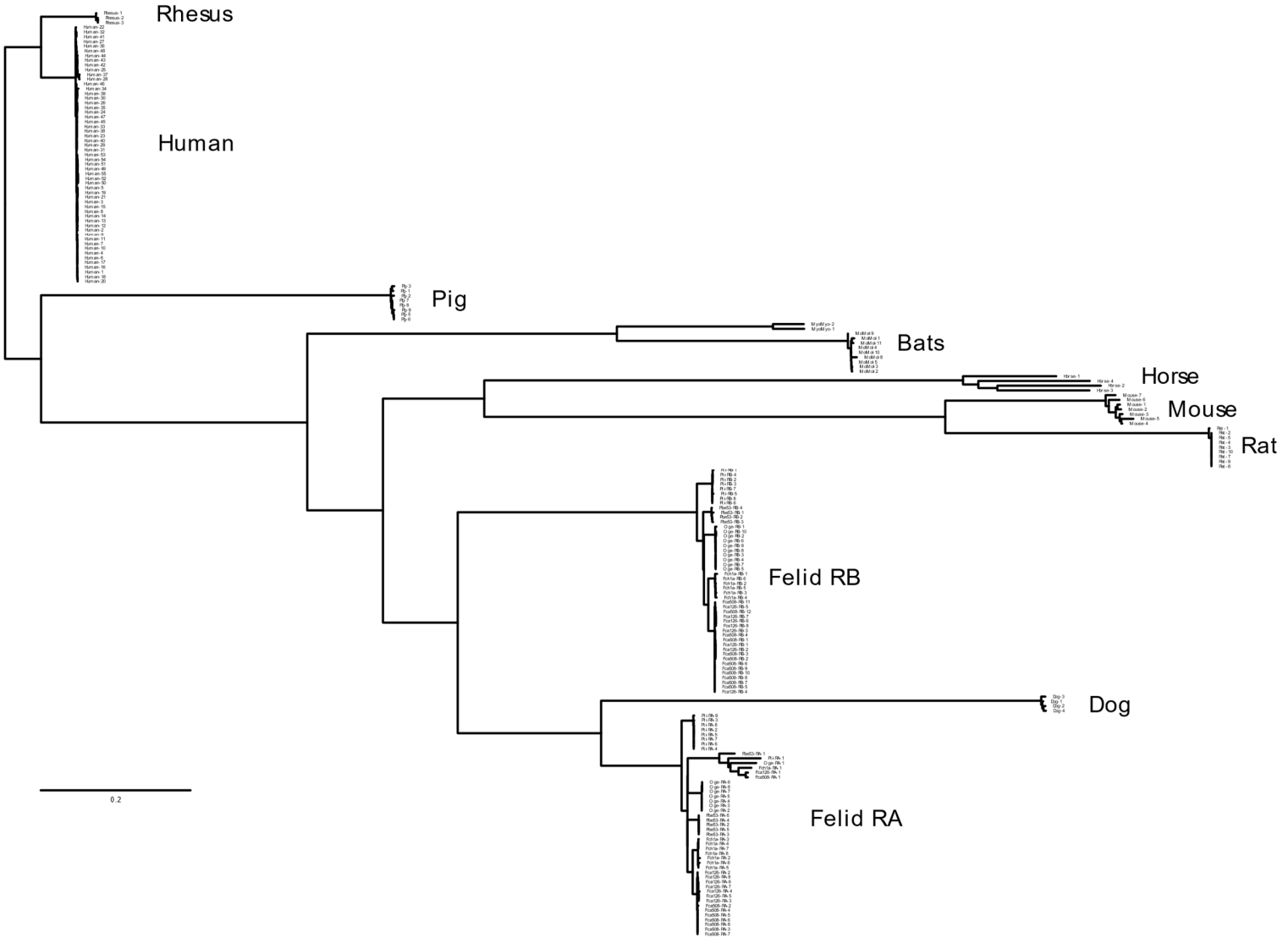


Figure B4.19. Maximum likelihood phylogeny generated from 10% gapped alignment of all repeat monomer CTCF binding motifs. Identical CTCF sequences collapsed into a single branch. Tree rooted by midpoint.

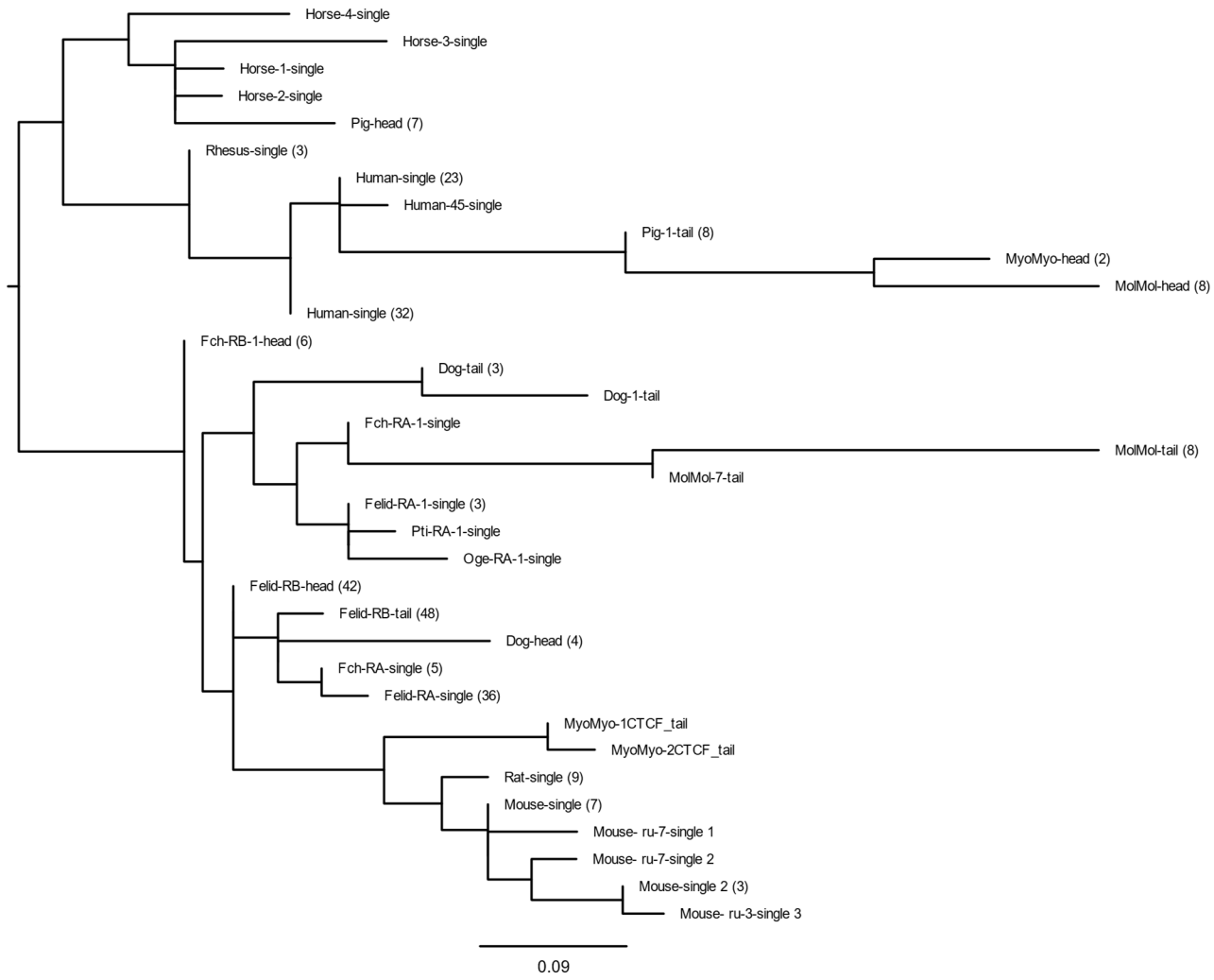


Figure B.A.1. Dotplot of de novo FelCha1.0 mitochondrial assembly and previous Jungle cat short-read mitochondrial assembly. Zoomed pane shows repetitive sequence gained in new assembly.

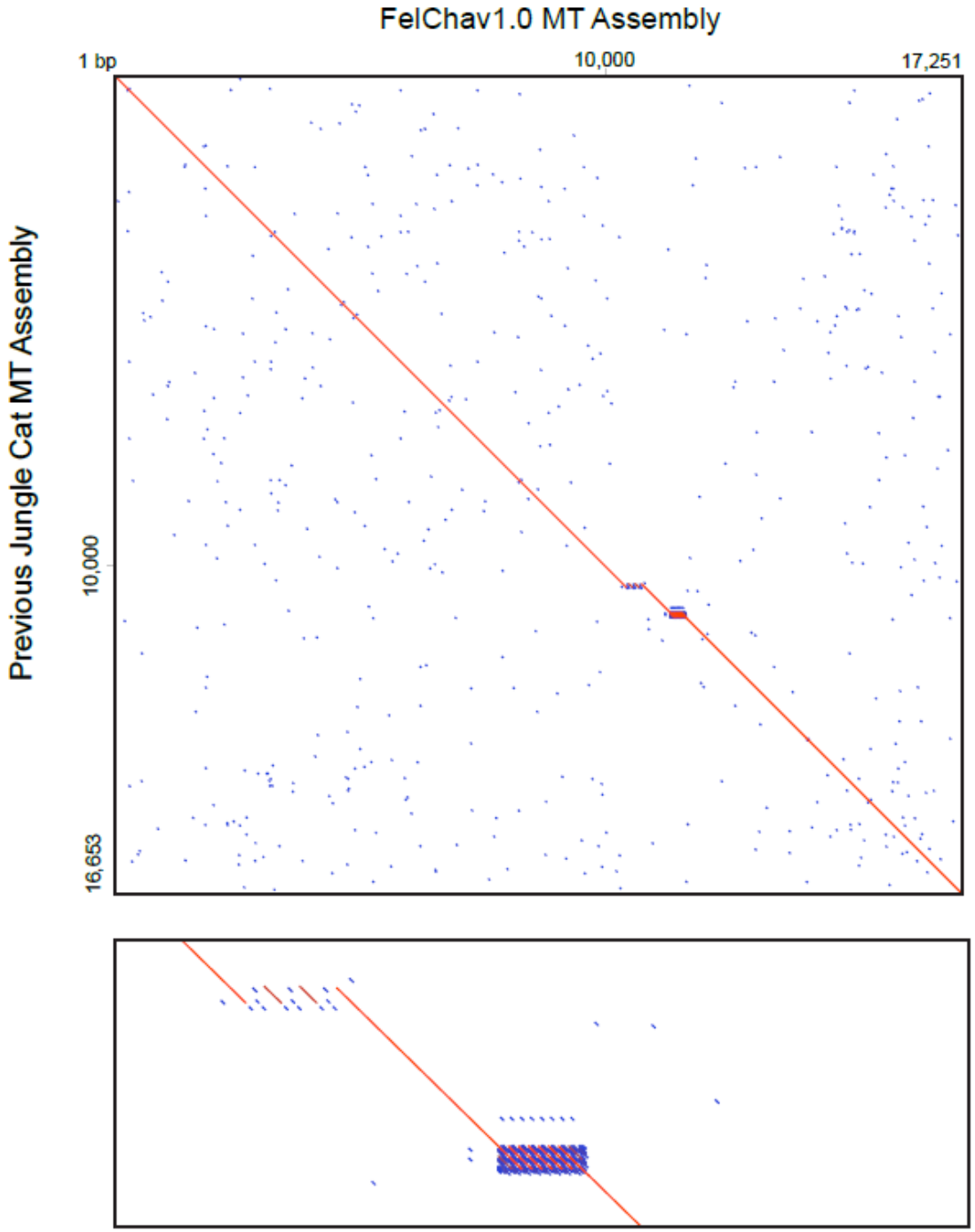


Figure B.A.2. Contig alignments to domestic cat single haplotype assembly (GCA_016509815.1). Chimeric contig ctg000159 representative of interchromosomal misjoin between B3 and E1 indicated by green circle.

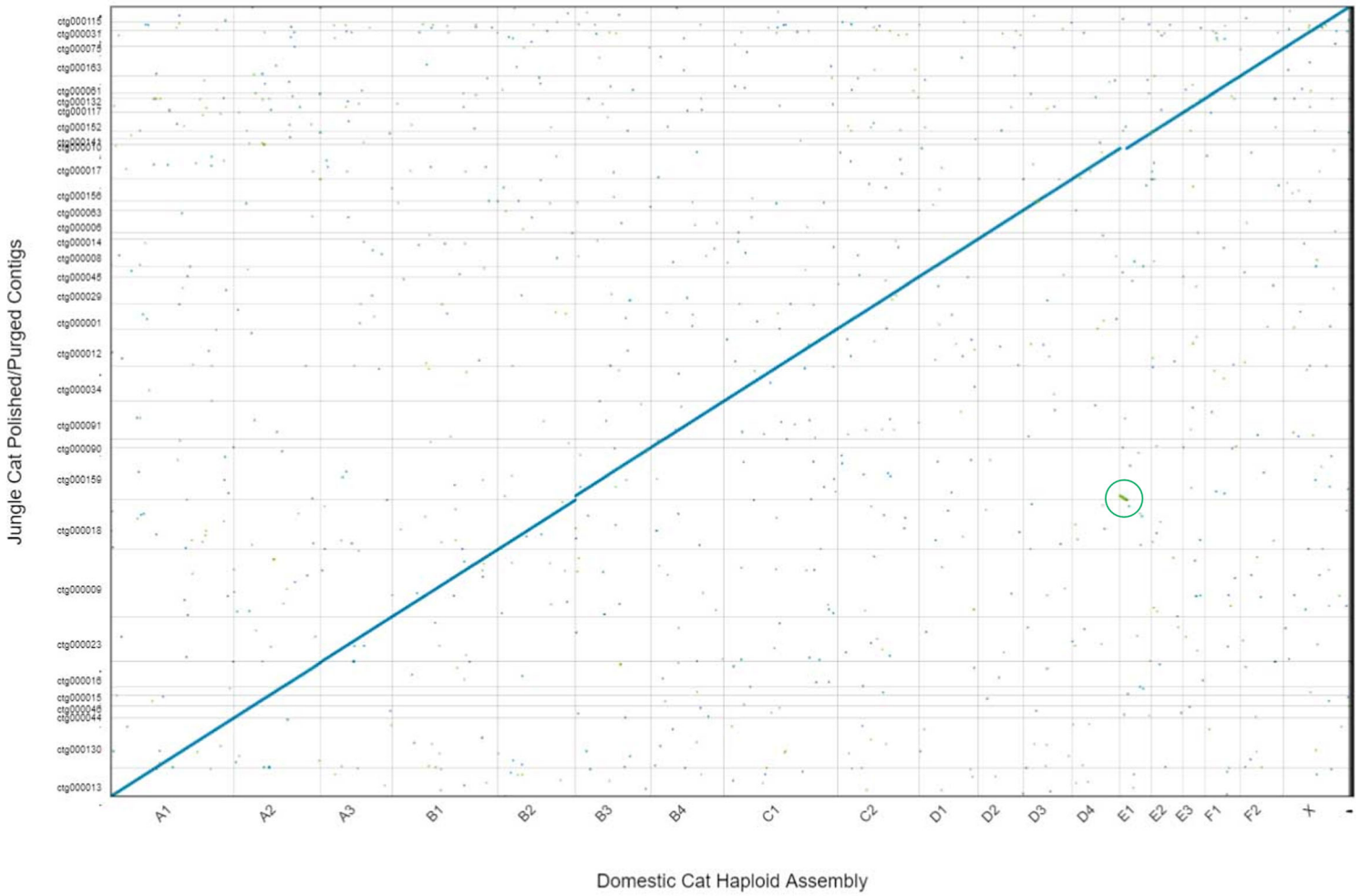
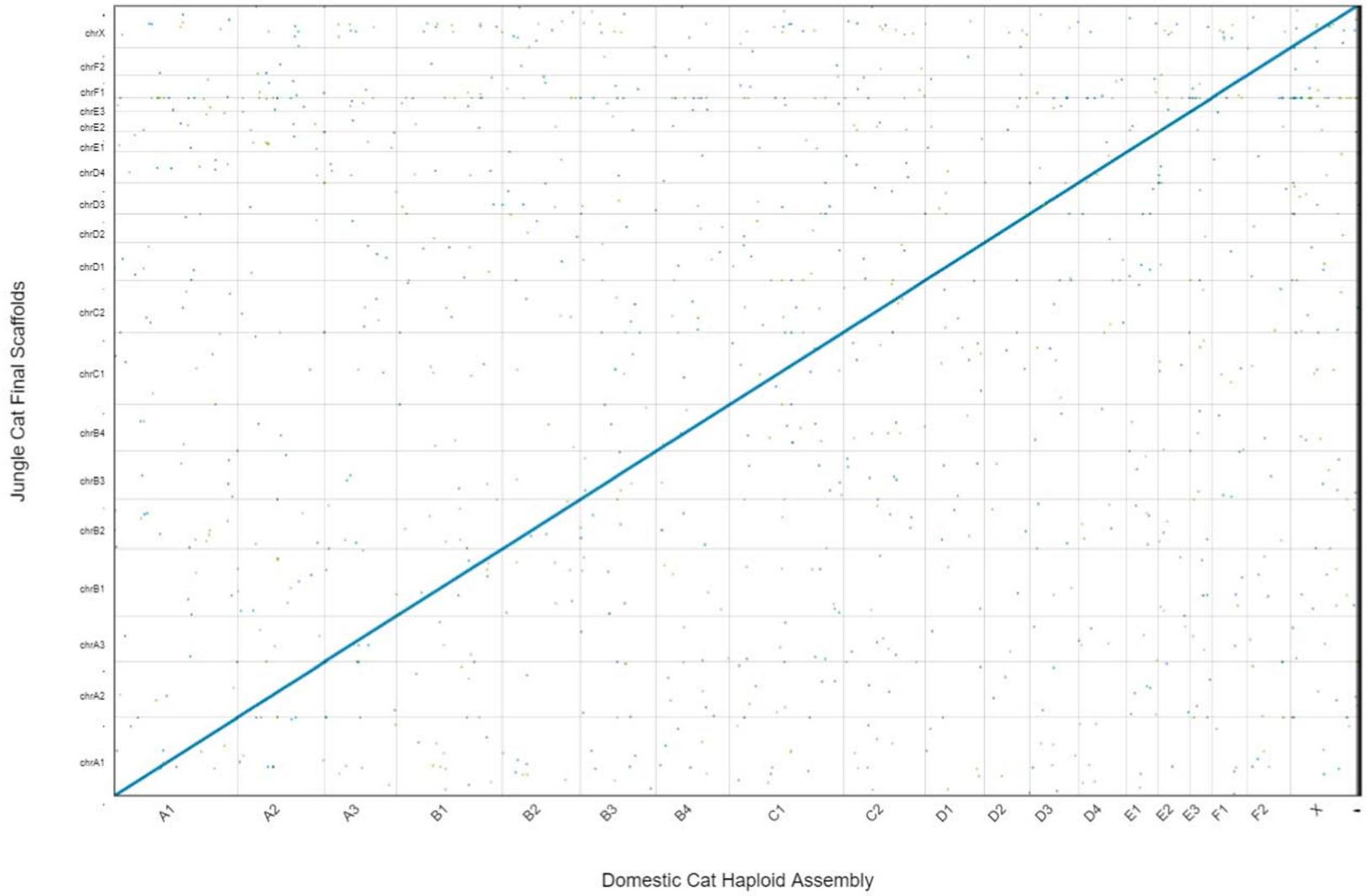


Figure B.A.3. Final scaffold alignments to domestic cat single haplotype assembly (GCA_016509815.1)



APPENDIX C SUPPLEMENTAL TABLES

Table C2.1. Sample accession information for additional Asian leopard cats (1-3), *Prionailurus* species (4-6), and domestic cat (7-9) used in the assembly QC and Phased Haplotype Analysis (PHA).

Sample ID	SRA ID	Species	SRA #	BioProject
1. Pbe-6	PBE	*Prionailurus b. bengalensis*	SRR2062628	PRJNA286910
2. Pbe-14	PBEP0014	*Prionailurus b. euptilurus*	SRR4426166	PRJNA348661
3. Pbe-38	PBEP0038	*Prionailurus b. bengalensis*	SRR4426179	PRJNA348661
4. Pja-5	PBEP0005	*Prionailurus javanensis*	SRR4426170	PRJNA348661
5. Pja-25	PBEP0025	*Prionailurus javanensis*	SRR4426162	PRJNA348661
6. Pvi-12	PVIP0012	*Prionailurus viverrinus*	SRR4426169	PRJNA348661
7. Pammi	100_Pammi	*Felis catus*	SRR8092623	PRJNA495843
8. CatIII	CatIII	*Felis catus*	SRR5815677	PRJNA393717
9. LilBub	LilBUB	*Felis catus*	SRR8377759	PRJNA512113

Table C2.2. Sample and sequence information for the domestic cat, Asian leopard cat, and two other *Prionailurus* species used for assembly, assembly QC and the Phased Haplotype Analysis.

Library Name/ ID	Avg. Read Length	Base Count	Organism	Subspecies	Sex	Instrument
Pbe-53	150	52,451,028,300	*Prionailurus bengalensis*	*euptilurus*	M	Illumina NovaSeq 6000
Pbe-14	125	32,821,274,250	*Prionailurus bengalensis*	*euptilurus*	F	Illumina HiSeq 2500
Pbe-38	125	37,174,242,250	*Prionailurus bengalensis*	*bengalensis*	F	Illumina HiSeq 2500
Pbe-6	101	64,245,439,662	*Prionailurus bengalensis*	*bengalensis*	F	Illumina HiSeq 2000
Pja-5	125	31,492,238,250	*Prionailurus javanensis**	*sumatranus*	M	Illumina HiSeq 2500
Pja-25	101	40,850,158,010	*Prionailurus javanensis**	*sumatranus*	M	Illumina HiSeq 2000
Pvi-12	101	39,586,764,968	*Prionailurus viverrinus*	*viverrinus*	M	Illumina HiSeq 2000
Fca-508	150	71,002,000,000	*Felis catus*	N/A	F	Illumina NovaSeq 6000
Pammi	151	146,300,000,000	*Felis catus*	N/A	F	Illumina HiSeq X Ten
CatIII	137	82,600,000,000	*Felis catus*	N/A	F	Illumina HiSeq X Ten
LilBub	155	93,110,000,000	*Felis catus*	N/A	F	Illumina NextSeq 500

*These samples were reclassified as *Prionailurus javanensis sumatranus* after they were uploaded to NCBI

Table C2.3. Raw sequencing output and haplotyping results.

PacBio Long Reads

Sample	Reads	Bases (bp)	Coverage	Subread N50 (bp)	Avg. Read Length (bp)
F1 Bengal	13,801,880	221,688,858,786	89x	25,561	16,062
Domestic Cat Haplotype	6,342,174	109,251,556,255	44x	25,541	17,226
Leopard Cat Haplotype	6,519,732	112,023,028,516	45x	25,585	17,182
Unknown Haplotype	11,876	20,811,380	.01x	1,667	1,752

Illumina Reads

Sample	Reads	Bases (bp)	Coverage
Domestic Cat Parent (Fca-508)	473,347,659	142,004,297,700	57x
Leopard Cat Parent (Pbe-53)	349,673,522	104,902,056,600	42x

***In situ* DNase Hi-C Reads**

Sample	Reads	Bases (bp)	Coverage*
F1 Bengal	481,008,096	144,302,428,800	58x
Domestic Cat Haplotype	273,559,977	82,067,993,100	33x
Leopard Cat Haplotype	278,398,530	83,519,559,000	33x

*The summed number of haplotyped reads is higher than the F1 raw total because reads mapping equally well to both assemblies were included in eachhaplotype (following Rice et al., 2020).

Table C2.4. Domestic cat single haplotype chromosome assembly.

Molecule	Total Length (bp)	Gaps (#)	Un-gapped (bp)
ALL	**2,422,299,418**	**60**	**2,422,283,418**
Chromosome A1	239,109,665	4	239,108,865
Chromosome A2	168,571,291	6	168,569,491
Chromosome A3	140,469,438	0	140,469,438
Chromosome B1	205,171,639	0	205,171,639
Chromosome B2	152,154,423	2	152,153,823
Chromosome B3	147,603,332	1	147,603,232
Chromosome B4	141,964,754	0	141,964,754
Chromosome C1	221,453,569	2	221,453,369
Chromosome C2	158,479,582	1	158,479,482
Chromosome D1	115,437,799	2	115,437,599
Chromosome D2	88,171,565	3	88,170,865
Chromosome D3	94,347,243	3	94,346,543
Chromosome D4	94,319,450	5	94,317,350
Chromosome E1	61,581,190	2	61,580,590
Chromosome E2	61,931,696	1	61,931,596
Chromosome E3	41,490,710	2	41,490,510
Chromosome F1	69,616,193	2	69,615,193
Chromosome F2	83,330,349	1	83,330,249
Chromosome X	127,194,472	17	127,190,772
Unplaced (n=52)	9,901,058	6	9,898,058

Table C2.5. Asian leopard cat single haplotype chromosome assembly.

Molecule	Total Length (bp)	Gaps (#)	Un-gapped (bp)
ALL	**2,435,702,060**	**56**	**2,435,689,660**
Chromosome A1	240,846,738	0	240,846,738
Chromosome A2	168,940,850	3	168,940,150
Chromosome A3	140,803,547	2	140,802,947
Chromosome B1	206,580,432	2	206,579,832
Chromosome B2	152,385,405	2	152,385,205
Chromosome B3	148,587,958	1	148,587,858
Chromosome B4	142,198,231	1	142,197,731
Chromosome C1	222,814,610	2	222,814,410
Chromosome C2	159,850,271	3	159,849,971
Chromosome D1	116,110,351	3	116,110,051
Chromosome D2	87,619,889	1	87,619,389
Chromosome D3	94,595,352	7	94,594,252
Chromosome D4	94,620,989	1	94,620,489
Chromosome E1	61,174,949	2	61,174,349
Chromosome E2	62,591,731	1	62,591,631
Chromosome E3	41,869,280	0	41,869,280
Chromosome E4	69,230,405	2	69,229,405
Chromosome F2	83,696,601	1	83,696,501
Chromosome X	129,104,405	19	129,100,905
Unplaced (n=64)	12,080,066	3	12,078,566

Table C2.6. RepeatMasker repeat analysis summary.

# bp masked / % of total sequence	Domestic Cat Scaffolds			Leopard Cat Scaffolds		
	813,861,694		33.60%	822,029,473		33.75%
Elements	Number of Elements	Length Occupied (bp)	Percentageof Sequence	Number of Elements	Length Occupied (bp)	Percentageof Sequence
SINEs	**469,804**	**69,316,698**	**2.86%**	**468,697**	**69,148,368**	**2.84%**
Alu/B1	0	0	0.00%	0	0	0.00%
MIRs	462,205	68,412,320	2.82%	461,113	68,245,362	2.80%
LINEs	**818,687**	**458,240,914**	**18.92%**	**821,304**	**467,024,407**	**19.17%**
LINE1	454,137	361,491,179	14.92%	456,511	370,292,026	15.20%
LINE2	310,052	84,612,805	3.49%	310,346	84,619,403	3.47%
L3/CR1	40,752	8,854,769	0.37%	40,765	8,853,628	0.36%
RTE	12,505	3,083,991	0.13%	12,430	3,061,807	0.13%
LTR elements	**285,140**	**108,443,228**	**4.48%**	**284,988**	**108,405,945**	**4.45%**
ERVL	87,026	39,290,813	1.62%	87,082	39,352,413	1.62%
ERVL-MaLRs	145,738	50,919,236	2.10%	145,760	50,904,494	2.09%
ERV_classI	28,377	12,425,503	0.51%	28,350	12,375,675	0.51%
ERV_classIII	0	0	0.00%	0	0	0.00%
DNA elements	**342,481**	**68,413,225**	**2.82%**	**342,709**	**68,441,325**	**2.81%**
hAT-Charlie	193,904	36,333,016	1.50%	193,801	36,337,024	1.49%
TcMar-Trigger	53,529	14,398,204	0.59%	53,607	14,400,538	0.59%
Unclassified	3,746	584,939	0.02%	3,687	583,111	0.02%
Total Interspersed Repeats		**704,999,004**	**29.10%**		**713,603,156**	**29.30%**
small RNA	140,966	10,719,095	0.44%	140,848	10,718,354	0.44%
Simple Repeats	1,463,437	69,960,000	2.89%	1,463,778	68,700,051	2.82%
Low complexity	521,881	28,033,555	1.16%	538,647	28,857,794	1.18%

Table C2.7. Assemblytics variant analysis comparing the leopard cat singlehaplotype assembly to the domestic cat single haploid assembly.

Variant	Count	Total (bp)
Insertion		
50-500	13,465	2,499,723
500-10,000	734	1,906,174
Total	14,199	4,405,897
Deletion		
50-500	13,288	2,189,922
500-10,000	564	1,090,944
Total	13,852	3,280,866
Tandem Expansion		
50-500	829	178,794
500-10,000	211	309,941
Total	1,040	488,735
Tandem Contraction		
50-500	611	136,116
500-10,000	78	82,503
Total	689	218,619
Repeat Expansion		
50-500	12,756	3,686,143
500-10,000	12,680	25,435,846
Total	25,436	29,121,989
Repeat Contraction		
50-500	11,482	3,187,343
500-10,000	10,340	18,142,475
Total	21,822	21,329,818

*Leopard cat gain/loss	9,187,318 bp
Total structural variants	77,038
Total affected bases	58.85 Mb

*Leopard cat gain/loss calculated by subtracting basepairs of sequence gained (insertions/expansions) from sequence lost (deletions/contractions).

Table C2.8. Results of annotation liftover of protein-coding genes between thefelCat9 reference and the new single haplotype assemblies reported in this paper.

	Protein Coding Genes		
	felCat9	Fca-508	Pbe-53
Chromosome A1	1,247	1,253	1,244
Chromosome A2	1,591	1,600	1,599
Chromosome A3	1,131	1,135	1,133
Chromosome B1	966	974	969
Chromosome B2	1,077	1,069	1,000
Chromosome B3	1,187	1,204	1,194
Chromosome B4	1,225	1,232	1,220
Chromosome C1	1,737	1,760	1,742
Chromosome C2	898	905	904
Chromosome D1	1,448	1,467	1,466
Chromosome D2	631	640	632
Chromosome D3	661	668	666
Chromosome D4	798	812	809
Chromosome E1	1,096	1,099	1,099
Chromosome E2	1,107	1,108	1,106
Chromosome E3	691	692	689
Chromosome F1/E4	681	690	685
Chromosome F2	422	424	422
Chromosome X	781	799	796
ChrUn	219	38	82
Total	**19,594**	**19,569**	**19,457**
Difference	**0.00%**	**0.13%**	**0.70%**

Table C2.9. Incorrectly sorted read sequence content from the PHA analysis.

Cross	Total reads sorted	# incorrectly sorted reads	<50% repetitive	% reads <10kb	% reads <5kb
CatIII x Pbe-38	12,873,782	523,782	87.39%	79.06%	57.25%
Fca-508 x Pbe-14	12,873,782	362,052	88.53%	80.29%	58.31%
LilBub x Pbe-53	12,873,782	429,613	87.65%	79.83%	56.99%
LilBub x Pbe-6	12,873,782	519,606	90.95%	79.70%	57.93%

Table C2.10. Percentage of incorrectly sorted reads separated by subtype. (M) = Maternal haplotype, (P)=Paternal haplotype, (U)=Unknown haplotype.

Cross	Mean Read Length (bp)	M-to-P	P-to-M	U-to-M	U-to-P	M-to-U	P-to-U
LilBubxPbe-53	6,992	44.85 %	53.38%	0.47%	0.75%	0.32%	0.23%
Fca-508xPbe-14	6,961	27.40%	70.65%	1.12%	0.54%	0.12%	0.17%
LilBubxPbe-6	7,206	40.42%	57.55%	0.67%	0.55%	0.37%	0.44%
CatIIIxPbe-38	7,365	47.17%	50.84%	0.65%	0.59%	0.36 %	0.39%

Table C3.1. Gene ontology results for genes downregulated in the testis of sterileChausie hybrids. The top ten most enriched, significant, processes involve meiotic division or spermatogenesis.

GO biological process complete	Fold Enrichment	Raw P-value	FDR
Meiotic sister chromatid segregation	7.78	1.25E-04	9.92E-03
Meiotic sister chromatid cohesion	7.78	1.25E-04	9.87E-03
piRNA metabolic process	7.41	9.00E-06	1.05E-03
Inner dynein arm assembly	7.3	2.54E-05	2.51E-03
Meiosis II	7	2.05E-04	1.46E-02
Meiosis II cell cycle process	7	2.05E-04	1.46E-02
DNA methylation involved in gamete generation	6.25	6.37E-05	5.55E-03
Epithelial cilium movement involved in extracellular fluid movement	5.71	2.16E-06	3.07E-04
Axonemal dynein complex assembly	5.63	2.45E-07	4.25E-05
Synaptonemal complex organization	5.53	2.86E-05	2.78E-03

Table C3.2. Raw sequencing output by platform for the Jungle cat assembly.

Library	Reads	Bases (bp)	Genome Coverage	subread N50 (bp)	Avg. Read Length (bp)
PacBio (Sequel)	7,594,421	122,681,343,250	49x	25,928	16,154
Illumina (150bp PE)	278,636,871	83,591,061,300	33x		
Hi-C (*in situ* DNase)	1,541,076,320	462,322,896,000	185x		

Table C3.3. Interspecific *P*-distance for *DXZ4* repeat A (RA) units with and without outlier RA-1, spacer sequence and repeat B (RB) units. *P*-distance was calculated for shared alignment sites only. Results indicate increased divergence across both RA and RB relative to the spacer sequence.

Interspecific Mean *P*-Distance

RA

	Domestic	Jungle
Domestic Cat		
Jungle Cat	0.0276	
Asian Leopard Cat	0.0401	0.0429

RA excluding RA-1

	Domestic	Jungle
Domestic Cat		
Jungle Cat	0.0165	
Asian Leopard Cat	0.0296	0.0320

Spacer

	Domestic	Jungle
Domestic Cat		
Jungle Cat	0.0103	
Asian Leopard Cat	0.0181	0.0176

RB

	Domestic	Jungle
Domestic Cat		
Jungle Cat	0.0203	
Asian Leopard Cat	0.0266	0.0304

Table C3.4. Repeat unit length summary for *DXZ4* tandem arrays from the Jungle cat *de novo*, domestic and Asian leopard cat single haplotype assemblies.

Species	Repeat A	Length (bp)
Fca	RA-1	4,481
Fca	RA-2	4,544
Fca	RA-3	4,544
Fca	RA-4	4,544
Fca	RA-5	4,544
Fca	RA-6	4,576
Fca	RA-7	4,544
Fca	RA-8	4,512
Fca Average (no RA-1)		**4,544**
Fch	RA-1	4,529
Fch	RA-2	4,576
Fch	RA-3	4,624
Fch	RA-4	4,576
Fch	RA-5	4,608
Fch	RA-6	4,608
Fch	RA-7	4,608
Fch	RA-8	4,608
Fch Average (no RA-1)		**4,601**
Pbe	RA-1	4,545
Pbe	RA-2	4,512
Pbe	RA-3	4,544
Pbe	RA-4	4,512
Pbe	RA-5	4,512
Pbe	RA-6	4,544
Pbe Average (no RA-1)		**4,525**
RA Average		**4,554**
RA StDev		**39**

Species	Repeat B	Length (bp)
Fca	RB-1	4,494
Fca	RB-2	4,554
Fca	RB-3	4,584
Fca	RB-4	4,584
Fca	RB-5	4,615
Fca	RB-6	4,584
Fca	RB-7	4,615
Fca	RB-8	4,615
Fca	RB-9	4,584
Fca	RB-10	4,614
Fca	RB-11	4,615
Fca	RB-12	4,615
Fca Average		**4,589**
Fch	RB-1	4,616
Fch	RB-2	4,590
Fch	RB-3	4,640
Fch	RB-4	4,589
Fch	RB-5	4,615
Fch	RB-6	4,615
Fch Average		**4,611**
Pbe	RB-1	4,673
Pbe	RB-2	4,630
Pbe	RB-3	4,651
Pbe	RB-4	4,651
Pbe Average		**4,651**
RB Average		**4,607**
RB StDev		**36**

Table C3.5. *DXZ4 in silico* copy number estimates.

		RA	RB	Total	Genome Coverage
Domestic Cats					
Fca-508 SHA*	Assembly	9	13	22	-
Iraq	Domestic Shorthair Outbred	28	9	37	26x
Mateo	Domestic Shorthair Outbred	12	12	24	35x
Danny Boy	Domestic Shorthair Outbred	9	13	22	36x
Portugal	Domestic Shorthair Outbred	7	7	14	24x
Sizzle	Domestic Shorthair Outbred	4	7	11	23x
Thailand	Domestic Shorthair Outbred	2	5	7	26x
Loki	Domestic Shorthair Outbred	1	4	5	50x
Outbred Average		**9**	**8**	**17**	**-**
Outbred SD		**8.6**	**3.1**	**10.4**	**-**
Rocket	Maine Coon	7	9	16	29x
Tennessee	Tennessee Rex	4	8	12	29x
Speckles	Peterbald	4	6	10	35x
Gannon	Egyptian Mau	3	8	11	28x
Marcus	Persian	2	7	9	36x
Breed Average		**4**	**8**	**12**	**-**
Breed SD		**1.7**	**1.0**	**2.4**	**-**
Domestic Average		**7**	**8**	**15**	**31x**
Domestic SD		**6.8**	**2.8**	**8.5**	**7.1x**
Jungle Cats					
Fch-1a (FelCha1.0)	Assembly	9	7	16	-
Fch-12		5	3	8	18x
Average		**7**	**5**	**12**	**18x**
Asian Leopard Cats					
Pbe-53 SHA*	Assembly	7	5	12	-
Pbe-38		6	3	9	13x
Average		**7**	**4**	**11**	**13x**

Table C3.6. Differential methylation of *DXZ4* across 8 felid testis or sorted germ cell samples. (Abbreviations: MF, methylation frequency)

Phenotype	Felid group	Sample ID	*DXZ4* gene Mean MF
Fertile	Domestic cat	Pachytene	0.181
	Domestic cat	Spermatid	0.151
	Domestic cat	FCA-4048	0.260
	Domestic cat	FCA-4415	0.268
	Chausie	JXD-049	0.246
	Chausie	JXD-080	0.262
Sterile	Chausie	JXD-019	0.302
	Chausie	JXD-061	0.322

Table C3.7. Windows with significant differential methylation of testes in *DXZ4* Repeat Array A on chromosome X between fertile (n, domestic cat=2, Chausie=2) and sterile (n,Chausie=2) felids. Significance was determined by 1-tailed t-test of unequal variance.

Window start	Window stop	*p-value*
94,185,541	94,185,607	0.0388
94,185,567	94,185,797	0.0485
94,185,601	94,185,817	0.0054
94,185,788	94,185,829	2.95×10^{-7}
94,185,816	94,185,839	7.30×10^{-5}
94,185,828	94,185,851	1.25×10^{-3}
94,185,838	94,185,866	6.26×10^{-4}
94,185,849	94,185,903	9.84×10^{-4}
94,185,865	94,185,952	5.22×10^{-3}
94,185,900	94,186,070	0.0231
94,185,939	94,186,117	7.23×10^{-3}
94,186,069	94,186,173	0.0133
94,186,114	94,186,200	9.92×10^{-3}
94,186,171	94,186,225	3.51×10^{-3}
94,186,199	94,186,399	3.92×10^{-3}
94,186,224	94,186,437	0.0297
94,186,423	94,186,481	0.0472
94,186,453	94,186,506	0.0211
94,186,923	94,186,975	0.0480
94,186,943	94,187,001	0.0134
94,186,972	94,187,487	0.0106
94,186,998	94,187,500	0.0210
94,187,910	94,188,198	0.0368
94,188,176	94,188,233	0.0485
94,188,297	94,188,653	1.33×10^{-3}
94,188,316	94,188,762	3.77×10^{-5}
94,188,651	94,188,790	1.02×10^{-3}
94,189,268	94,189,317	0.0422
94,189,291	94,189,317	0.0176

Table C3.8. SRA accessions of individuals used for *in silico DXZ4* copy number estimations.

Species	Identifier	SRA	Sex	Ancestry
Domestic Cat	Sizzle	SRR5055407	Male	Established Breed
Domestic Cat	Loki	SRR5055386	Male	Established Breed
Domestic Cat	Rocket	SRR5051106	Male	Established Breed
Domestic Cat	Gannon	SRR5051108	Male	Established Breed
Domestic Cat	Marcus	SRR2224864	Male	Established Breed
Domestic Cat	Flowmaster	SRR5051112	Male	Established Breed
Domestic Cat	TennesseeTom	SRR5051114	Male	Established Breed
Domestic Cat	Speckles	SRR5051122	Male	Established Breed
Domestic Cat	Fca-508	SRR12914279	Female	Domestic Shorthair Outbred
Domestic Cat	Mateo	SRR11392568	Male	Domestic Shorthair Outbred
Domestic Cat	DannyBoy	SRR11392571	Male	Domestic Shorthair Outbred
Domestic Cat	Portugal-350	SRR5040114	Male	Domestic Shorthair Outbred
	Portugal-550	SRR5040108		
Domestic Cat	Iraq-350	SRR5040113	Male	Domestic Shorthair Outbred
	Iraq-550	SRR5040123		
Domestic Cat	Thailand-350	SRR5040120	Male	Domestic Shorthair Outbred
	Thailand-550	SRR5040112		
Jungle Cat	Fch-1a	SRR13340505	Male	
Jungle Cat	Fch-12	SRR2062187	Female	
Asian Leopard Cat	Pbe-53	SRR12914278	Male	
Asian Leopard Cat	Pbe-38	SRR4426179	Female	

Table C3.9. Meta-data for each of the felid testis samples included in this study.

Sample ID	Phenotype	Age at neuter (years)	Breeder estimated Jungle cat ancestry
Whole testis tissue			
FCA-4415	Fertile Domestic Cat	3	
FCA-4048	Fertile Domestic Cat	2.5	
JXD-019	Sterile Chausie	1.5	14%
JXD-049	Fertile Chausie	1.75	13%
JXD-061	Sterile Chausie	2	14%
JXD-080	Fertile Chausie	3	14%
Germ cells			
Pachytene	Fertile Domestic Cat	1.5	
Spermatid	Fertile Domestic Cat	1.5	

Table C3.10. Number of RRBS raw and uniquely mapped reads, percent mappability (M) to the reference genome, bisulfite conversion (BSconv) proportions, percent of cytosines within each methylation motif (mCG, mCHG, mCHH), and methylationfrequency (MF) across all chromosomes.

Sample ID	Batch	Raw reads	Uniquely aligned reads	M	BSconv	mCG	mCHG	mCHH	MF*
FCA-4415	1	34,135,894	28,227,690	82.7	0.94	77.7	11	9.8	0.231
FCA-4048	1	33,582,352	27,711,201	82.5	0.94	78.1	9.8	8.7	0.228
JXD-019	1	38,683,216	30,015,515	77.6	0.95	72.3	8.8	7.9	0.207
JXD-049	1	34,555,044	27,925,207	80.8	0.94	77	12.4	11.3	0.228
JXD-061	1	32,703,912	24,615,157	75.3	0.94	74.9	11.9	10.9	0.219
JXD-080	1	31,531,246	25,114,703	79.7	0.94	74.9	11	9.8	0.227
Pachytene	2	27,264,057	16,041,684	58.8	0.97	64.8	5.1	4.3	--
Spermatids	2	28,818,849	16,917,143	58.7	0.97	59.6	3.1	2.6	--

*After filtering sites for 10x coverage and uniting all testes data

Table C4.1. Raw sequencing output and haplotyping results.

	F1 Safari Cat	Domestic Haplotype	Geoffroy's Haplotype	F1 Liger	Tiger Haplotype	Lion Haplotype
PacBio						
Reads	18,607,818	8,218,109	10,389,709	21,059,837	10,489,759	10,570,078
Bases (bp)	395,174,858,902	182,737,769,633	212,437,089,269	383,502,214,277	194,491,582,536	189,010,631,741
Haplotype	100%	46%	54%	100%	51%	49%
Coverage	158x	73x	85x	153x	78x	76x
subread N50	31,629	32,222	31,036	28,253	28,447	28,059
Average Read Length	21,199	22,236	20,447	18,151	18,541	17,882
Illumina Short Read		***	***			
Reads	n/a	473,347,659	434,480,212	n/a	218,145,036	230,689,660
Bases		142,004,297,700	108,620,053,000		65,443,510,800	69,206,898,000
Coverage		57x	43x		26x	28x
HiC					***	***
Reads	651,877,707	361,178,387	368,313,790	n/a	119,690,697	255,801,455
Bases	195,563,312,100	108,353,516,100	110,494,137,000		35,907,209,100	76,740,436,500
Coverage	78x	43x	44x		14x	31x

***SRA Data

Table C4.2. Lion Y-linked contigs identified using BLAST and female read coverage. Ctg000044 captures the pseudoautosomal and single copy region. Annotations were mapped to each contig using liftover from the domestic cat published Y chromosome sequence.

Contig	Scaffold	Length (bp)	Annotations
ctg000044	chrY	8,609,702	*PAR, TETY2, UTY, DDX3Y, USP9Y, Cyorf_orig, AMELY, EIF2S3Y, ZFY, EIF1AY, RPS4Y-A/B/C, HSFY*
ctg000061	chrY_random_scaffold_78	669,012	None
ctg000071	chrY_random_scaffold_79	504,101	None
ctg000060	chrY_random_scaffold_80	459,065	*Cyorf_orig*
ctg000049	chrY_random_scaffold_81	432,664	None
ctg000088	chrY_random_scaffold_82	393,145	None
ctg000087	chrY_random_scaffold_83	285,082	None
ctg000025	chrY_random_scaffold_84	268,049	SRY
ctg000018	chrY_random_scaffold_85	267,247	None
ctg000006	chrY_random_scaffold_86	247,052	*SMCY, RPS4Y-A/B/C*
ctg000045	chrY_random_scaffold_87	199,686	None
ctg000024	chrY_random_scaffold_88	199,165	*SRY, RPS4Y-B*
ctg000050	chrY_random_scaffold_89	194,577	None
ctg000023	chrY_random_scaffold_90	136,774	None
ctg000048	chrY_random_scaffold_91	55,398	None

Table C4.3. Domestic cat genome assembly derived from F1 Safari Cat hybrid.

Molecule	Total Length (bp)	Gaps	Ungapped Length (bp)
ALL	**2,425,730,028**	**39**	**2,425,722,928**
Chromosome A1	239,367,248	1	239,367,148
Chromosome A2	169,388,855	7	169,387,755
Chromosome A3	140,443,288	0	140,443,288
Chromosome B1	205,367,284	2	205,367,084
Chromosome B2	151,959,158	0	151,959,158
Chromosome B3	148,491,486	3	148,490,386
Chromosome B4	142,168,536	0	142,168,536
Chromosome C1	221,611,373	3	221,611,073
Chromosome C2	158,578,461	1	158,578,361
Chromosome D1	115,366,949	1	115,366,449
Chromosome D2	88,083,857	0	88,083,857
Chromosome D3	94,435,393	3	94,435,093
Chromosome D4	95,154,158	2	95,153,958
Chromosome E1	61,876,196	4	61,875,796
Chromosome E2	61,988,844	1	61,988,744
Chromosome E3	41,437,797	1	41,437,697
Chromosome F1	69,239,673	1	69,239,573
Chromosome F2	83,466,477	0	83,466,477
Chromosome X	127,842,892	9	127,840,392
Unlocalized (n=39)	6,905,449	0	6,905,449
Unplaced (n=12)	2,556,654	0	2,556,654

Table C4.4. Geoffroy's cat genome assembly derived from F1 Safari Cat hybrid.

Molecule	Total Length (bp)	Gaps	Ungapped Length (bp)
ALL	**2,426,370,816**	**45**	**2,426,362,316**
Chromosome A1	239,694,388	1	239,694,288
Chromosome A2	169,709,481	6	169,708,081
Chromosome A3	140,630,183	1	140,630,083
Chromosome B1	205,836,458	1	205,836,358
Chromosome B2	152,606,360	1	152,605,860
Chromosome B3	148,130,213	0	148,130,213
Chromosome B4	142,838,203	3	142,837,503
Chromosome C1	222,052,948	3	222,052,648
Chromosome C2	158,996,348	0	158,996,348
Chromosome C3	153,706,148	1	153,705,648
Chromosome D1	115,783,437	4	115,782,637
Chromosome D2	88,472,993	5	88,472,493
Chromosome D3	94,884,351	5	94,883,451
Chromosome D4	94,461,142	0	94,461,142
Chromosome E1	61,591,894	1	61,591,394
Chromosome E2	62,031,975	0	62,031,975
Chromosome E3	42,047,855	3	42,047,555
Chromosome X	128,930,408	10	128,928,608
Unlocalized (n=27)	3,937,321	0	3,937,321
Unplaced (n=1)	28,710	0	28,710

Table C4.5. Tiger genome assembly derived from F1 Liger hybrid.

Molecule	Total Length (bp)	Gaps	Ungapped Length (bp)
ALL	**2,408,678,698**	**65**	**2,408,668,598**
Chromosome A1	238,220,823	1	238,220,323
Chromosome A2	167,824,101	4	167,823,301
Chromosome A3	139,420,020	1	139,419,520
Chromosome B1	204,240,962	3	204,240,662
Chromosome B2	151,172,648	3	151,172,348
Chromosome B3	146,942,463	4	146,941,663
Chromosome B4	140,883,259	1	140,883,159
Chromosome C1	219,880,392	4	219,879,992
Chromosome C2	157,485,960	1	157,485,860
Chromosome D1	114,653,560	3	114,652,860
Chromosome D2	86,999,983	1	86,999,483
Chromosome D3	94,015,955	9	94,014,655
Chromosome D4	93,600,330	0	93,600,330
Chromosome E1	61,289,036	7	61,287,936
Chromosome E2	61,603,977	0	61,603,977
Chromosome E3	41,508,437	3	41,508,137
Chromosome F2	82,905,707	1	82,905,607
Chromosome F3	69,613,107	2	69,612,907
Chromosome X	127,706,282	16	127,704,682
Unlocalized (n=52)	8,483,976	1	8,483,476
Unplaced (n=3)	227,720	0	227,720

Table C4.6. Lion genome assembly derived from F1 Liger hybrid.

Molecule	Total Length (bp)	Gaps	Ungapped Length (bp)
ALL	**2,297,552,363**	**55**	**2,297,542,863**
Chromosome A1	238,914,584	3	238,913,484
Chromosome A2	167,961,581	5	167,960,681
Chromosome A3	139,988,245	1	139,987,745
Chromosome B1	205,061,116	3	205,060,816
Chromosome B2	151,445,352	5	151,444,452
Chromosome B3	147,402,474	7	147,401,774
Chromosome B4	141,504,389	1	141,503,889
Chromosome C1	220,622,163	4	220,621,763
Chromosome C2	158,330,212	2	158,330,012
Chromosome D1	115,343,376	3	115,343,076
Chromosome D2	87,363,192	3	87,362,892
Chromosome D3	94,488,104	6	94,487,504
Chromosome D4	94,013,389	1	94,012,889
Chromosome E1	61,380,078	4	61,379,678
Chromosome E2	61,787,680	0	61,787,680
Chromosome E3	41,728,670	4	41,727,870
Chromosome F2	83,205,205	1	83,205,105
Chromosome F3	69,419,297	0	69,419,297
Chromosome Y	8,609,702	0	8,609,702
Unlocalized (n=30)	8,253,736	1	8,253,236
Unplaced (n=4)	729,818	1	729,318

Table C4.7. *P*-Distance relationships for RA and RB monomers, between and within species.

RA Mean Group Distance (10% stripped)							**Within Species**	**Within RA/RB**	**Between RA/RB**
	Fca126	Fca508	Fch	Pbe	Oge	Pti			
Fca126	**0.0114**								
Fca508	0.0113	**0.0117**					**0.0120 +/- 0.0012**	**0.0372**	**0.67**
Fch	0.0255	0.0259	**0.0141**						
Pbe	0.0362	0.0361	0.0384	**0.0110**					
Oge	0.0443	0.0441	0.0452	0.0446	**0.0127**				
Pti	0.0533	0.0532	0.0546	0.0544	0.0544	**0.0107**			
RA Mean Group Distance, Excluding RA-1 (10% stripped)									
	Fca126	Fca508	Fch	Pbe	Oge	Pti			
Fca126	**0.0003**								
Fca508	0.0013	**0.0000**					**0.0010 +/- 0.0006**	**0.0269**	**0.67**
Fch	0.0157	0.016	**0.0020**						
Pbe	0.0271	0.027	0.0291	**0.0000**					
Oge	0.0354	0.035	0.0360	0.036	**0.0008**				
Pti	0.0450	0.045	0.0460	0.046	0.046	**0.0008**			
Spacer Between Mean Group Distance (10% stripped)									
	Fca126	Fca508	Fch	Pbe	Oge	Pti			
Fca126									
Fca508	0.0017								
Fch	0.0106	0.0101							
Pbe	0.0180	0.0175	0.0170						
Oge	0.0181	0.0178	0.0174	0.0190					
Pti	0.0244	0.0240	0.0236	0.0249	0.0224				
RB Mean Group Distance (10% stripped)									
	Fca126	Fca508	Fch	Pbe	Oge	Pti			
Fca126	**0.0017**								
Fca508	0.0015	**0.0010**					**0.0015 +/- 0.0007**	**0.0255**	**0.67**
Fch	0.0212	0.0212	**0.0023**						
Pbe	0.0266	0.0264	0.0302	**0.0022**					
Oge	0.0270	0.0269	0.0308	0.0250	**0.0006**				
Pti	0.0394	0.0395	0.0413	0.0372	0.0385	**0.0013**			

Table C4.8. Felid RA monomer lengths across species.

Species	Repeat Unit	Length (bp)
Fca126	RA-1	4,525
Fca126	RA-2	4,536
Fca126	RA-3	4,524
Fca126	RA-4	4,536
Fca126	RA-5	4,581
Fca126	RA-6	4,502
Fca126	RA-7	4,513
Fca126	RA-8	4,513
Average		**4,529**
StDev		**23**
Fca508	RA-1	4,481
Fca508	RA-2	4,544
Fca508	RA-3	4,544
Fca508	RA-4	4,544
Fca508	RA-5	4,544
Fca508	RA-6	4,576
Fca508	RA-7	4,544
Fca508	RA-8	4,512
Average		**4,536**
StDev		**26**
Fch	RA-1	4,529
Fch	RA-2	4,576
Fch	RA-3	4,624
Fch	RA-4	4,576
Fch	RA-5	4,608
Fch	RA-6	4,608
Fch	RA-7	4,608
Fch	RA-8	4,608
Average		**4,592**
StDev		**29**

Table C4.8. Continued

Species	Repeat Unit	Length (bp)
Pbe	RA-1	4,545
Pbe	RA-2	4,512
Pbe	RA-3	4,544
Pbe	RA-4	4,512
Pbe	RA-5	4,512
Pbe	RA-6	4,544
Average		**4,528**
StDev		**16**
Oge	RA-1	4,590
Oge	RA-2	4,588
Oge	RA-3	4,513
Oge	RA-4	4,589
Oge	RA-5	4,551
Oge	RA-6	4,588
Oge	RA-7	4,589
Oge	RA-8	4,588
Average		**4,575**
StDev		**26**
Pti	RA-1	4,746
Pti	RA-2	4,772
Pti	RA-3	5,021
Pti	RA-4	4,938
Pti	RA-5	4,744
Pti	RA-6	4,911
Pti	RA-7	4,938
Pti	RA-8	4,772
Pti	RA-9	4,800
Average		**4,849**
StDev		**97**

Table C4.9. Felid RB monomer lengths across species.

Species	Repeat Unit	Length (bp)
Fca126	RB-1	4,616
Fca126	RB-2	4,570
Fca126	RB-3	4,616
Fca126	RB-4	4,615
Fca126	RB-5	4,570
Fca126	RB-6	4,683
Fca126	RB-7	4,661
Fca126	RB-8	4,683
Average		**4,627**
StDev		**42**
Fca508	RB-1	4,494
Fca508	RB-2	4,554
Fca508	RB-3	4,584
Fca508	RB-4	4,584
Fca508	RB-5	4,615
Fca508	RB-6	4,584
Fca508	RB-7	4,615
Fca508	RB-8	4,615
Fca508	RB-9	4,584
Fca508	RB-10	4,614
Fca508	RB-11	4,615
Fca508	RB-12	4,615
Average		**4,589**
StDev		**35**
Fch	RB-1	4,616
Fch	RB-2	4,590
Fch	RB-3	4,640
Fch	RB-4	4,589
Fch	RB-5	4,615
Fch	RB-6	4,615
Average		**4,611**
StDev		**17**

Table C4.9. Continued

Species	Repeat Unit	Length (bp)
Pbe	RB-1	4,673
Pbe	RB-2	4,630
Pbe	RB-3	4,651
Pbe	RB-4	4,651
Average		**4,651**
StDev		**15**
Oge	RB-1	4,673
Oge	RB-2	4,672
Oge	RB-3	4,638
Oge	RB-4	4,639
Oge	RB-5	4,672
Oge	RB-6	4,639
Oge	RB-7	4,672
Oge	RB-8	4,640
Oge	RB-9	4,671
Oge	RB-10	4,639
Average		**4,656**
StDev		**17**
Pti	RB-1	4,716
Pti	RB-2	4,675
Pti	RB-3	4,675
Pti	RB-4	4,675
Pti	RB-5	4,715
Pti	RB-6	4,715
Pti	RB-7	4,675
Pti	RB-8	4,715
Average		**4,695**
StDev		**20**

Table C4.10. Tiger copy number variation in the RA embedded GGGAA microsatellite.

Tiger Monomer	Copies	Length (bp)
RA-1	66	330
RA-2	75	375
RA-3	124	620
RA-4	107	535
RA-5	64	320
RA-6	99	495
RA-7	108	540
RA-8	70	350
RA-9	75	375
Average	**88**	**438**
StDev	**21**	**104**

Table C4.11. Lengths of spacer sequence within each species.

Species	Length (bp)
Fca126	32,303
Fca508	32,152
Fch	34,978
Pbe	34,574
Oge	33,974
Pti	34,925
Average	**34,136**
StDev	**1,019**
StDev Excluding Domestic Cats	**400**

Table C4.12. All monomers annotated and aligned to generate the maximum likelihood phylogeny shown in Figure 5 (Un-collapsed Supplementary Figure 18).

Order	Species	ID	Start	End	Length (bp)	CTCF
Canidae	Dog	ru-1	90185723	90181179	4545	2
Canidae	Dog	ru-2	90190219	90185724	4496	2
Canidae	Dog	ru-3	90268730	90273211	4482	2
Canidae	Dog	ru-4	90264233	90268729	4497	2
Average Length ± SD					**4505 ± 24**	
Cetartiodactyla	Pig	ru-1	95026435	95033117	6683	2
Cetartiodactyla	Pig	ru-2	95033118	95039799	6682	2
Cetartiodactyla	Pig	ru-3	95039800	95046399	6600	2
Cetartiodactyla	Pig	ru-5	95093328	95099958	6631	2
Cetartiodactyla	Pig	ru-6	95099959	95106589	6631	2
Cetartiodactyla	Pig	ru-7	95106590	95113274	6685	2
Cetartiodactyla	Pig	ru-8	95113275	95119959	6685	2
Cetartiodactyla	Pig	ru-9	95119960	95126643	6684	2
Average Length ± SD					**6660 ± 32**	
Chiroptera	Molossus	ru-1	44793189	44798116	4928	2
Chiroptera	Molossus	ru-2	44798117	44803044	4928	2
Chiroptera	Molossus	ru-3	44803045	44807971	4927	2
Chiroptera	Molossus	ru-4	44807972	44812898	4927	2
Chiroptera	Molossus	ru-5	44812899	44817826	4928	2
Chiroptera	Molossus	ru-8	44931150	44926260	4891	2
Chiroptera	Molossus	ru-9	44936103	44931151	4953	2
Chiroptera	Molossus	ru-10	44940963	44936104	4860	2
Chiroptera	Molossus	ru-11	44945886	44940964	4923	2
Average Length ± SD					**4918 ± 25**	
Chiroptera	Myotis myotis	ru-1	45192790	45189004	3787	2
Chiroptera	Myotis myotis	ru-2	45189003	45185247	3757	2
Average Length ± SD					**3772 ± 15**	
Felidae	Domestic Cat-126	RA-1	94454159	94449635	4525	1
Felidae	Domestic Cat-126	RA-2	94458695	94454160	4536	1
Felidae	Domestic Cat-126	RA-3	94463219	94458696	4524	1
Felidae	Domestic Cat-126	RA-4	94467755	94463220	4536	1

Felidae	Domestic Cat-126	RA-5	94472336	94467756	4581	1
Felidae	Domestic Cat-126	RA-6	94476838	94472337	4502	1
Felidae	Domestic Cat-126	RA-7	94481351	94476839	4513	1
Felidae	Domestic Cat-126	RA-8	94485864	94481352	4513	1
Felidae	Domestic Cat-126	RB-1	94522013	94526628	4616	2
Felidae	Domestic Cat-126	RB-2	94526629	94531198	4570	2
Felidae	Domestic Cat-126	RB-3	94531199	94535814	4616	2
Felidae	Domestic Cat-126	RB-4	94535815	94540429	4615	2
Felidae	Domestic Cat-126	RB-5	94540430	94544999	4570	2
Felidae	Domestic Cat-126	RB-6	94545000	94549682	4683	2
Felidae	Domestic Cat-126	RB-7	94549683	94554343	4661	2
Felidae	Domestic Cat-126	RB-8	94554344	94559026	4683	2
Average Length ± SD					**4578**	**60**
Felidae	Domestic Cat-508	RA-1	94189489	94185009	4481	1
Felidae	Domestic Cat-508	RA-2	94194033	94189490	4544	1
Felidae	Domestic Cat-508	RA-3	94198577	94194034	4544	1
Felidae	Domestic Cat-508	RA-4	94203121	94198578	4544	1
Felidae	Domestic Cat-508	RA-5	94207665	94203122	4544	1
Felidae	Domestic Cat-508	RA-6	94212241	94207666	4576	1
Felidae	Domestic Cat-508	RA-7	94216785	94212242	4544	1
Felidae	Domestic Cat-508	RA-8	94221297	94216786	4512	1
Felidae	Domestic Cat-508	RB-1	94257516	94262009	4494	2
Felidae	Domestic Cat-508	RB-2	94262010	94266563	4554	2
Felidae	Domestic Cat-508	RB-3	94266564	94271147	4584	2
Felidae	Domestic Cat-508	RB-4	94271148	94275731	4584	2

Felidae	Domestic Cat-508	RB-5	94275732	94280346	4615	2
Felidae	Domestic Cat-508	RB-6	94280347	94284930	4584	2
Felidae	Domestic Cat-508	RB-7	94284931	94289545	4615	2
Felidae	Domestic Cat-508	RB-8	94289546	94294160	4615	2
Felidae	Domestic Cat-508	RB-9	94294161	94298744	4584	2
Felidae	Domestic Cat-508	RB-10	94298745	94303358	4614	2
Felidae	Domestic Cat-508	RB-11	94303359	94307973	4615	2
Felidae	Domestic Cat-508	RB-12	94307974	94312588	4615	2
Average Length ± SD					**4568**	**41**
Felidae	Jungle Cat	RA-1	93374298	93369770	4529	1
Felidae	Jungle Cat	RA-2	93378874	93374299	4576	1
Felidae	Jungle Cat	RA-3	93383498	93378875	4624	1
Felidae	Jungle Cat	RA-4	93388074	93383499	4576	1
Felidae	Jungle Cat	RA-5	93392682	93388075	4608	1
Felidae	Jungle Cat	RA-6	93397290	93392683	4608	1
Felidae	Jungle Cat	RA-7	93401898	93397291	4608	1
Felidae	Jungle Cat	RA-8	93406506	93401899	4608	1
Felidae	Jungle Cat	RB-1	93444449	93449064	4616	2
Felidae	Jungle Cat	RB-2	93449065	93453654	4590	2
Felidae	Jungle Cat	RB-3	93453655	93458294	4640	2
Felidae	Jungle Cat	RB-4	93458295	93462883	4589	2
Felidae	Jungle Cat	RB-5	93462884	93467498	4615	2
Felidae	Jungle Cat	RB-6	93467499	93472113	4615	2
Average Length ± SD					**4600 ± 26**	
Felidae	Geoffroy's Cat	RA-1	95773101	95768512	4590	1

Felidae	Geoffroy's Cat	RA-2	95777689	95773102	4588	1
Felidae	Geoffroy's Cat	RA-3	95782202	95777690	4513	1
Felidae	Geoffroy's Cat	RA-4	95786791	95782203	4589	1
Felidae	Geoffroy's Cat	RA-5	95791342	95786792	4551	1
Felidae	Geoffroy's Cat	RA-6	95795930	95791343	4588	1
Felidae	Geoffroy's Cat	RA-7	95800519	95795931	4589	1
Felidae	Geoffroy's Cat	RA-8	95805107	95800520	4588	1
Felidae	Geoffroy's Cat	RB-1	95842965	95847637	4673	2
Felidae	Geoffroy's Cat	RB-2	95847638	95852309	4672	2
Felidae	Geoffroy's Cat	RB-3	95852310	95856947	4638	2
Felidae	Geoffroy's Cat	RB-4	95856948	95861586	4639	2
Felidae	Geoffroy's Cat	RB-5	95861587	95866258	4672	2
Felidae	Geoffroy's Cat	RB-6	95866259	95870897	4639	2
Felidae	Geoffroy's Cat	RB-7	95870898	95875569	4672	2
Felidae	Geoffroy's Cat	RB-8	95875570	95880209	4640	2
Felidae	Geoffroy's Cat	RB-9	95880210	95884880	4671	2
Felidae	Geoffroy's Cat	RB-10	95884881	95889519	4639	2
Average Length ± SD					**4620 ± 46**	
Felidae	Asian Leopard Cat	RA-1	95634730	95630186	4545	1
Felidae	Asian Leopard Cat	RA-2	95639242	95634731	4512	1
Felidae	Asian Leopard Cat	RA-3	95643786	95639243	4544	1
Felidae	Asian Leopard Cat	RA-4	95648298	95643787	4512	1

Felidae	Asian Leopard Cat	RA-5	95652810	95648299	4512	1
Felidae	Asian Leopard Cat	RA-6	95657354	95652811	4544	1
Felidae	Asian Leopard Cat	RB-1	95694532	95699204	4673	2
Felidae	Asian Leopard Cat	RB-2	95699205	95703834	4630	2
Felidae	Asian Leopard Cat	RB-3	95703835	95708485	4651	2
Felidae	Asian Leopard Cat	RB-4	95708486	95713136	4651	2
Average Length ± SD					**4577 ± 62**	
Felidae	Tiger	RA-1	94563438	94558693	4746	1
Felidae	Tiger	RA-2	94568210	94563439	4772	1
Felidae	Tiger	RA-3	94573231	94568211	5021	1
Felidae	Tiger	RA-4	94578169	94573232	4938	1
Felidae	Tiger	RA-5	94582913	94578170	4744	1
Felidae	Tiger	RA-6	94587824	94582914	4911	1
Felidae	Tiger	RA-7	94592762	94587825	4938	1
Felidae	Tiger	RA-8	94597534	94592763	4772	1
Felidae	Tiger	RA-9	94602334	94597535	4800	1
Felidae	Tiger	RB-1	94641394	94646109	4716	2
Felidae	Tiger	RB-2	94646110	94650784	4675	2
Felidae	Tiger	RB-3	94650785	94655459	4675	2
Felidae	Tiger	RB-4	94655460	94660134	4675	2
Felidae	Tiger	RB-5	94660135	94664849	4715	2
Felidae	Tiger	RB-6	94664850	94669564	4715	2
Felidae	Tiger	RB-7	94669565	94674239	4675	2
Felidae	Tiger	RB-8	94674240	94678954	4715	2
Average Length ± SD					**4777**	**105**
Perissodactyla	Horse	ru-1	95802564	95798294	4271	1
Perissodactyla	Horse	ru-2	95806878	95802565	4314	1
Perissodactyla	Horse	ru-3	95810708	95806879	3830	1
Perissodactyla	Horse	ru-4	95814978	95810709	4270	1
Average Length ± SD					**4171**	**198**

Primate	Human	RA-1	114136373	114133517	2857	1
Primate	Human	RA-2	114139329	114136375	2955	1
Primate	Human	RA-3	114142268	114139331	2938	1
Primate	Human	RA-4	114145240	114142270	2971	1
Primate	Human	RA-5	114148212	114145242	2971	1
Primate	Human	RA-6	114151184	114148214	2971	1
Primate	Human	RA-7	114154172	114151186	2987	1
Primate	Human	RA-8	114157128	114154174	2955	1
Primate	Human	RA-9	114160100	114157130	2971	1
Primate	Human	RA-10	114163055	114160102	2954	1
Primate	Human	RA-11	114166044	114163057	2988	1
Primate	Human	RA-12	114168999	114166046	2954	1
Primate	Human	RA-13	114171955	114169001	2955	1
Primate	Human	RA-14	114174910	114171957	2954	1
Primate	Human	RA-15	114177866	114174912	2955	1
Primate	Human	RA-16	114180854	114177868	2987	1
Primate	Human	RA-17	114183826	114180856	2971	1
Primate	Human	RA-18	114186782	114183828	2955	1
Primate	Human	RA-19	114189754	114186784	2971	1
Primate	Human	RA-20	114192726	114189756	2971	1
Primate	Human	RA-21	114195682	114192728	2955	1
Primate	Human	RA-22	114198621	114195684	2938	1
Primate	Human	RA-23	114201593	114198623	2971	1
Primate	Human	RA-24	114204565	114201595	2971	1
Primate	Human	RA-25	114207537	114204567	2971	1
Primate	Human	RA-26	114210525	114207539	2987	1
Primate	Human	RA-27	114213497	114210527	2971	1
Primate	Human	RA-28	114216485	114213499	2987	1
Primate	Human	RA-29	114219457	114216487	2971	1
Primate	Human	RA-30	114222429	114219459	2971	1
Primate	Human	RA-31	114225385	114222431	2955	1
Primate	Human	RA-32	114228357	114225387	2971	1
Primate	Human	RA-33	114231329	114228359	2971	1
Primate	Human	RA-34	114234301	114231331	2971	1
Primate	Human	RA-35	114237273	114234303	2971	1
Primate	Human	RA-36	114240245	114237275	2971	1
Primate	Human	RA-37	114243217	114240247	2971	1
Primate	Human	RA-38	114246189	114243219	2971	1
Primate	Human	RA-39	114249161	114246191	2971	1
Primate	Human	RA-40	114252116	114249163	2954	1
Primate	Human	RA-41	114255088	114252118	2971	1
Primate	Human	RA-42	114258044	114255090	2955	1
Primate	Human	RA-43	114261016	114258046	2971	1

Primate	Human	RA-44	114264004	114261018	2987	1
Primate	Human	RA-45	114266976	114264006	2971	1
Primate	Human	RA-46	114269948	114266978	2971	1
Primate	Human	RA-47	114272937	114269950	2988	1
Primate	Human	RA-48	114275909	114272939	2971	1
Primate	Human	RA-49	114278881	114275911	2971	1
Primate	Human	RA-50	114281853	114278883	2971	1
Primate	Human	RA-51	114284825	114281855	2971	1
Primate	Human	RA-52	114287797	114284827	2971	1
Primate	Human	RA-53	114290769	114287799	2971	1
Primate	Human	RA-54	114293741	114290771	2971	1
Primate	Human	RA-55	114296713	114293743	2971	1
Average Length ± SD					**2966 ± 19**	
Primate	Rhesus	ru-1	112598557	112595646	2912	1
Primate	Rhesus	ru-2	112601483	112598558	2926	1
Primate	Rhesus	ru-3	112604410	112601484	2927	1
Average Length ± SD					**2922 ± 7**	
Rodentia	Mouse	ru-1	74772474	74776367	3894	1
Rodentia	Mouse	ru-2	74776368	74781372	5005	2
Rodentia	Mouse	ru-3	74781373	74787165	5793	3
Rodentia	Mouse	ru-4	74787166	74792031	4866	2
Rodentia	Mouse	ru-5	74792032	74796851	4820	2
Rodentia	Mouse	ru-6	74796852	74800744	3893	1
Rodentia	Mouse	ru-7	74800745	74805980	5236	2
Average Length ± SD					**4787 ± 640**	
Rodentia	Rat	ru-1	111714420	111711555	2866	1
Rodentia	Rat	ru-2	111717306	111714421	2886	1
Rodentia	Rat	ru-3	111720171	111717307	2865	1
Rodentia	Rat	ru-4	111723056	111720172	2885	1
Rodentia	Rat	ru-5	111725922	111723057	2866	1
Rodentia	Rat	ru-7	111772230	111769352	2879	1
Rodentia	Rat	ru-8	111775108	111772231	2878	1
Rodentia	Rat	ru-9	111777987	111775109	2879	1
Rodentia	Rat	ru-10	111780833	111777988	2846	1
Average Length ± SD					**2872 ± 12**	

Table C4.13. Repeat unit features summary.

Species	Count	CTCF Profile	Average Length (bp)	SD (bp)	SD % of Length
Dog	4	RB-like	4505	24	0.53%
Pig	8	RB-like	6660	32	0.48%
Molossus	9	RB-like	4918	25	0.52%
Myotis	2	RB-like	3772	15	0.40%
Domestic Cat-126	16	Both	4578	60	1.30%
Domestic Cat-508	20	Both	4568	41	0.90%
Jungle Cat	14	Both	4600	26	0.57%
Geoffroy's Cat	18	Both	4620	46	0.99%
Asian Leopard Cat	10	Both	4577	62	1.36%
Tiger	17	Both	4777	105	2.21%
Horse	4	RA-like	4171	198	4.74%
Human	55	RA-like	2966	19	0.63%
Rhesus	3	RA-like	2922	7	0.23%
Mouse	7	Neither	4787	640	13.37%
Rat	9	RA-like	2872	12	0.42%

Table C4.14. Mean group *P*-distance based on 10% masked mafft alignment. Within-group represented along grey diagonal. Green shading identifies Felid monomer columns. Bolded values represent lowest *P*-distance between felid array and non-felid monomers.

Species	Human	Rhesus	Mouse	Rat	Horse	Dog	Felid-RA1	Felid-RA	Felid-RB	Pig	Bat-myomyo	Bat-molmol
Human	0.0016											
Rhesus	0.0927	0.0014										
Mouse	1.0665	1.0567	0.0183									
Rat	1.1219	1.1371	0.5113	0.0007								
Horse	1.0218	1.0506	1.0963	1.0648	0.2239							
Dog	0.9143	0.9437	1.0854	1.1128	0.9924	0.0051						
Felid-RA1	0.8067	0.8256	0.9673	**1.0482**	0.9714	0.5678	0.0413					
Felid-RA	**0.7664**	**0.7901**	**0.9382**	1.0539	0.9228	**0.5358**	0.0669	0.0183				
Felid-RB	0.7896	0.8035	0.9912	1.0605	**0.8765**	0.8495	0.5623	0.5066	0.0186			
Pig	0.5442	0.5697	1.0904	1.1751	1.0704	0.9529	0.9022	**0.8648**	0.8876	0.0033		
Bat-myomyo	0.8865	0.8814	1.0499	1.0642	1.0392	0.9645	0.8605	0.8222	**0.8176**	0.9385	0.0036	
Bat-molmol	0.9063	0.9262	1.034	1.1033	1.0242	0.9752	0.8652	**0.8316**	0.8702	0.9775	0.451	0.0624

Table C4.15. Mammalian Assemblies.

Common Name	Sex	Assembly	Year	Accession	Ref	Annotations	Data
Southern two-toed sloth	Male	mChoDid1.pri	2020	GCF_015220235.1	Yes	Yes	PacBio Sequel I CLR; Illumina NovaSeq; Arima Genomics Hi-C; Bionano Genomics DLS
African elephant	Female	Loxafr3.0	2009	GCF_000001905.1	Yes	Yes	Sanger, illumina
Human	Female	T2T	2021	GCA_009914755.2	No	Yes, LiftOver GRCh38.p13	PacBio Sequel II HiFi; Oxford Nanopore MinION; Illumina NovaSeq
Chimpanzee	Male	Clint_PTRv2	2018	GCF_002880755.1	Yes	Yes	?
Rhesus monkey	Male	rheMacS_1.0	2019	GCA_008058575.1	No	Yes, LiftOver Mmul_10	PacBio Sequel
Rabbit	Female	OryCun3.0	2020	GCA_013371645.1	No	Yes, LiftOver OryCun2.0	PacBio RSII; Illumina
Mouse	?	GRCm39	2020	GCA_000001635.9	Yes	Yes	?
Rat	Male	mRatBN7.2	2020	GCF_015227675.2	Yes	Yes	PacBio Sequel; 10X Genomics Chromium; BioNano; Arima Hi-C
Horse	Female	EquCab3.0	2018	GCF_002863925.1	Yes	Yes	Sanger; Illumina HiSeq; PacBio
Dog	Male	ROS_Cfam_1.0	2020	GCA_014441545.1	Yes	Yes	PacBio Sequel; Illumina

Domestic Cat-508	Trio-Maternal	Fcat_Pben_1.0_maternal_alt	2020	GCA_016509815.1	No	Yes, Liftover FelCat9	PacBio Sequel, NovaSeq, *in situ* Hi-C
Domestic Cat-126	Trio-Maternal	F.catus_Fca126_mat1.0	2021	GCA_018350175.1	No	Yes, Liftover FelCat9	PacBio Sequel II, NovaSeq, *in situ* Hi-C
Jungle Cat	Male	n/a	2021	n/a	na	Yes, Liftover FelCat9	PacBio Sequel, NovaSeq, *in situ* Hi-C
Leopard Cat	Trio-Paternal	Fcat_Pben_1.0_paternal_pri	2020	GCA_016509475.1	na	Yes, Liftover FelCat9	PacBio Sequel, NovaSeq, *in situ* Hi-C
Geoffroy's Cat	Trio-Paternal	O.geoffroyi_Oge1_pat1.0	2021	GCA_018350155.1	na	Yes, Liftover FelCat9	PacBio Sequel II, NovaSeq, *in situ* Hi-C
Tiger	Trio-Maternal	P.tigris_Pti1_mat1.0	2021	GCA_018350195.1	na	Yes, Liftover FelCat9	PacBio Sequel II, NovaSeq, *in situ* Hi-C
Pig	Male	Ninghe_Sus_1	2020	GCA_015776825.1	No	Yes, LiftOver Sscroffa11.1	Illumina HiSeq; PacBio RSII
Cow	Trio-Paternal	ARS_UNL_Btau-highland_paternal_1.0	2019	GCA_009493655.1	No	Yes, LiftOver from UCD1.2	PacBio Sequel; Illumina HiSeq
Yak	Trio-Maternal	ARS_UNL_BGru_maternal_1.0	2019	GCA_009493645.1	No	Yes, LiftOver from UCD1.2	PacBio Sequel; Illumina HiSeq
Goat	Male	Saanen_v1	2020	GCA_015443085.1	No	Yes, LiftOver ARS1	PacBio; Hi-C; Illumina
Sheep	Male	ASM1117029v1	2020	GCA_011170295.1	No	Yes	Oxford Nanopore PromethION; Illumina HiSeq

Myotis bat	Female	mMyoMyo1.p	2020	GCF_014108235.1	na	Yes	PacBio Sequel CLR; 10X Genomics chromium linked reads; Bionano Genomics; Phase Genomics HiC; PacBio Sequel IsoSeq
Pallas's mastiff bat	Male	mMolMol1.p	2020	GCF_014108415.1	na	Yes	PacBio Sequel CLR; 10X Genomics chromium linked reads; Bionano Genomics DLE1; Arima Genomics Hi-C kit; PacBio Sequel IsoSeq

Table C4.16. X Chromosome attributes and support for investigated species.

Organism	Scientific Name	Total Length (bp)	Total Length (Mb)	Relative Position (0-1)	Centromere Classification	Support
Sloth	*Choloepus didactylus*	193,839,925	194	?	Metacentric?	Atlas
Elephant	*Loxodonta africana*	Broken	Scaffold		Sub-metacentric	Atlas
Human	*Homo sapien*	154,259,625	154	0.39	Sub-metacentric	NCBI, Atlas
Chimpanzee	*Pan troglodytes*	151,576,176	152	0.38	Sub-metacentric	NCBI, Atlas, Human Comparison
Rhesus monkey	*Macaca mulatta*	152,195,021	152	0.45	Metacentric	NCBI, Atlas
Rabbit	*Oryctolagus cuniculus*	136,722,232	137	0.29	Sub-metacentric	NCBI
Mouse	*Mus musculus*	169,476,592	169	n/a	Acrocentric	NCBI, Atlas
Rat	*Rattus norvegicus*	152,453,651	152	n/a	Acrocentric	Atlas, Publication
Horse	*Equus caballus*	128,206,784	128	?	Sub-metacentric	Atlas, Publication
Dog	*Canis familiaris*	127,069,619	127	?	Metacentric	Atlas, Published
Domestic Cat 126	*Felis catus*	127,842,892	128	0.40	Sub-metacentric	Atlas
Domestic Cat 508	*Felis catus*	127,194,472	127	0.40	Sub-metacentric	Atlas
Jungle Cat	*Felis chaus*	126,453,223	126	?	Sub-metacentric	Atlas
Leopard Cat	*Prionailurus bengalensis*	129,104,405	129	?	Sub-metacentric	Atlas
Geoffroy's Cat	*Oncifelis geoffroyi*	128,930,408	129	?	Sub-metacentric	Atlas
Tiger	*Panthera tigris*	127,706,282	128	?	Sub-metacentric	Atlas
Pig	*Sus scrofa*	125,099,353	125	0.40	Sub-metacentric	Atlas, NCBI Gap
Cow	*Bos taurus*	136,391,934	136	0.28	Sub-metacentric	Atlas, Publication
Yak	*Bos grunniens*	132,421,397	132	?	Sub-metacentric	Atlas, Publication
Goat	*Capra hircus*	142,353,804	142	n/a	Acrocentric	Atlas, Publication
Sheep	*Ovis aries*	145,453,424	145	n/a	Acrocentric	Atlas, Publication
Pallas's Mastiff Bat	*Molossus molossus*	148,092,577	Scaffold?			
Myotis Bat	*Myotis myotis*	130,331,158	Scaffold?			

Table CA.1. Y-linked contigs identified using BLAST. Ctg000067-2 is the Y specific region originally included in ctg000067, which composed the pseudoautosomal region. ctg000067-2, ctg000078, and ctg000135 all contained genes belonging to the singlecopy region of the Y chromosome. These were manually scaffolded and identified in the final assembly as chrY. Annotations were mapped to each contig using liftover from the domestic cat published Y chromosome sequence.

Contig ID	Length (bp)	Annotations
ctg000067-2	206,618	*FLJ36031Y-d340, FLJ36031Y-e268*
ctg000078	564,111	*TETY2, UTY, DDX3Y, USP9Y*
ctg000135	482,115	*AMELY, EIF2S3Y, ZFY, EIF1AY, RPS4Y-A, RPS4Y-B, RPS4Y-C*
ctg000072	409,643	*LOC109496917, LOC111561459*
ctg000074	72,713	*TSPY-b274*
ctg000077	317,206	*TSPY-a244, TSPY-c264, Cyorf_fusion, SRY*
ctg000100	214,609	None
ctg000101	326,606	*FLJ36031Y-a260, FLJ36031Y-b321, FLJ36031Y-c254, FLJ36031Y-d340, FLJ36031Y-e268*
ctg000128	63,486	*TETY1*
ctg000136	229,506	*RPS4Y-A, RPS4Y-B, RPS4Y-C, HSFY*
ctg000137	214,959	*RPS4Y-A, RPS4Y-C, HSFY*
ctg000138	48,132	*RPS4Y-A, RPS4Y-B, RPS4Y-C, HSFY*
ctg000150	69,303	*Cyorf_orig*
ctg000171	256,672	*FLJ36031Y-d340, FLJ36031Y-e268*
Total Length	**3,475,679**	

Table CA.2. Jungle cat genome assembly (FelCha1.0).

Molecule	Total Length (bp)	Gaps	Ungapped Length (bp)
ALL	**2,428,287,114**	**57**	**2,428,281,414**
Chromosome A1	240,008,610	2	240,008,410
Chromosome A2	169,335,317	8	169,334,517
Chromosome A3	140,691,898	1	140,691,798
Chromosome B1	205,710,267	0	205,710,267
Chromosome B2	152,756,071	0	152,756,071
Chromosome B3	148,552,997	1	148,552,897
Chromosome B4	142,338,932	1	142,338,832
Chromosome C1	222,028,171	4	222,027,771
Chromosome C2	159,171,038	1	159,170,938
Chromosome D1	115,689,139	2	115,688,939
Chromosome D2	88,529,372	2	88,529,172
Chromosome D3	95,225,590	2	95,225,390
Chromosome D4	95,074,675	4	95,074,275
Chromosome E1	61,214,427	3	61,214,127
Chromosome E2	61,992,405	1	61,992,305
Chromosome E3	41,346,658	0	41,346,658
Chromosome F1	69,875,921	1	69,875,821
Chromosome F2	83,746,879	0	83,746,879
Chromosome X	126,453,223	19	126,451,323
Chromosome Y (SCR)	1,253,044	2	1,252,844
Chromosome Y Unlocalized (n=11)	2,222,635	0	2,222,635
Unplaced (n=21)	5,069,845	3	5,069,545

Table CA.3. RepeatMasker repeat analysis summary.

Total Sequence Masked	**817,518,325 (bp)**		**33.67%**
Elements	**Number of Elements**	**Length Occupied (bp)**	**Percentage**
SINEs	**469,392**	**69,228,429**	**2.85%**
Alu/B1	-	-	0.00%
MIRs	461,742	68,320,782	2.81%
LINEs	**820,170**	**461,379,223**	**19.00%**
LINE1	455,565	364,763,156	15.02%
LINE2	310,104	84,495,495	3.48%
L3/CR1	40,779	8,846,541	0.36%
RTE	12,488	3,076,198	0.13%
LRT elements	**285,101**	**108,386,633**	**4.46%**
ERVL	86,921	39,232,030	1.62%
ERVL-MaLRs	145,860	50,918,315	2.10%
ERV_classI	28,392	12,440,852	0.51%
ERV_classIII	-	-	0.00%
DNA elements	**342,818**	**68,433,349**	**2.82%**
hAT-Charlie	193,901	36,348,164	1.50%
TcMar-Trigger	53,706	14,407,736	0.59%
Unclassified	3,729	587,155	0.02%
Total interspersed Repeats		**708,014,789**	**29.16%**
small RNA	141,131	10,704,033	0.44%
Satellites	2	462	0.00%
Simple Repeats	1,465,939	70,199,559	2.89%
Low complexity	531,854	28,455,522	1.17%

Table CA.4. Summary of annotation liftover between the felCat9 reference and the Jungle cat assembly.

	Protein Coding Genes	
	felCat9	**Jungle Cat**
Chromosome A1	1,247	1,249
Chromosome A2	1,591	1,614
Chromosome A3	1,131	1,141
Chromosome B1	966	970
Chromosome B2	1,077	1,069
Chromosome B3	1,187	1,200
Chromosome B4	1,225	1,225
Chromosome C1	1,737	1,756
Chromosome C2	898	906
Chromosome D1	1,448	1,460
Chromosome D2	631	650
Chromosome D3	661	673
Chromosome D4	798	815
Chromosome E1	1,096	1,098
Chromosome E2	1,107	1,111
Chromosome E3	691	696
Chromosome F1	681	696
Chromosome F2	422	425
Chromosome X	781	792
Chromosome Y	21	13
ChrUn (n=32)	219	52
Total	**19,594**	**19,611**
Jungle Cat Copies	**203**	**1.04%**

Table CA.5. Additional Jungle cat gene copies identified by Liftoff relative to felCat9.0 with Y chromosome single copy region annotations (Li et al., 2013).

Gene Name	Extra Copies	Gene Name	Extra Copies
A_RPS4Y	3	LOC101101494	1
ARV1	2	LOC102901082	1
B_RPS4Y	2	LOC102902070	1
C_RPS4Y	5	LOC105259672	1
CD8A	1	LOC105259679	1
d340_FLJ36031Y	3	LOC105259841	1
e268_FLJ36031Y	3	LOC105260280	1
GIMAP2	1	LOC105260290	2
GIMAP6	1	LOC105260348	2
GP1BB	1	LOC105260366	6
HSFX4	1	LOC105260391	1
HSFY	2	LOC105260530	1
IFNW4	1	LOC105260634	1
LOC101080402	1	LOC105260966	1
LOC101080615	1	LOC105260993	1
LOC101080874	1	LOC105261116	1
LOC101081393	1	LOC109493961	1
LOC101081959	1	LOC109493962	1
LOC101083228	1	LOC109496557	5
LOC101083281	2	LOC109496917	22
LOC101084708	1	LOC109496987	1
LOC101085613	1	LOC111556525	36
LOC101089503	2	LOC111556663	1
LOC101089637	1	LOC111556757	19
LOC101089971	1	LOC111556934	4
LOC101090249	1	LOC111557541	9
LOC101090443	1	LOC111558267	1
LOC101090556	1	LOC111558710	1
LOC101091647	1	LOC111558783	3
LOC101091653	1	LOC111560404	4
LOC101091766	1	LOC111561459	4
LOC101094994	1	LOC111561460	7
LOC101097517	1	LOC111561651	1
LOC101097790	1	LOC111561809	1
LOC101098529	1	MZT2B	1
LOC101099158	1	PNMA6A	1
LOC101100737	1	SEPT5	1
LOC101101043	1	SMPD4	1
LOC101101243	1	TMEM211	1

Table CA.6. Assemblytics variant analysis comparing the Jungle cat assembly to the Fca-508 single haplotype domestic cat assembly.

Variant	Count	Total (bp)
Insertion		
50-500	18,943	3,647,523
500-10,000	1,016	2,708,148
Total	19,959	6,355,671
Deletion		
50-500	16,586	3,011,021
500-10,000	863	1,945,977
Total	17,449	4,956,998
Tandem Expansion		
50-500	1,647	336,364
500-10,000	271	446,894
Total	1,918	783,258
Tandem Contraction		
50-500	1,366	296,395
500-10,000	242	296,372
Total	1,608	592,767
Repeat Expansion		
50-500	11,526	3,390,482
500-10,000	8,371	15,139,307
Total	19,897	18,529,789
Repeat Contraction		
50-500	9,592	2,692,419
500-10,000	6,595	11,673,785
Total	16,187	14,366,204
Total structural variants		**77,018**
Total affected bases		**45.59 Mb**
Total Gain		**25.67 Mb**
Total Loss		**19.92 Mb**
Difference		**5.75 Mb**