nature genetics

Article

https://doi.org/10.1038/s41588-023-01548-y

Single-haplotype comparative genomics provides insights into lineage-specific structural variation during cat evolution

Received: 23 June 2023

Accepted: 20 September 2023

Published online: 02 November 2023

Check for updates

Kevin R. Bredemeyer[1,2,10], LaDeana Hillier[3,10], Andrew J. Harris[1,2,10], Graham M. Hughes[4], Nicole M. Foley[1], Colleen Lawless[4], Rachel A. Carroll[5], Jessica M. Storer[6], Mark A. Batzer[7], Edward S. Rice[5], Brian W. Davis[1,2], Terje Raudsepp[1,2], Stephen J. O'Brien[8], Leslie A. Lyons[9], Wesley C. Warren[5] ✉ & William J. Murphy[1,2] ✉

The role of structurally dynamic genomic regions in speciation is poorly understood due to challenges inherent in diploid genome assembly. Here we reconstructed the evolutionary dynamics of structural variation in five cat species by phasing the genomes of three interspecies F1 hybrids to generate near-gapless single-haplotype assemblies. We discerned that cat genomes have a paucity of segmental duplications relative to great apes, explaining their remarkable karyotypic stability. X chromosomes were hotspots of structural variation, including enrichment with inversions in a large recombination desert with characteristics of a supergene. The X-linked macrosatellite *DXZ4* evolves more rapidly than 99.5% of the genome clarifying its role in felid hybrid incompatibility. Resolved sensory gene repertoires revealed functional copy number changes associated with ecomorphological adaptations, sociality and domestication. This study highlights the value of gapless genomes to reveal structural mechanisms underpinning karyotypic evolution, reproductive isolation and ecological niche adaptation.

Comparative genomics is a powerful approach for inferring the genetic basis of adaptation and speciation. Its success depends on accurate and representative whole-genome alignments that precisely quantify genetic similarities and differences between evolutionary lineages to make predictions regarding the impact of genomic divergence on phenotypic evolution and diversification. The application of long-read sequencing has enabled increasingly precise comparisons between taxa, facilitating the assembly of 92–96% of a diploid genome sequence into chromosomes[1,2]. However, tracing the evolutionary history of regions of high structural complexity and allelic divergence has remained challenging. Until the completion of the human telomere-to-telomere (T2T) project[3–5], genomic 'dark matter' (refs. 6,7) that encompasses satellite arrays, centromeres, segmental duplications (SDs) and complex gene families had been missing from nearly all comparative genomic studies. Consequently, for most species, we still have a limited understanding of the evolutionary dynamics of the

[1]Veterinary Integrative Biosciences, Texas A&M University, College Station, TX, USA. [2]Interdisciplinary Program in Genetics & Genomics, Texas A&M University, College Station, TX, USA. [3]Department of Genome Sciences, University of Washington, Seattle, WA, USA. [4]School of Biology & Environmental Sciences, University College Dublin, Dublin, Ireland. [5]Department of Animal Sciences, University of Missouri, Columbia, MO, USA. [6]Institute for Systems Biology, Seattle, WA, USA. [7]Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, USA. [8]Guy Harvey Oceanographic Center, Nova Southeastern University, Fort Lauderdale, FL, USA. [9]Department of Veterinary Medicine & Surgery, University of Missouri, Columbia, MO, USA. [10]These authors contributed equally: Kevin R. Bredemeyer, LaDeana Hillier, Andrew J. Harris. ✉e-mail: warrenwc@missouri.edu; wmurphy@cvm.tamu.edu

most repetitive genomic sequences and how their divergence manifests in reproductive isolation and phenotypic innovation.

The cat family Felidae represents a speciose and successful apex predator radiation that occupies diverse biomes across the globe. Previous comparative genomic studies have illuminated their rapid diversification in the Miocene[8,9], frequent postspeciation gene flow[9,10], the impacts of demographic changes on genetic diversity and fitness[11–13], and the genetic consequences of domestication[14]. Here we applied the trio-binning approach[15] to three divergent interspecific crosses amenable to high-resolution haplotype phasing (Fig. 1a) to generate near-gapless genome assemblies from multiple species pairs along the felid phylogeny. Comparisons of these assemblies provided an unprecedented glimpse into the properties of large and complex gene families and functional repetitive elements that were previously inaccessible[14,16,17]. We describe insights into the cauldron of repetitive genetic variation with potentially large effects on chromosome function and speciation.

Results

Phased genome assembly reveals remarkable collinearity

We used long-read PacBio sequencing to phase and assemble six single-haplotype genomes from five cat species (domestic cat, leopard cat, Geoffroy's cat, tiger and lion) through the application of trio-binning to three F1 interspecies hybrids[15]. The parent species of the crosses diverged ≥4 million years ago (MYA; Fig. 1a), enabling >99.5% of the long sequence reads to be accurately phased into the parental haplotypes[18] (Fig. 1a, Supplementary Figs. 1–4 and Supplementary Table 1). De novo assembly produced ultracontiguous assemblies with contig N50 = 77–104 Mb (Table 1 and Fig. 1b). At least 99.6% of the euchromatic sequence was assembled into chromosome-length scaffolds using high-throughput chromosome conformation capture (Hi-C; Supplementary Fig. 5), with an average of just 53 gaps per genome assembly, 15 gapless chromosomes across all species and 62% of the assembled autosomes containing two or fewer gaps (Fig. 1c and Supplementary Fig. 6), exceeding comparable parameters from all other domestic species reference assemblies (Fig. 1b). The canonical telomeric sequence shared by vertebrates is TTAGGG[19]; however, different blocks of microsatellites are found in telomeres of other species of the generalized pattern (TxAyGz)[20]. To determine which chromosome assemblies extended into one or both telomeres, we searched for telomere-like repeat sequences by requiring 80% of the terminal 100 bases of the chromosome to be labeled as a repeat family or a tandem repeat. Then, we extended the search window progressively. About 61% of the chromosomes in the six assemblies likely extend into both telomeres, 32% extend into one telomere and the remaining 7% lack terminal repeats and are likely incomplete. Only 32% of the assembled chromosomes possess the canonical TTAGGG tandem array at the telomere, while 21 chromosomes terminated with the FA-satellite[21,22] (Supplementary Table 2).

Pairwise whole-genome alignments between the five species' assemblies revealed near-complete karyotypic stasis after they diverged from a common ancestor ~11 to 15 MYA[9,10] (Fig. 1d). The only change in chromosome number is a single Robertsonian translocation of two small acrocentrics (chrF1 and chrF2), producing a medium-size metacentric (chrC3) shared by all species of the neotropical cat genus *Leopardus* (Fig. 1e,f)[23]. Close inspection of alignments between *Leopardus geoffroyi* and *Felis catus* showed that chromosome C3 was the product of a centric fusion, followed by a near chromosome arm-length inversion that reoriented >99% of C3q relative to the ancestral chrF2 homolog (Fig. 1g). All other chromosomal rearrangements between species were inversions several orders of magnitude smaller in size (<2 Mb; Fig. 2a and Supplementary Table 3). We identified 172 fixed inversions >50 bp (Fig. 2a) across the five species phylogeny that samples >50 million years of independent branch length. By comparison, great ape genomes contain the products of 1,326 fixed inversions >50 bp (ref. 1) (Fig. 2a). Felids and great apes diverged on a very similar evolutionary timescale, matching nearly 1:1 for divergence events (Fig. 2a). Given the similarity in sampled evolutionary history, great ape genomes possess 7.7-fold more rearrangements than felids suggesting that great ape genomes are more structurally prone to chromosome rearrangement than felids.

SDs have been hypothesized to be major drivers of chromosome evolution and disease susceptibility in the great ape lineage by promoting nonallelic homologous recombination[24,25], particularly because of their uniquely interspersed distribution[26]. In support of this hypothesis, SDs flank 82–86% of known primate inversions[27]. To determine whether SDs might be a primary driver of felid inversions, we used SEDEF[28] to identify SDs in each cat haplotype. The total bases in felid SDs range from 25 to 35 Mb, or 1% to 1.5% of each genome (Supplementary Fig. 7). By comparison, the SD frequency (7%) estimated in the human T2T genome[29] is five- to seven-fold higher than in felid genomes. Compared to great apes, the similar-fold reduction in chromosomal rearrangements and SD frequency in felid genomes supports the hypothesis that the overall frequency of SDs is the primary driver of chromosome evolution in these two lineages. Future analysis of near-gapless genomes in other mammalian lineages with highly variable rates of karyotypic evolution will enable the testing of this hypothesis.

Structural variation is enriched on chromosome X

The hemizygous nature of the X chromosome (chrX) in male heterogametic taxa promotes faster rates of evolution relative to the autosomes and the accumulation of loci associated with reproductive isolation and speciation[30,31]. Previous studies revealed a higher fixation rate of inversions on chrX relative to autosomes[32,33]. In cats, chrX was an outlier in terms of the number of inversions relative to chromosome length (Fig. 2b). For each branch in the phylogeny, the mean inversion was significantly larger on chrX than the autosomes (Fig. 2c). Inversions accumulated disproportionately in an ~45-Mb recombination cold spot on chrX that is enriched for barriers to gene flow across multiple felid lineages[10] (Fig. 2d). Two thirds (24/36) of the X-linked inversions were fixed versus polymorphic (Supplementary Table 3). About 70% of fixed inversions harbored at least one protein-coding gene (mean 1.3 genes/fixed inversion). In contrast, only 33% of polymorphic inversions spanned or overlapped with a single protein-coding gene. In half of these cases, the inversion was located within a long intron (Supplementary Table 4). These results support previous observations in insects[33] and suggest that the fixed X-linked inversions within the 45-Mb recombination cold spot may harbor beneficial alleles given their longer length and enrichment with protein-coding genes. Previous studies of small and big cats identified signatures of natural selection within the large recombination cold spot[14,34]. We hypothesize that this gene-rich, inversion-rich region is a major X-linked supergene locus underpinning felid reproductive isolation that warrants future comparative genomic analyses.

Satellite elements have been implicated in speciation but are poorly represented in diploid genome assemblies[35,36]. Cat chrX harbors the only X-linked speciation gene identified in mammals; the macrosatellite repeat *DXZ4* (ref. 37). *DXZ4* has been well studied regarding its putative role in mammalian chrX inactivation (XCI). Human *DXZ4* consists of a single 3-kb tandem repeat array containing 56 monomers, where each repeat contains a single CTCF-binding site[4] (Fig. 3a). Long noncoding RNAs (*DANT1* and *DANT2*) expressed from *DXZ4* on the inactive chrX (Xi) promote superlooping with other macrosatellites on the Xi[38] and facilitate the localization of the Barr body in female placental mammals to the nucleolar membrane[39] (Fig. 3a). The human T2T genome assembly first completely resolved the *DXZ4* array structure, but a complete assembly of *DXZ4* sequences in other mammalian taxa is largely lacking, clouding our understanding of its evolution and function. *DXZ4* was resolved in all six cat assemblies, revealing a unique

https://doi.org/10.1038/s41588-023-01548-y

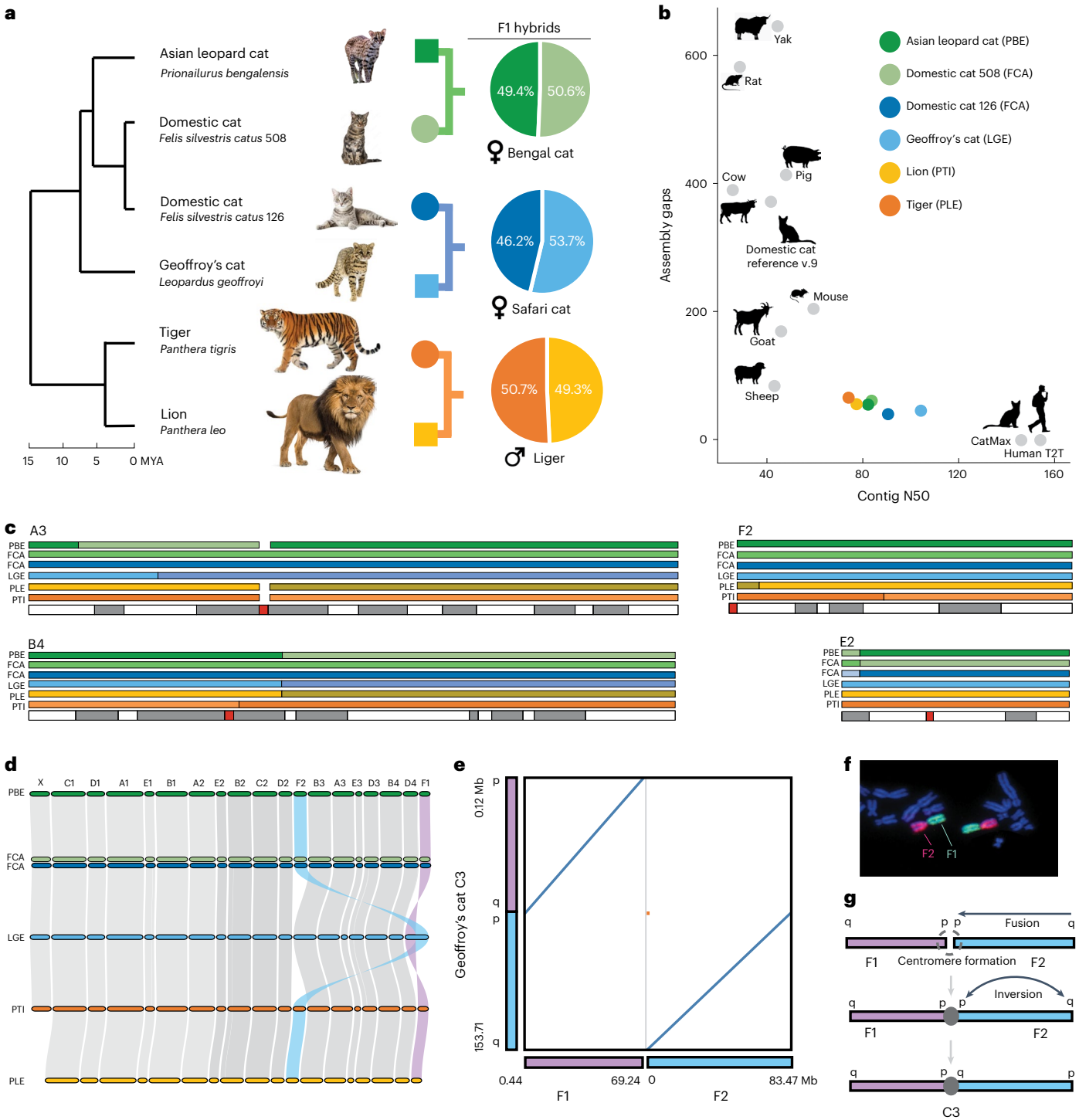

Fig. 1 | Assembly and synteny comparisons among the genomes of five cat species. a, Phylogeny and timescale of the parent species of the three hybrid trios used for assembly and comparative analysis. Pie charts illustrate the phasing results (% of total reads) for the F1 PacBio CLR long reads. **b**, Comparison of contig N50 statistics and number of assembly gaps against other highly contiguous mammalian reference genomes from domestic species. CatMax refers to the theoretical N50 maximum based on domestic cat chromosome sizes. PBE, *Prionailurus bengalensis*; FCA, *Felis catus*; LGE, *Leopardus geoffroyi*; PTI, *Panthera tigris*; PLE, *Panthera leo*. **c**, Contig alignments for the six felid single-haplotype assemblies from chrsA3, B4, E2 and F2/C3 to the felCat9 diploid domestic cat long-read genome assembly, depicted on the bottom as a G-banded ideogram. Inferred centromere locations are indicated by red bars. The bars above each ideogram are colored by species and represent assembly contigs > 1 Mb. Breaks between contigs are indicated by a black line and a shift in color contrast. The full set of chromosome alignments is found in Supplementary Fig. 6. **d**, Synteny plot[70] illustrating extensive collinearity of the five species assemblies. Blue and purple alignment tracks highlight the only chromosome number change in Felidae, the Robertsonian fusion of chrF1 and chrF2 present in all felid genera, and the derived C3 chromosome observed in Geoffroy's cat and all species of the genus *Leopardus*. **e**,**f**, Dot plot alignment (left) of Geoffroy's cat chrC3 and domestic cat chrF1 and chrF2 (**e**) (illustrated with multicolor FISH in **f**). **g**, Note the orange alignment fragment (in **e**) indicating a small centromeric fragment of chrF2 that defines the inversion breakpoint on the ancestral chrF2.

https://doi.org/10.1038/s41588-023-01548-y

Table 1 | Felid single-haplotype genome assembly statistics

Species/hybrid sequenced	Domestic cat-508 (Bengal cat F1)	Domestic cat-126 (Safari cat F1)	Asian leopard cat (Bengal cat F1)	Geoffroy's cat (Safari cat F1)	Lion (liger F1)	Tiger (liger F1)
Sex of parent haplotype	♀	♀	♀	♀	♂	♀
Chromosomes	18, X	18, X	18, X	17, X	18, Y	18, X
Contigs	123	103	132	88	103	135
Largest contig	205,171,639	172,124,406	240,846,738	239,106,607	166,870,000	166,130,000
Ungapped assembly length (Mb)	2,422,283,418	2,425,722,929	2,435,689,660	2,426,362,316	2,297,542,863	2,408,668,598
Contig N50 (Mb)	84,507,663	92,686,623	83,696,501	104,474,415	77,781,637	74,360,613
Scaffolds	71	70	83	46	53	74
Total assembly length (Mb)	2,422,299,418	2,425,747,038	2,435,718,761	2,426,370,816	2,297,568,983	2,408,695,688
Scaffold N50 (Mb)	147,603,332	148,491,486	148,587,958	152,606,360	147,402,474	146,942,463
Chromosome gaps	60	39	56	45	55	65
Complete BUSCO genes (mammalia_odb10)	8,621	8,619	8,621	8,612	8,417	8,630
Percent complete	93.4	93.4	93.4	93.3	91.2	93.5
Single copy	8,599	8,596	8,599	8,592	8,383	8,601
Duplicated	154	160	154	152	143	147
Missing	451	447	451	462	666	449
Complete+partial (%)	95.1	95.2	95.1	95.0	92.8	95.1

compound tandem repeat composed of two highly divergent (mean *P* distance = 0.67) repeat arrays, RA and RB (Fig. 3b). Both monomer types contain CTCF-binding sites, but notably differ in the number and orientation of the sites that are important for CTCF-binding affinity and loop extrusion directionality[40], suggesting divergent superlooping functions between the arrays. The human and mouse genomes notably lack the RB array.

Studies using interspecific backcross hybrids of the domestic cat and Jungle cat (*Felis chaus*) identified *DXZ4* as a major-effect hybrid male sterility locus, with a likely role in reproductive isolation within the *Felis* genus[37]. The testicular germ cells of sterile male hybrid cats possess RA-specific methylation defects and *DANT1* misregulation, culminating in the failure of meiotic sex chromosome inactivation (MSCI) and meiotic arrest, hallmark phenotypes in mammalian interspecies hybrids[31]. Evidence that *DXZ4* functions in male meiotic silencing was intriguing, given the parallels between the heterochromatic Barr body formed during female XCI and the condensed X–Y body in male MSCI. Although the hybrid sterility phenotype was attributed to *DXZ4* interspecific divergence, the precise mechanism is not well understood. Here our expanded sampling of felid genomes demonstrates that the compound RA and RB repeat structure is copy number variable across all species (Fig. 3b), suggesting copy number-mediated expression effects may have an important role in speciation in other felids. In addition, StainedGlass[41] plots illustrate the rapidity of *DXZ4* repeat array sequence divergence (Fig. 3c). RA and RB arrays evolve two- to three-fold faster than the flanking and intervening noncoding spacer sequences. Notably, a genome-wide analysis of pairwise interspecific genetic divergence calculated across 28,312 5-kb alignment windows (94.1% of the multispecies alignment) placed *DXZ4* RA in the top 0.5% of the most rapidly evolving genomic loci (Fig. 3d), supporting its role as a speciation gene[37].

To determine whether the compound *DXZ4* array structure in cats is the exception or the rule in placental mammals, we searched for *DXZ4* arrays in long-read genome assemblies from species representing divergent superorders (Fig. 3e and Supplementary Figs. 8–11). Most assemblies possessed a gap within or adjacent to the predicted position of *DXZ4* (Supplementary Figs. 12 and 13). We were able to recover sufficient repeat array resolution at the edge of some assembly gaps to characterize the CTCF array. Although the *DXZ4* monomer sequence diverges rapidly to the point of phylogenetic saturation and lack of phylogenetic patterning (Supplementary Fig. 14), we observed the conservation of the CTCF-binding motif patterns across species from different ordinal lineages. Euarchontoglires (for example, primates, rodents and rabbits) possessed only RA or RB, while members of Laurasiatheria possess RA, RB or both types (Fig. 3e). RA and RB were, therefore, present in the most recent common ancestor of boreoeutherian mammals. Moreover, the repeat unit length is relatively constrained (between 3 and 4.9 kb) across species despite rapid sequence divergence and little conservation outside the CTCF motif[42]. Given this unusual combination of spatial and structural evolutionary conservation and an extremely fast rate of sequence evolution, we predict that *DXZ4* satellite divergence may have a more widespread role in establishing and maintaining species boundaries in other mammalian clades.

Intriguingly, all sampled species from the family Bovidae lack *DXZ4* in their assembly, suggesting they may have evolved compensatory mechanisms for its putative loss. Multiple studies have shown that ablation of *DXZ4* has no significant impact on the silenced state of the inactive chrX in mouse and human cells[40,43]. Nonetheless, the high degree of syntenic, CTCF[42] and spatial conservation of the *DXZ4* repeat array over the past 104 million years of the placental mammal radiation suggest that *DXZ4* expression and long-range chromatin interactions are functionally important for some heretofore unidentified cellular role during XCI and MSCI[44]. Pan-autosomal gene downregulation is one noteworthy cellular phenotype shared by in vivo *DXZ4*-knock-out mice[45] and sterile feline interspecific hybrid testes[37]. These observations raise the possibility that *DXZ4*, acting alone or in concert with other X-linked macrosatellites, may function in RNA-dependent, chrX-autosomal crosstalk associated with the chrX 'counting' process in XCI[45] and proper sequestration of the DNA damage response factors from the autosomes to the X–Y body during MSCI[46,47]. Gapless chrX assemblies from a diverse sampling of mammalian genomes will be

https://doi.org/10.1038/s41588-023-01548-y

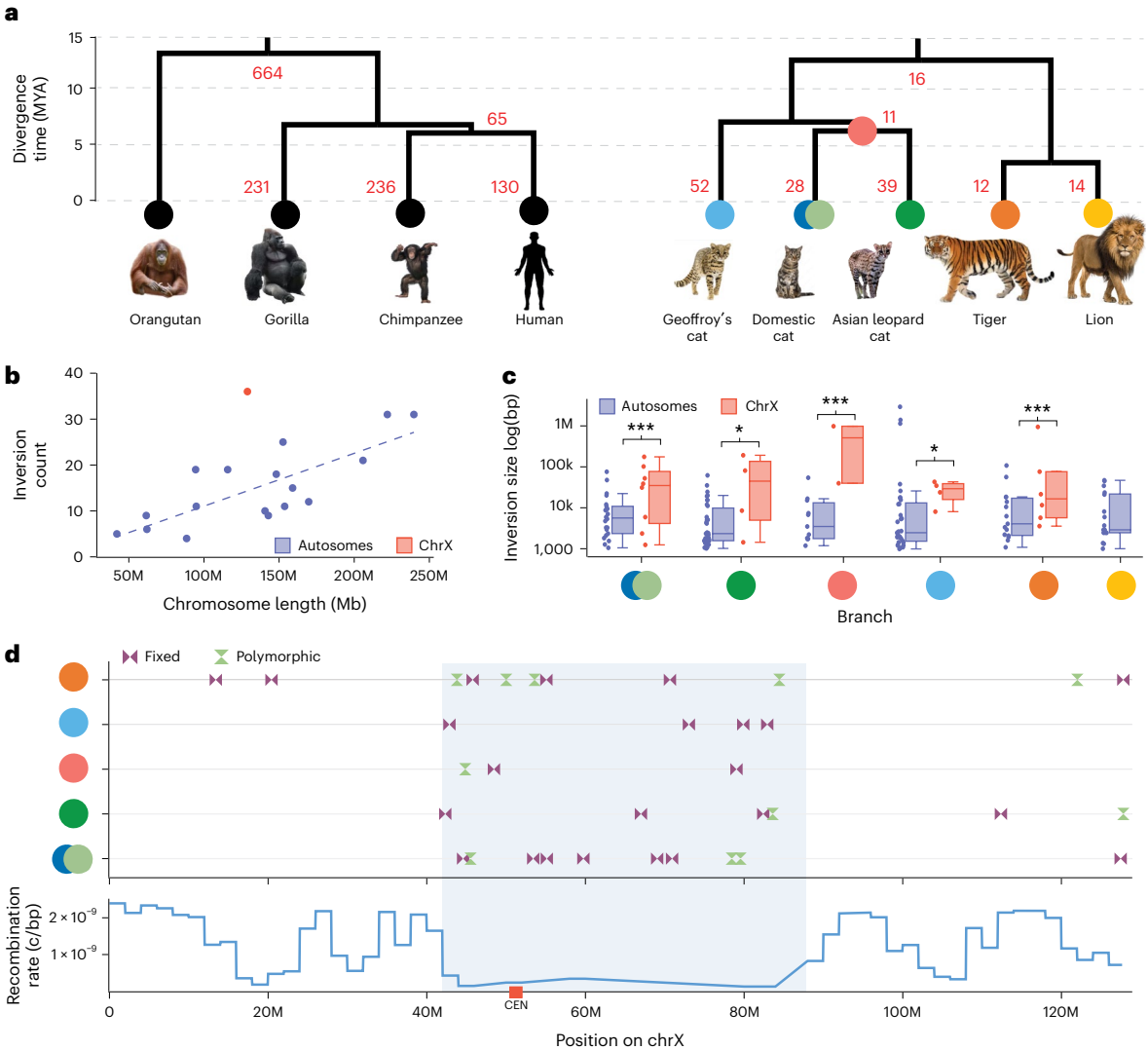

Fig. 2 | Felid structural variation. a, Comparison of fixed inversions (red numbers) plotted on branches of the phylogeny of felids (right) and great apes (left) (5). Note the similar divergence times between ape and felid species sampled. **b**, Per chromosome inversion counts plotted against chromosome length. Autosomes are indicated with blue dots and chrX in red. **c**, Comparison of inversion size between the autosomes and chrX for each branch of the phylogeny (colored dots) shown in **a** (except for the lion genome, which is derived from the paternal haplotype of the male F1 liger; Supplementary Table 3). A one-sided Wilcoxon rank sum test determined significance (*$P < 0.05$, ***$P < 0.01$). Domestic cat ($n = 40$ autosomal inversions, $n = 11$ chrX inversions, $U = 2.52$, $P = 5.9 \times 10^{-3}$), Geoffroy's cat ($n = 33$ autosomal inversions, $n = 4$ chrX inversions, $U = 2.15$, $P = 1.6 \times 10^{-2}$), Asian leopard cat ($n = 40$ autosomal inversions, $n = 6$ chrX inversions, $U = 1.92$, $P = 2.7 \times 10^{-2}$), domestic cat + Asian leopard cat ($n = 17$ autosomal inversions, $n = 3$ chrX inversions, $U = 2.59$, $P = 4.8 \times 10^{-3}$), tiger ($n = 34$ autosomal inversions, $n = 11$ chrX inversions, $U = 2.54$, $P = 5.6 \times 10^{-3}$), lion ($n = 34$ autosomal inversions). Box plots show the interquartile range with the center line representing the median. Whiskers indicate the highest and lowest value within the upper and lower fences (upper fence = 75% quantile + 1.5× interquartile range, lower fence = 50% quantile − 1.5× interquartile range). **d**, The physical distribution of fixed and polymorphic inversions (Supplementary Table 4) on chrX for each branch of the phylogeny relative to the tiger genome. The chrX genome sequences are otherwise collinear across species. A tiger recombination map estimated from population genomic data (Supplementary Fig. 30) is depicted at the bottom (Methods) and is highly conserved with the recombination rate profile of the domestic cat chrX[9,71]. The shaded area refers to a large recombination cold spot shared with domestic cats, humans and pigs[9,10].

critical to understanding the functional relevance of *DXZ4* in the chrX biology of placental mammals.

Variation in centromere structure and size

Current human and great ape centromere sequence models portray large tandem repeat arrays of α satellites flanked by other satellite repeat types, SDs, transposable elements and even some genes[48]. Whether centromere structure is conserved across mammalian lineages is poorly understood because they are not sequence-resolved in most genome assemblies. Therefore, we sought to determine whether our assemblies possessed genomic signatures characteristic of centromeric satellites[5]. Given the absence of previously annotated cat centromeric sequences, we first characterized the overall landscape of feline repetitive elements to enable de novo prediction of the most probable centromeric satellites (Supplementary Fig. 15). Interspersed repeats comprise 38% of each genome with a marked distinction between Felinae (*Felis*, *Prionailurus* and *Leopardus*) and *Panthera*, with Felinae showing an average SINE insertion rate of ~2.7× higher than *Panthera*, while conversely, the LINE insertion rate in *Panthera* is ~1.6× higher than Felinae (Supplementary Fig. 16).

Next, we searched for novel repeat enrichment within narrowly defined chromosomal regions for which we had a strong priori evidence classifying that region as centromere-containing based on integrative analysis of comparative mapping approaches[9,14,17]

https://doi.org/10.1038/s41588-023-01548-y

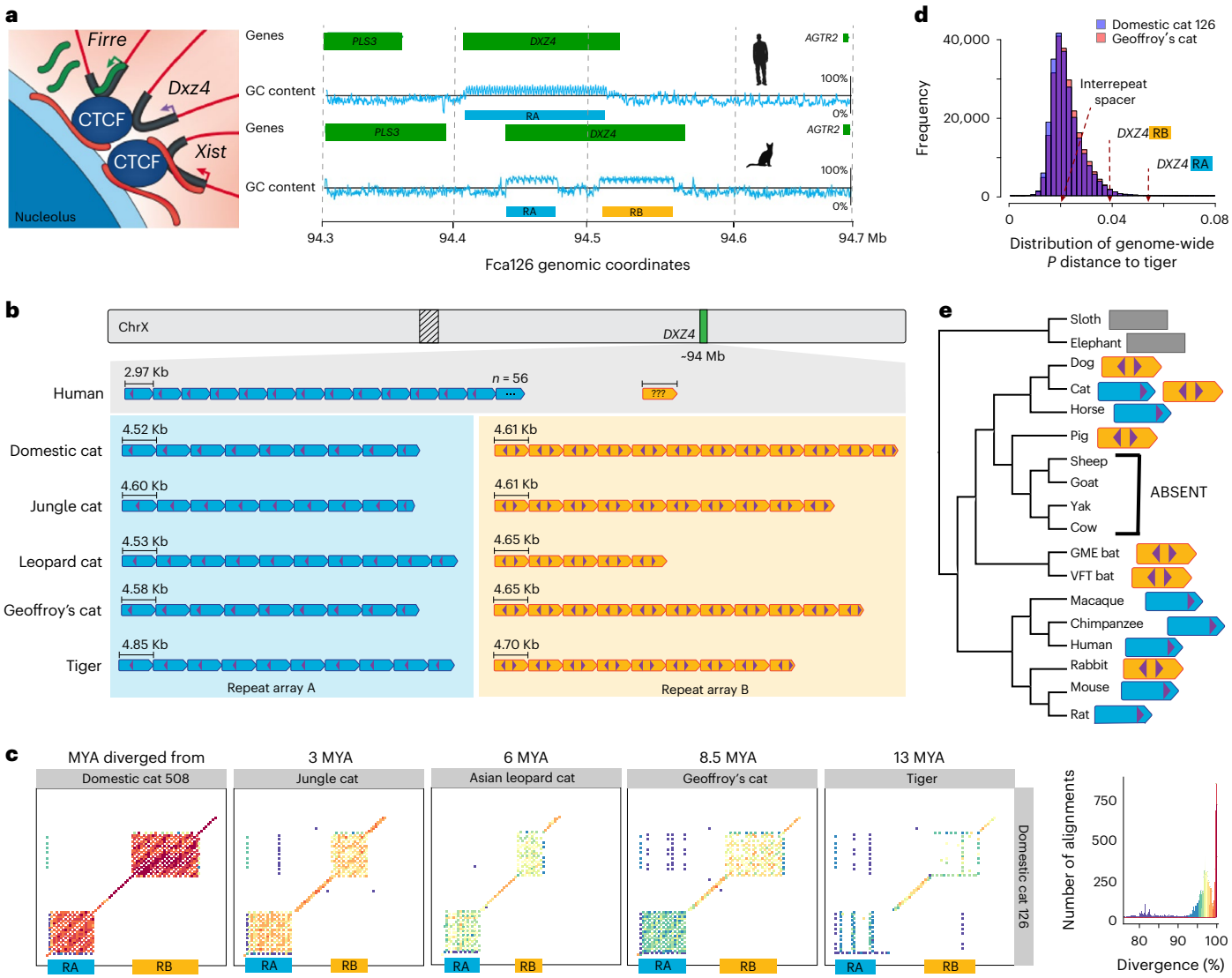

Fig. 3 | *DXZ4* evolution in placental mammals. a, Left, X-linked lncRNAs from *Dxz4*, *Xist* and *Firre* cooperatively interact in 3D space to anchor the inactive chrX to the nucleolus (figure modified from ref. 72); right: comparison of the human and domestic cat *DXZ4* repeat structure and GC content shown in genomic context to flanking genes *PLS3* and *AGTR2*. Felids possess two distinct repeat arrays, RA (blue) and RB (yellow), while humans only possess the RA type. **b**, *DXZ4* repeat unit size, CTCF-binding site composition (purple arrows), and copy number in human (top) and sequenced cat species. The Jungle cat data are from a single-haplotype chrX assembly (27). **c**, StainedGlass (version 59) dot plots showing *DXZ4* repeat array divergence between the domestic cat (Fca126) and other cat species (the percentage of identity between species alignments is shown to the right) in increasing order of evolutionary divergence. Note higher conservation across the central and flanking regions adjacent to the RA and RB arrays. **d**, Distribution of genomic divergence rates between tiger-Geoffroy's cat and tiger-domestic cat across 28,312 5-kb alignment windows. Pairwise divergence values for *DXZ4* RA and RB and the internal spacer region are shown for comparison. **e**, Phylogeny of placental mammals with *DXZ4* repeat array presence (blue = RA type, yellow = RB type, gray = ambiguous) inferred from each genome assembly.

(Supplementary Fig. 17). This strategy identified a single, most probable centromere-containing interval for each chromosome enriched >1,000-fold with a small class of tandem repeats (Supplementary Fig. 18). The location of these intervals was highly conserved across species and consistent with stability of the felid karyotype. Like human and ape centromeres, several better-resolved cat centromeres (for example, chrE3; Fig. 4a) consisted of a central satellite array of higher-order repeats. The predominant satellite repeat was 113 bp in length, ~25% smaller than the 151-bp α satellite typical of great ape centromeres[5,48] (Supplementary Fig. 19). StainedGlass analysis of these candidate satellite arrays revealed patterns of monomer divergence similar to great ape centromere arrays, with more divergent monomers flanking higher identity monomers within the central satellite array (Fig. 4a). The Geoffroy's cat possessed the largest centromeric repeat arrays on most chromosomes (Supplementary Fig. 20). This species' karyotype also has the distinct C3 metacentric chromosome, a product of a Robertsonian chromosome fusion between chrF1 and chrF2, which occurred in the ancestor of the *Leopardus* lineage ≥3 MYA (refs. 9,10). StainedGlass and syntenic alignment plots (Fig. 4b and Supplementary Fig. 21) reveal that Geoffroy's cat chrC3 centromeric region retains the highest pattern and sequence similarity to the ancestral chrF1 centromeric satellite array.

Centromere sizes and repeat composition varied markedly between chromosomes and across felid species. Although we cannot exclude incomplete/collapsed sequences for some of this variation (Supplementary Figs. 22–25), the centromeric regions of three autosomes were gapless in all six felid genomes (chrs. B4, D4 and E2), likely due to reduced satellite array repeat complexity. For example, *Felis*

https://doi.org/10.1038/s41588-023-01548-y

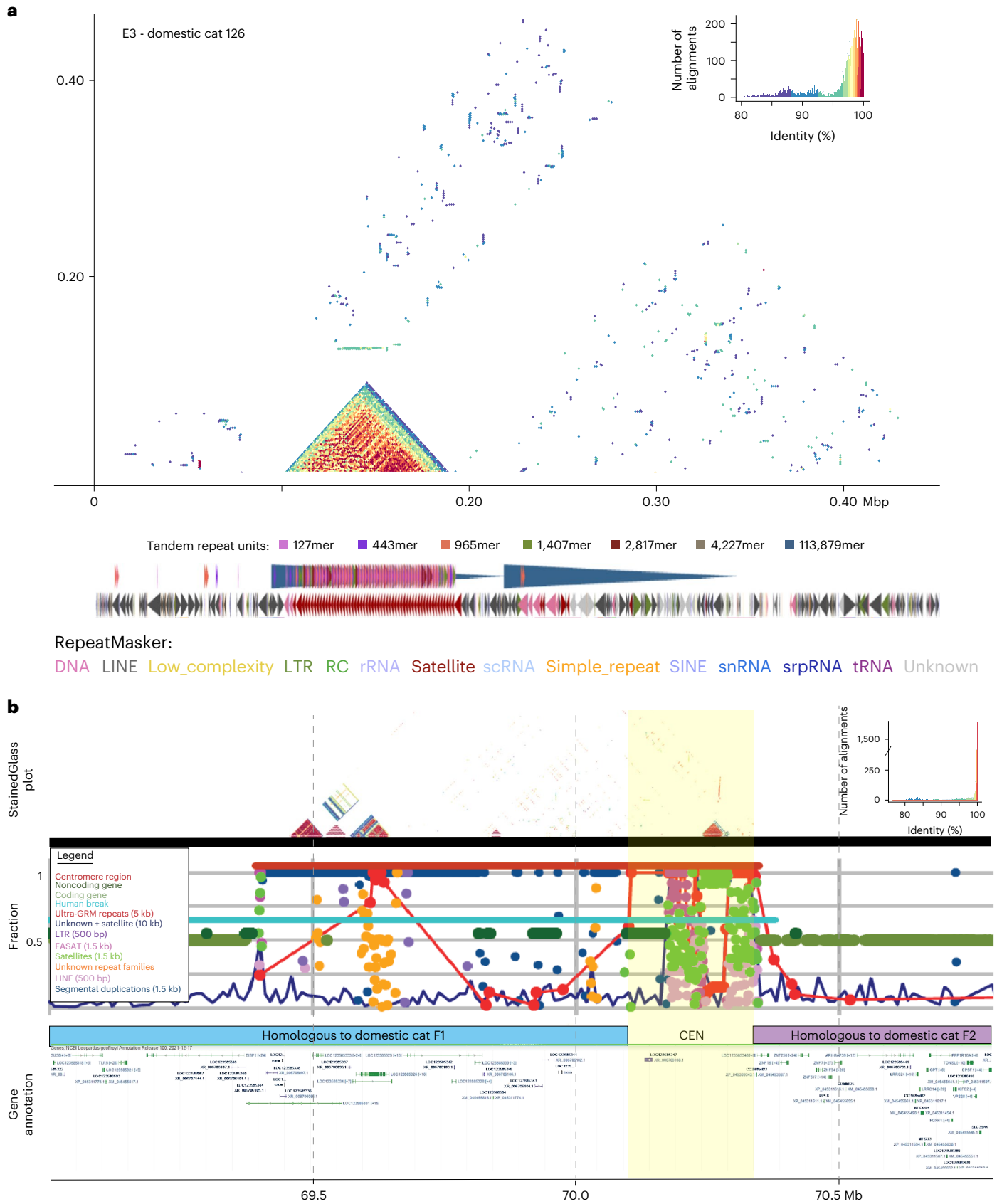

Fig. 4 | Centromere annotation and evolution. a, StainedGlass[41] dot plot of domestic cat-126 chrE3 centromere region showing percent identity of self-alignments within the satellite repeat array (colored triangle, with % identity scale and distribution shown in the upper right). Below the chromosome are tracks for tandem repeat annotations (colors indicate different GRM-defined repeat units) and RepeatMasker annotations (key at bottom). **b**, Geoffroy's cat chrC3 centromere region. The lower two panels display NCBI CpG and gene annotations and inferred homology to the domestic cat F1 and F2 centromeric regions. The top tracks show StainedGlass plots and repeat annotations (and fractions observed on *y* axis). The most probable centromeric repeat array is highlighted in yellow and supported by alignments in Supplementary Fig. 21. CEN, centromere.

https://doi.org/10.1038/s41588-023-01548-y

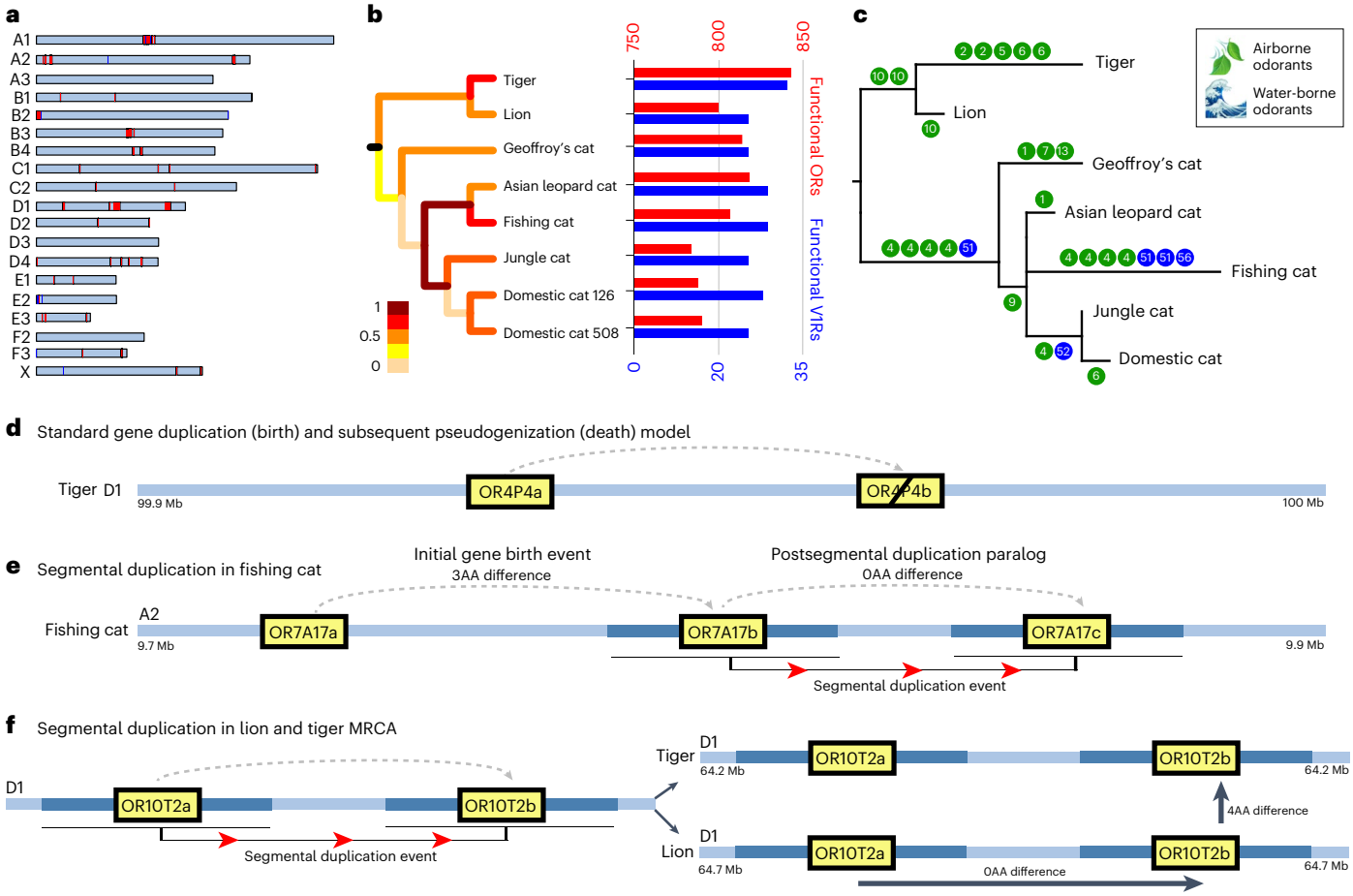

Fig. 5 | ORG and V1R gene evolution in cats. a, Chromosomal distribution of ORG (red) and V1R gene (blue) within the domestic cat genome. **b**, Phylogeny and rate of ORG family duplications (scale to lower left). Barplots to the right illustrate per species ORG (red) and V1R (blue) functional gene copy number. **c**, Number of per branch unique ORG retention, classified into class I (blue = 'water-borne') and class II (green = 'airborne') receptor types. Each circle represents a uniquely retained gene, with its subfamily classification depicted by the number. **d**–**f**, Models of ORG birth and death with specific examples—the standard birth and death (pseudogenization) model illustrated by tiger chrD1 (OR4P4a and OR4P4b) (**d**); a gene birth followed by paralog birth via SD in the fishing cat (**e**) and a gene birth via SDs in the *Panthera* ancestor preceding speciation of the lion and tiger lineages (**f**).

chrB4 possesses a narrower centromeric interval and lacks the large satellite arrays observed on other chromosomes (Supplementary Fig. 26). Some mammalian families, such as equids (donkeys, onagers and zebras), also exhibit considerable variability in the presence/absence of satellite repeats at their centromeres[49,50]. By contrast, the chrD4 centromere possesses a mostly conserved satellite array and illustrates the rapidity with which the central satellite monomer array sequences diverge relative to the flanking sequence (Supplementary Figs. 27 and 28), similar to great apes[5]. These new assemblies pave the way to exploring the potential role of interspecific centromeric satellite variation in felid speciation[51].

Evolutionary innovations in sensory supergene families

Olfactory receptor genes (ORGs) encode receptors that detect odorants and represent the largest gene superfamily, dispersed across the majority of mammalian chromosomes[52] (Fig. 5a). Variation in repertoire size and functional content has been linked to shifts in ecology, diet and life history traits, which are likely crucial components of adaptation to new environments[53,54]. Most comparative studies of ORG variation were based on short-read assemblies, which confound allelic discrimination and gene copy number differences. Indeed, the previous enumeration of differences in OR repertoire sizes between cats and tigers produced opposing results[14,54]. We quantified the functional ORG and vomeronasal receptor (V1R) gene profiles within each genome assembly and added published repertoire reconstructions from the jungle cat (*Felis chaus*)[37] and a fishing cat (*Prionailurus viverrinus*) based on Hi-Fi reads[55]. These assemblies showed gapless ORG and V1R gene cluster inclusion with contiguity metrics approaching the single-haplotype assemblies (mean cN50 = 80 versus 91 Mb).

We observed large ORG copy differences (>10% of the maximum repertoire size) between species (Fig. 5b and Supplementary Table 5). Felids retain >70% functional ORGs (Supplementary Table 6), larger than most mammals[54]. This elevated functional repertoire may reflect their predatory behaviors, with an acute sense of smell to track and locate prey across great physical distances[56]. The tiger is solitary, with among the largest home range sizes and habitat diversity of any living felid[57]. It possesses the most extensive functional ORG repertoire and the highest number of gene duplications of any sampled species for airborne Class II ORGs (Fig. 5b,c and Supplementary Tables 6 and 7). Several ORGs that are known to bind volatile compounds in the blood (*OR1G1*: nonanal, *OR2W1* and *OR51V1*: hexanal)[58,59], and the pheromone androstenone (*OR7D4*)[58] had relatively high copy numbers (Supplementary Fig. 29). The tiger and Geoffroy's cat lineages both possessed specific duplications in ORGs associated with blood-associated odorants. By contrast, the ancestor of the domestic cat lineage had the fewest ORG duplication events, potentially reflecting relaxed evolutionary pressure on olfaction before or during domestication.

Class I ORG families (OR51, OR52, OR55 and OR56) are generally considered the 'water-borne' odorant-binding class, and selection for functional copies is usually rare in terrestrial mammals. The fishing cat (*Prionailurus viverrinus*) is one of two felids with pronounced aquatic adaptations such as foot webbing and other otter-like morphological adaptations to the head and tail[56]. The fishing cat possesses one of the largest relative percentages of functional water-borne ORGs (75%), similar to the two domestic cats (74% and 76%) and higher than the other wild felids (lion: 67%, tiger: 71%, Geoffroy's cat: 72% and leopard cat and jungle cat: 73%; Supplementary Table 8). Notably, the adaptive importance of water-borne OR receptors to the fishing cat is reflected in the lack of any class I-specific pseudogenization events within its lineage and the retention of three functional class I ORGs that have subsequently been pseudogenized in all other felid species (Fig. 5c and Supplementary Table 9).

ORG sequences evolve through an evolutionary pattern known as the birth-and-death model[60] (Fig. 5d). This model assumes new ORGs are 'born' through tandem gene duplication and retained via subfunctionalization or neofunctionalization[61]. Gene death occurs from nonsense mutations or larger-scale genic deletions. Analysis of the chromosomal regions flanking ORG clusters revealed that while many of the inferred duplication events consisted of the ORG sequence alone, 18 of 198 detected lineage-specific gene duplications (9.1%) were the product of larger SDs spanning ≥2,000 bp (Fig. 5e,f), similar to the frequency (10%) of SD-driven ORG duplications in humans[62]. A mean rate of 2.73 amino acid mutations was observed between functional segmentally duplicated ORGs compared to 2.3 amino acids in gene-specific duplicates, suggesting differences in the rate of natural selection acting on ORG evolution may be dependent on the duplication mechanism. This distinction is important because all genes duplicated as part of a larger block may not be targets of selection. SD likely explains some of the more extensive ORG repertoires observed in mammals, as in the African elephant, which is estimated to possess over 2,000 functional genes but more than 1,000 pseudogenes[63]. Future analyses of sensory genes in T2T genomes will allow further exploration of this model of ORG evolution in a range of vertebrate taxa.

V1R detects pheromones and other sociochemicals. We recovered complete V1R gene repertoires for each species, ranging from 67 genes in the jungle cat to 85 genes in the tiger (Fig. 5b), with ~36% of V1R genes retaining function across species (Supplementary Tables 10 and 12). The Tiger genome possessed the most functional V1R loci. Like their large functional ORG repertoire, this is potentially attributable to the large physical distances necessary for tigers to detect scent marks and discriminate potential conspecific and reproductively receptive mates[64]. Most of the estimated gene duplication events occurred in tiger and lion genomes. They may reflect divergent adaptations to the use of social/sexual cues in both solitary and social life histories. Interestingly, we observed the highest frequency of nonfunctional (68%) V1R genes within the lion genome. Because lions live in highly cooperative groups in physical proximity, we hypothesize that the increased pseudogenization rate may be the product of relaxed selection on the use of chemical cues for determining sexual status and identifying mates relative to solitary species. Furthermore, while there were no unique lineage or species-specific retention of functional V1R genes like in the ORG family, the only unique V1R gene loss event occurred in the ancestor of the domestic cats, evidence of relaxed selective pressures during domestication[14].

Discussion

Here we applied feline hybrid models to produce multiple well-annotated and near-gapless sequence assemblies spanning the felid radiation. Despite their similar evolutionary ages, great ape and felid lineages possess distinct differences in SD densities that provide a genomic explanation for the striking karyotypic stability observed across the cat radiation. Resolving recalcitrant sequence structures also clarifies how natural selection continues to shape different axes of genomic diversity. The chemosensory system is particularly relevant in this sense, as gene family variation has large fitness effects, and here we showed that precisely resolved gene repertoires allow for discriminating the ecological relevance of gene birth and death. Notably, large differences in ORG and V1R gene repertoires between the closely related lion and tiger likely mirror the outcome of natural selection on evolved differences in social versus solitary life histories. The private retention of aquatic-borne odorant receptors in the fishing cat also helps to clarify the role of natural selection in ecological niche adaptation. Future studies of sensory gene repertoire variation within species occupying broad geographic ranges and habitats (for example, tiger, puma and bobcat) using phased assembly approaches will provide critical insights into the genetic basis of local sensory adaptation.

Speciation studies typically focus on the landscape of divergence, seeking outlier loci or 'islands of speciation' to uncover the genetic barriers that maintain species boundaries in the face of gene flow[65]. Our study illustrates the rapidity with which functional satellite elements evolve relative to background rates of gene sequence variation and provides additional evidence as to the role of *DXZ4*'s exceptional divergence in felid speciation. Yet satellites are often invisible to divergence scans as these highly repetitive regions are typically missing[4,37] or misassembled in most diploid genome assemblies. Future genomic prospecting from T2T genomes[3,66] promises to lend new insights into the landscape of genomic and structural divergence in adaptive phenotypic variation. We anticipate exciting breakthroughs inferring the genetic mechanisms of speciation and enabling genomically informed biodiversity conservation[67–69].

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information, details of author contributions and competing interests and statements of data and code availability are available at https://doi.org/10.1038/s41588-023-01548-y.

References

1. Kronenberg, Z. N. et al. High-resolution comparative analysis of great ape genomes. *Science* **360**, eaar6343 (2018).
2. Rhie, A. et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).
3. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
4. Miga, K. H. et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020).
5. Logsdon, G. A. et al. The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021).
6. Sedlazeck, F. J., Lee, H., Darby, C. A. & Schatz, M. C. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.* **19**, 329–346 (2018).
7. Ahmad, S. F. et al. Dark matter of primate genomes: satellite DNA repeats and their evolutionary dynamics. *Cells* **9**, 2714 (2020).
8. Johnson, W. E. et al. The late Miocene radiation of modern Felidae: a genetic assessment. *Science* **311**, 73–77 (2006).
9. Li, G., Davis, B. W., Eizirik, E. & Murphy, W. J. Phylogenomic evidence for ancient hybridization in the genomes of living cats (Felidae). *Genome Res.* **26**, 1–11 (2016).
10. Li, G., Figueiró, H. V., Eizirik, E. & Murphy, W. J. Recombination-aware phylogenomics reveals the structured genomic landscape of hybridizing cat species. *Mol. Biol. Evol.* **36**, 2111–2126 (2019).
11. Dobrynin, P. et al. Genomic legacy of the African cheetah, *Acinonyx jubatus*. *Genome Biol.* **16**, 277 (2015).

12. Abascal, F. et al. Extreme genomic erosion after recurrent demographic bottlenecks in the highly endangered Iberian lynx. *Genome Biol.* **17**, 251 (2016).
13. Johnson, W. E. et al. Genetic restoration of the Florida panther. *Science* **329**, 1641–1645 (2010).
14. Montague, M. J. et al. Comparative analysis of the domestic cat genome reveals genetic signatures underlying feline biology and domestication. *Proc. Natl Acad. Sci. USA* **111**, 17230–17235 (2014).
15. Koren, S. et al. De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* **36**, 1174–1182 (2018).
16. Cho, Y. S. et al. The tiger genome and comparative analysis with lion and snow leopard genomes. *Nat. Commun.* **4**, 2433 (2013).
17. Buckley, R. M. et al. A new domestic cat genome assembly based on long sequence reads empowers feline genomic medicine and identifies a novel gene for dwarfism. *PLoS Genet.* **16**, e1008926 (2020).
18. Bredemeyer, K. R., Harris, A. J., Li, G. & Zhao, L. Ultracontinuous single haplotype genome assemblies for the domestic cat (*Felis catus*) and Asian leopard cat (*Prionailurus bengalensis*). *J. Hered.* **197**, 165–173 (2021).
19. Meyne, J., Ratliff, R. L. & Moyzis, R. K. Conservation of the human telomere sequence (TTAGGG)n among vertebrates. *Proc. Natl Acad. Sci. USA* **86**, 7049–7053 (1989).
20. Peska, V. & Garcia, S. Origin, diversity, and evolution of telomere sequences in plants. *Front. Plant Sci.* **11**, 117 (2020).
21. Fanning, T. G. Origin and evolution of a major feline satellite DNA. *J. Mol. Biol.* **197**, 627–634 (1987).
22. Santos, S., Chaves, R. & Guedes-Pinto, H. Chromosomal localization of the major satellite DNA family (FA-SAT) in the domestic cat. *Cytogenet. Genome Res.* **107**, 119–122 (2004).
23. Wurster-Hill, D. H. & Centerwall, W. R. The interrelationships of chromosome banding patterns in canids, mustelids, hyena, and felids. *Cytogenet. Cell Genet.* **34**, 178–192 (1982).
24. Bailey, J. A., Baertsch, R., Kent, W. J., Haussler, D. & Eichler, E. E. Hotspots of mammalian chromosomal evolution. *Genome Biol.* **5**, R23 (2004).
25. Marques-Bonet, T., Ryder, O. A. & Eichler, E. E. Sequencing primate genomes: what have we learned? *Annu. Rev. Genomics Hum. Genet.* **10**, 355–386 (2009).
26. Cantsilieris, S. et al. An evolutionary driver of interspersed segmental duplications in primates. *Genome Biol.* **21**, 202 (2020).
27. Mao, Y. et al. A high-quality bonobo genome refines the analysis of hominid evolution. *Nature* **594**, 77–81 (2021).
28. Numanagic, I. et al. Fast characterization of segmental duplications in genome assemblies. *Bioinformatics* **34**, i706–i714 (2018).
29. Vollger, M. R. et al. Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 (2022).
30. Charlesworth, D. & Charlesworth, B. Sex chromosomes: evolution of the weird and wonderful. *Curr. Biol.* **15**, R129–R131 (2005).
31. Larson, E. L., Keeble, S., Vanderpool, D., Dean, M. D. & Good, J. M. The composite regulatory basis of the large X-effect in mouse speciation. *Mol. Biol. Evol.* **34**, 282–295 (2017).
32. Charlesworth, B., Coyne, J. A. & Barton, N. H. The relative rates of evolution of sex chromosomes and autosomes. *Am. Nat.* **130**, 113–146 (1987).
33. Cheng, C. & Kirkpatrick, M. Inversions are bigger on the X chromosome. *Mol. Ecol.* **28**, 1238–1245 (2019).
34. Figueiró, H. V. et al. Genome-wide signatures of complex introgression and adaptive evolution in the big cats. *Sci. Adv.* **3**, e1700299 (2017).
35. Ferree, P. M. & Prasad, S. How can satellite DNA divergence cause reproductive isolation? Let us count the chromosomal ways. *Genet. Res. Int.* **2012**, 430136 (2012).
36. Bayes, J. J. & Malik, H. S. Altered heterochromatin binding by a hybrid sterility protein in *Drosophila* sibling species. *Science* **326**, 1538–1541 (2009).
37. Bredemeyer, K. R. et al. Rapid macrosatellite evolution promotes X-linked hybrid male sterility in a Feline interspecies cross. *Mol. Biol. Evol.* **38**, 5588–5609 (2021).
38. Figueroa, D. M., Darrow, E. M. & Chadwick, B. P. Two novel *DXZ4*-associated long noncoding RNAs show developmental changes in expression coincident with heterochromatin formation at the human (*Homo sapiens*) macrosatellite repeat. *Chromosome Res.* **23**, 733–752 (2015).
39. Dossin, F. & Heard, E. The molecular and nuclear dynamics of X-chromosome inactivation. *Cold Spring Harb. Perspect. Biol.* **14**, a040196 (2022).
40. Bonora, G. et al. Orientation-dependent Dxz4 contacts shape the 3D structure of the inactive X chromosome. *Nat. Commun.* **9**, 1445 (2018).
41. Vollger, M. R., Kerpedjiev, P., Phillippy, A. M. & Eichler, E. E. StainedGlass: interactive visualization of massive tandem repeat structures with identity heatmaps. *Bioinformatics* **38**, 2049–2051 (2022).
42. Horakova, A. H. et al. The mouse *DXZ4* homolog retains Ctcf binding and proximity to *Pls3* despite substantial organizational differences compared to the primate macrosatellite. *Genome Biol.* **13**, R70 (2012).
43. Froberg, J. E., Pinter, S. F., Kriz, A. J., Jégu, T. & Lee, J. T. Megadomains and superloops form dynamically but are dispensable for X-chromosome inactivation and gene escape. *Nat. Commun.* **9**, 5004 (2018).
44. Brashear, W. A., Bredemeyer, K. R. & Murphy, W. J. Genomic architecture constrained placental mammal X chromosome evolution. *Genome Res.* **31**, 1353–1365 (2021).
45. Andergassen, D. et al. In vivo *Firre* and *Dxz4* deletion elucidates roles for autosomal gene regulation. *eLife* **8**, e47214 (2019).
46. Abe, H. et al. Active DNA damage response signaling initiates and maintains meiotic sex chromosome inactivation. *Nat. Commun.* **13**, 7212 (2022).
47. Abe, H. et al. The initiation of meiotic sex chromosome inactivation sequesters DNA damage signaling from autosomes in mouse spermatogenesis. *Curr. Biol.* **30**, 408–420 (2020).
48. Altemose, N. et al. Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022).
49. Carbone, L. et al. Evolutionary movement of centromeres in horse, donkey, and zebra. *Genomics* **87**, 777–782 (2006).
50. Raudsepp, T., Finno, C. J., Bellone, R. R. & Petersen, J. L. Ten years of the horse reference genome: insights into equine biology, domestication and population dynamics in the post-genome era. *Anim. Genet.* **50**, 569–597 (2019).
51. Henikoff, S., Ahmad, K. & Malik, H. S. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293**, 1098–1102 (2001).
52. Young, J. M. & Trask, B. J. The sense of smell: genomics of vertebrate odorant receptors. *Hum. Mol. Genet.* **11**, 1153–1160 (2002).
53. Hayden, S. et al. Ecological adaptation determines functional mammalian olfactory subgenomes. *Genome Res.* **20**, 1–9 (2010).
54. Hughes, G. M. et al. The birth and death of olfactory receptor gene families in mammalian niche adaptation. *Mol. Biol. Evol.* **35**, 1390–1406 (2018).
55. Carroll, R. A. et al. A novel fishing cat reference genome for the evaluation of potential germline risk variants. Preprint at *bioRxiv* https://doi.org/10.1101/2022.11.17.516921 (2022).
56. Sunquist, M. & Sunquist, F. *Wild Cats of the World* (Univ. Chicago Press, 2017).
57. Nel, J. A. J. *Handbook of the Mammals of the World, Vol. 1: Carnivores* (Lynx Edicions, 2009).

58. Dunkel, A. et al. Nature's chemical signatures in human olfaction: a foodborne perspective for future biotechnology. *Angew. Chem. Int. Ed.* **53**, 7124–7143 (2014).
59. Moran, Y., Barzilai, M. G., Liebeskind, B. J. & Zakon, H. H. Evolution of voltage-gated ion channels at the emergence of Metazoa. *J. Exp. Biol.* **218**, 515–525 (2015).
60. Nei, M. & Rooney, A. P. Concerted and birth-and-death evolution of multigene families. *Annu. Rev. Genet.* **39**, 121–152 (2005).
61. Zhao, J., Teufel, A. I., Liberles, D. A. & Liu, L. A generalized birth and death process for modeling the fates of gene duplication. *BMC Evol. Biol.* **15**, 275 (2015).
62. Newman, T. & Trask, B. J. Complex evolution of 7E olfactory receptor genes in segmental duplications. *Genome Res.* **13**, 781–793 (2003).
63. Niimura, Y., Matsui, A. & Touhara, K. Corrigendum: extreme expansion of the olfactory receptor gene repertoire in African elephants and evolutionary dynamics of orthologous gene groups in 13 placental mammals. *Genome Res.* **25**, 926 (2015).
64. Soso, S. B. & Koziel, J. A. Characterizing the scent and chemical composition of *Panthera leo* marking fluid using solid-phase microextraction and multidimensional gas chromatography–mass spectrometry-olfactometry. *Sci. Rep.* **7**, 5137 (2017).
65. Nosil, P. & Feder, J. L. Genomic divergence during speciation: causes and consequences. *Phil. Trans. R. Soc. B* **367**, 332–342 (2012).
66. Miga, K. H. & Sullivan, B. A. Expanding studies of chromosome structure and function in the era of T2T genomics. *Hum. Mol. Genet.* **30**, R198–R205 (2021).
67. Wold, J. et al. Expanding the conservation genomics toolbox: incorporating structural variants to enhance genomic studies for species of conservation concern. *Mol. Ecol.* **30**, 5949–5965 (2021).
68. Formenti, G. et al. The era of reference genomes in conservation genomics. *Trends Ecol. Evol.* **37**, 197–202 (2022).
69. Mérot, C., Oomen, R. A., Tigano, A. & Wellenreuther, M. A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends Ecol. Evol.* **35**, 561–572 (2020).
70. Lovell, J. T. et al. GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *eLife* **11**, e78526 (2022).
71. Li, G. et al. A high-resolution SNP array-based linkage map anchors a new domestic cat draft genome assembly and provides detailed patterns of recombination. *G3* **6**, 1607–1616 (2016).
72. Jégu, T., Aeby, E. & Lee, J. T. The X chromosome in space. *Nat. Rev. Genet.* **18**, 377–389 (2017).

Methods

Biological materials and genome sequencing

Fibroblast cell lines were established at the National Cancer Institute under protocols approved under contract N01-CO-12400. The parent-offspring trio of the Safari cat was composed of a random-bred domestic cat (*Felis silvestris catus*) dam, a Geoffroy's cat (*Leopardus geoffroyi*) sire and a female F1 offspring. Cell lines were karyotyped to confirm species identity and F1 status (Supplementary Fig. 31). The details of the Bengal cat F1 trio were previously reported[18,73,74]. The parent-offspring trio of the liger was composed of a tiger dam, a lion sire and a male F1 offspring (LxT-3). A karyotype of the F1 male liger was generated (Supplementary Fig. 32).

High molecular weight genomic DNA was extracted from cells using a modified salting-out protocol[75]. PacBio SMRT libraries were size selected (>20-kb) and sequenced on the Sequel IIe instrument to yield approximately 158× and 153× coverage for the Safari and Liger F1, respectively. The Bengal F1 reads[18] were sequenced on the Sequel I platform to 90× coverage.

Illumina fragment libraries (~300-bp average insert size) were prepared for the parent samples of trios using the NEBNext Ultra II FS DNA Library Kit (New England Biolabs). Samples were sequenced to ~30× coverage with 2 × 150-bp reads on the NovaSeq 6000 platform.

Hi-C library preparation and sequencing

Fibroblasts were fixed as a monolayer using 1% formaldehyde, divided into ~4.2×10^6 cell aliquots, snap-frozen in liquid nitrogen and stored at −80 °C (ref. 76). Cells were lysed, resuspended in 200 µl of 0.5× DNase I digestion buffer and chromatin digested with 1.5 units of DNase I for 4 min. Downstream library preparation was performed as described[76] and sequenced on the Illumina NovaSeq 6000 to ~78× coverage.

Genome assembly and annotation

Haplotype binning. All Illumina data were processed with FastQC v0.11.8 (ref. 77) and adapter trimming using Trim Galore! v0.6.4. Illumina sequences were unavailable for the parents of the F1 Safari cat. Therefore, we used the domestic cat parent (Fca-508) of the Bengal F1 hybrid and published Geoffroy's cat Illumina data (Oge-3: SRR6071645)[10] for phasing. Long reads were phased into haplotype bins using the trio-binning feature of Canu v1.8 (TrioCanu)[15,78].

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. Lion Y chromosome contigs were identified using published procedures[37].

Contig polishing and QC. NextPolish v1.3.0 (ref. 79) and NextDenovo corrected long reads were used to polish the raw contigs. 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 the conversion of consensus sequence to reflect the alternate haplotypes not present in the F1. The lgs options within the configuration file were 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 (ref. 80) with the --fast run option selected. BUSCO v4.0.6 (ref. 81) was used to assess genome completeness, with the -m genome setting with -l mammalia_odb10 database selected (9,226 single copy genes). Visual assessment of the assemblies was performed through alignment to the domestic cat assembly Fcat_Pben_1.0_maternal_alt (Fca-508: GCA_016509815.1)[18] using nucmer (mummer3.23 package)[82] with default settings. Delta files were used to generate dot plots using Dot: interactive dot plot viewer for genome–genome alignments (DNAnexus).

We also assessed assembly quality based on *k*-mer accuracy and completeness. Illumina data from each respective F1 hybrid were used to generate Meryl (v1.3) *k*-mer databases for the two parents and a child. Resulting Meryl databases were then used to generate hapmer databases using Merqury's (v1.3) hapmer script ($ sh $MERQURY/trio/hapmers.sh). The parental hapmer databases and child database were then passed to Merqury to evaluate assembly quality. We also assessed assembly quality using Inspector (https://github.com/Maggi-Chen/Inspector; v1.0.2).

Scaffolding. Polished contigs from the domestic and Geoffroy's cat were scaffolded using Hi-C data generated from the F1 Safari cat hybrid fibroblasts. Hi-C reads were binned into parental haplotypes prior to scaffolding by aligning the offspring reads to both polished parental assemblies using bwa mem v0.7.17 (ref. 83) and the classify_by_alignment (https://github.com/esrice/trio_binning/; v0.2.0) program as described in ref. 84. Haplotyped reads were mapped to polished contigs using the pipeline and scripts described in ref. 84 (https://github.com/esrice/slurm-hic/) using SALSA v2.2 (refs. 85,86) with parameters -e none -m yes. We removed all Y chr contigs prior to scaffolding to prevent incorporation of repetitive Y chromosome contigs into paralogous autosomal regions. Previously published Hi-C data for tiger (SRR8616865) and lion (SRR10075807/SRR10075808) (DNA Zoo[87]) were used to scaffold their respective assemblies with SALSA parameters -e GATC -m yes. The resulting scaffolds were inspected using QUAST, nucmer and Hi-C contact maps. RagTag v1.0.1 (ref. 88) was used to align scaffolds relative to Fcat_Pben_1.0_maternal_alt (Fca-508: GCA_016509815.1). Selected RagTag parameters included –remove-small, -f 10000 and -j unplaced.txt. RagTag scaffolds were manually inspected with Hi-C maps generated using Juicer v1.5.7 (ref. 89) with option -s for compatibility with DNase Hi-C libraries. Maps were visualized using Juicebox v1.11.08 (ref. 90) and Juicebox Assembly Tools with scripts from 3d-dna v.180922.

Genome annotation. The NCBI annotation pipeline provided the final assembly annotations used in our analyses. Identification and annotation of *DXZ4* repeat units were performed manually using GC content traces, CTCF motif annotations and self–self dot plots for the region using Geneious Prime v2021.0.3 and FlexiDot v1.06 (ref. 91). CTCF motifs were annotated using the Geneious Annotate & Predict tool with a sequence motif of GAGTTTCGCTTGATGGCAGTGTTGCACCACGAAT, based on the conserved CTCF motif logo[92], 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[93]. Independent repeat units were aligned using the Mafft Multiple Aligner v1.4.0, and maximum likelihood (ML) trees were generated with RAxML v8.2.11 (ref. 94) under a GTR+I+G model of sequence evolution. Trees were pruned using Mesquite v3.61 (ref. 95) and visualized using FigTree v1.4.4. Mean within- and between-group *P* distances for masked (10% gaps masked) *DXZ4* repeat unit alignments were calculated using Mega-X v10.0.5 (ref. 96). To compare the rate of *DXZ4* repeat evolution to the remainder of the genome, we created a multiple-sequence alignment with the domestic cat genome (Fca126) and Geoffroy's cat aligned to the tiger SHA reference. The alignment was passed to Tree House Explorer (v1.0.2)[97] where the THExBuilder function was used to calculate *P* distances in 10 kb windows with a strict missing data threshold of 0.0.

Comparative genomic analyses of *DXZ4* were assessed with contiguous long-read genome assemblies from all placental mammal superordinal clades[98] downloaded from NCBI. We chose male assemblies,

where available, due to their single chrX haplotype. Reference gene annotations for *PLS3* and *AGTR2* were used with Liftoff to identify the location of *DXZ4* (ref. 92). Centromere positions were identified using a combination of NCBI annotations, interspecific alignments and the Atlas of Mammalian Chromosomes, Seond Edition[99]. Dot plots were generated using FlexiDot. We determined the presence/absence of *DXZ4* based on the presence of repeat structure, CTCF-binding motifs and location relative to *PLS3* and *AGTR2*. Human, cat and pig *DXZ4* repeat monomers were also queried against the chrX using the discontiguous megablast BLAST algorithm.

Repetitive landscape, centromere annotation and analysis

Repeats. Repeats in each of the genomes were masked using RepeatMasker (RepeatMasker-4.1.2-p1; Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013–2015 <http://www.repeatmasker.org>) with the Dfam3.5+RepBase (rbrm-20181026) libraries where RepeatMasker was configured to use trf 4.09.1 for identifying tandem repeats, rmblastn (2.10.0+) to generate alignments and called with the -species cat option to mask using the cat-specific libraries. All repeats identified with that RepeatMasker run using the standard cat libraries were then masked as Ns, and RepeatModeler2 (ref. 100) (RepeatModeler 2 v2.0.2a; rmblast 2.10.0+; TRF 4.10, RECON, RepeatScout 1.0.6, RepeatMasker 4.1.2 = −p1; LTR Structural Analysis: Enabled (GenomeTools 1.6.2, LTR_Retriever v2.9.0, Ninja 1.10.2, MAFFT 7.453, CD-HIT 4.8.1)) was used to model additional repetitive elements with the LTRstruct option enabled (LTR_retriever v2.9.0 configured to use rmblast2.10.+; RepeatMasker; hmmer3.3.2; cdhit4.8.1). All identified repeats were masked, RepeatModeler2 was run again and the genomes were N-masked. Finally, to be certain the centromeres had been fully sampled, centromere regions from the final N-masked genomes were used as the input to RepeatModeler to create a final set of repeat models that were added to the Dfam3.5 + RepBase + the two previous rounds of RepeatModeler. The RepeatModeler consensus sequences were extended when the full repeat was not modeled and trimmed when the repeat model ran into a neighboring exon, concatenated, and redundancy was removed.

SDs. Before identifying SDs, repetitive elements identified using the RepeatMasker/RepeatModeler approach described above, as well tandem repeats identified by GRM[101] and ULTRA[102] (version 0.99.17ultra; using period=10, period=100 and period=4000), were masked. SDs were defined using SEDEF[28] with default parameters.

Centromeres. Initial outer bounds for the centromere region of each chromosome were defined by aligning known bounding markers[103] against each cat genome using blat[104]. The location was further refined by identifying human/cat synteny breakpoints by aligning each cat genome to the human genome (GCA_000001405.27_GRCh38.p12_genomic.fna) using nucmer[82] with default parameters, and then filtered using a 70% identity filter (delta-filter -i 70). Many felid chromosome arms are painted by separate human chromosomes using Zoo-FISH data, hence synteny breaks should define centromeric regions[105]. Reciprocal best alignments were extracted (show-coords -cT) and human/cat breakpoints were identified. To identify the centromere boundary, beginning at the human/cat alignment breakpoint, we move into the centromere analyzing the repeat density of Unknown+Satellite repeats in 25 kb windows in 1 kb steps. When that repeat density exceeded 0.3, we stepped 'back' to the base of the repeat density peak. To identify the position at which there was a significant change in the Unknown+Satellite repeat density, we identified the change point with a probability of at least 0.75 (ref. 106). From that point, we again walked 'away' from the centromere using a window size of 1.5 kb on the density of all repetitive elements that were enriched >500× within the centromere to incorporate any missed elements (density > 0.25) within 30 kb and to incorporate missed tandem repeats (repeat unit sizes 100 to 4,000, window size = 5k; density > 0.20). Finally, we checked that any boundary was between and not within a predicted gene.

Sensory receptor annotation and analysis. To identify both OR and V1R genes, we combined both the BLAST[107,108] and the Olfactory Receptor Assigner[53] algorithms into a single workflow. Initially, genomic regions containing putative sequences were identified by mapping a set of mammal-annotated ORG and V1R sequences, available via RefSeq, to each genome using blastn. A minimum of 85% sequence identity and 200 bp covered per hit were used to highlight potential sensory gene sequences and exclude nonspecific GPCR-like regions. Genomic regions for each hit were extracted with an additional 500 bp up and downstream to ensure start and stop codons were included. ORA uses a set of reference profile hidden Markov models (HMMs) to annotate ORG/V1R genes for each region extracted. Profile HMMs specific to V1Rs were generated using HMMR3 (ref. 109). ORG/V1R genes were classified as nonfunctional if they contained an in-frame stop codon or if they were less than 650 bp in length (that is, not long enough to complete the seven-transmembrane domain). Identified ORG/V1R sequences were mapped to the original RefSeq data to confirm they were definitive sensory genes. All ORG/V1R genes were mapped (blastn) between felid genomes to ensure no sequences were missing between species. ORA was used to classify all ORG and V1R genes into 13 subfamilies (OR1/OR3/OR7, OR2/OR13, OR4, OR5/OR8/OR9, OR6, OR10, OR11, OR12, OR14, OR51, OR52, OR55 and OR56) and eight subfamilies (V1R1, V1R2, V1R3, V1R4, V1R5, V1R48, V1R90 and V1R100), respectively.

ML gene trees per gene family per chromosome were inferred using IQTREE v.1.6.12(GTR+I+G)[110] based on multiple-sequence alignments generated with Clustal Omega[111]. The number of lineage-specific gene duplication events per species was estimated using Notung[112]. Additionally, by splitting gene trees into all possible subtrees via the 'ape' package in R[113], gene presence/absence per subtree was used to characterize putative one-to-one orthologs across species. Ambiguous orthologous relationships were further resolved using both genomic coordinates and blast hits. To determine if lineage-specific ORG/V1R gene duplications consisted of only the specific receptor gene or represented the duplication of a larger chromosomal region, 1,000 bp both up and downstream of each sequence was extracted and analyzed for SDs as described above.

Tiger recombination map. Publicly available short-read data for four individual *Panthera tigris jacksoni* (SRR7152390, SRR7152389, SRR7152391 and SRR715294) were trimmed, filtered and mapped to the *Panthera tigris* (P.tigris_Pti1_mat1.1) reference genome. Mapping results were evaluated and summarized using the Qualimap function bamqc[114]. Samtools[115] was used to remove duplicate reads. Base quality score recalibration was performed using GATK[116,117] by generating an initial reference set of SNPs from the dataset itself. Variants were then called, and all samples were jointly genotyped. Variants were filtered to remove variants in repeatmasked regions using GATK. Variants were further filtered, removing variants within 5 bp of an indel and those which did not meet the following quality criteria: -e'%QUAL<30 | INFO/DP<16 | INFO/DP>62 | QD<2 | FS>60 | SOR>10 | ReadPosRankSum <-8 | MQRankSum <-12.5 | MQ<40' in bcftools (https://github.com/samtools/bcftools). VCFtools (https://vcftools.github.io/man_latest.html) was used to remove indels, leaving 3,067,994 biallelic SNPs for further analysis. ReLERNN v.1.0.0, a deep learning approach that uses recurrent neural networks, was used to model the genome-wide recombination rate[118]. A mutation rate of 2.2×10^{-9} (ref. 119), was used. ReLERNN was run using the simulate, train, predict and bscorrect modules with default settings. Inferred recombination rates were averaged in 2 Mb blocks in 50 kb sliding windows.

Structural variant/inversion identification and analysis

Initial inversion call set detection with PAV. An initial variant call set was generated using PAV[120] (GitHub commit: 24efbea) with minimap2 (v2.24)[121] parameters '-x asm20 --secondary=no -a -t {params.cpu} --eqx -Y -B 2 -z 10000,50 --end-bonus=100' and PAV configuration settings 'inv_region_limit: 3000000', henceforth referred to as the PAV-mm2 call set. The 'sv_inv.bed.gz' bed files containing inversion calls for each sample were then used for downstream filtration and validation. As an additional line of validation, we also ran PAV using Long-Read Aligner (LRA) (v1.3.2)[122] with parameters '-CONTIG -p s -t'. The resulting 'sv_inv.bed.gz' inversion bed file was used for validation of the PAV-mm2 initial call set. Inversions overlapping regions identified as collapsed SDs identified by SDA[123] were removed from the analysis.

PBSV. CLR reads were mapped to Geoffroy's cat reference assembly (O.geoffroyi_Oge1_pat1.0) using pbmm2 (v1.9.0) using the parameters '--sort --median-filter'. The variant call set was generated using PBSV (v2.8.0) by first identifying signatures of structural variants using the discover command 'pbsv discover --tandem-repeats tandem_repeats.bed <input.bam> <output.svsig.gz>', where tandem repeats were identified by GRM and ULTRA. Then, variants are called using the call command 'pbsv call <reference.fasta> <output.svsig.gz> <output.vcf>'.

Sniffles. CLR reads were mapped to Geoffroy's cat reference assembly (O.geoffroyi_Oge1_pat1.0) using pbmm2 (v1.9.0) using the parameters '--sort --median-filter'. Variants were then called using Sniffles (v2.0.7)[124,125] with parameters '-t <cpu_count> -i <input.bam> -v <output.sniffles.vcf> --tandem-repeats <reference-repeats.bed>'.

Long-read mapping-based call set filtration. Call sets from PAV–LRA, PBSV and Sniffles were used to filter the initial PAV-mm2 call set by removing variants that were not supported by at least one of the three additional variant call sets. We utilized BEDTools (v2.30.0)[126] to call inversion variants with a 50% reciprocal overlap. Inversions identified on unplaced scaffolds were excluded. We identified large inversions (>500 kbp) not called by PAV with SafFire (https://github.com/mrvollger/SafFire). Input paf files were generated by mapping each assembly to the Geoffroy's cat reference assembly (O.geoffroyi_Oge1_pat1.0) with minimap2 (v2.24) with parameters '-x asm20 -t <cpu_count> -c --eqx' and then rustybam (https://github.com/mrvollger/rustybam - bioconda v0.1.31) parameters 'rb trim-paf sample.paf | rb break-paf --max-size 5000 | rb orient | rb filter --paired-len 100000 | rb stats --paf > sample.SafFire.bed'. Inversions greater than 500 kbp were called if supported by both SafFire- and Nucmer-based[82] dot plots.

Short-read genotyping and inversion classification. Pangenie (v1.0.1)[127] classified inversions as species/lineage-specific, paraphyletic with breakpoint use or polymorphic. Paired-end Illumina datasets for the lion ($n = 14$), tiger ($n = 14$), domestic cat (n = 10) and Asian leopard cat ($n = 10$) were downloaded from NCBI's SRA database and interleaved utilizing Seqkit's (v0.16.0)[128] concat function. The interleaved FASTQ files and fully-phased VCF files were then passed to Pangenie using the parameters '-u -s <sample_name> -o <sample_name> -i <sample_interleaved_fastq> -r <reference_assembly> -v <fully_phased_PAV_inversions.vcf>'. We could not genotype Geoffroy's cat-specific inversions using Illumina data. They were called if supported by inverted alignments to all query species. An initial phylogenetic matrix was constructed by merging inversions across all samples based on 50% reciprocal overlap (calculated by pybedtools v0.9.0)[126,129].

Annotation of SV-overlapping/containing SDs, gaps, genes and repetitive elements. Further, pybedtools (v0.9.0) intersected the breakpoint positions of the inversions with the coordinates of SDs, gaps, genes and repetitive elements. SciPy's (v1.7.3)[130] rank sum function (one-sided, greater) determined if inversions flanked by SDs were significantly larger than inversions not flanked by SDs. Inversions flanked by repetitive elements sharing more than 90% identity were identified using pandas (v1.4.0). Repetitive elements within 100 kb of the inversion breakpoints were aligned using biopython's (v1.79)[131] pairwise2.align.globalmx (upstream_seq, downstream_seq, 1, 0, score_only=True).

Statistics and reproducibility. The one-sided Wilcoxon rank sum test was used to determine differences in inversion sizes between the autosomes and chrXs. In this study, no statistical method was used to predetermine sample size, no data were excluded from the analyses and the experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Assemblies are available in NCBI under accession numbers GCA_016509475.2, GCA_016509815.2, GCA_018350155.1, GCA_018350175.1, GCA_018350195.2 and GCA_018350215.1. OR gene sequences and *DXZ4* alignments are found at: https://figshare.com/s/68266360874d5078bdf5.

Code availability

Publicly available software and packages were used in this study. No custom code was used. All software and packages used in this study are described within Methods section.

References

73. Menotti-Raymond, M. et al. A genetic linkage map of microsatellites in the domestic cat (*Felis catus*). *Genomics* **57**, 9–23 (1999).
74. Menotti-Raymond, M. et al. Second-generation integrated genetic linkage/radiation hybrid maps of the domestic cat (*Felis catus*). *J. Hered.* **94**, 95–106 (2003).
75. Miller, S. A., Dykes, D. D. & Polesky, H. F. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* **16**, 1215 (1988).
76. Ramani, V. et al. Mapping 3D genome architecture through in situ DNase Hi-C. *Nat. Protoc.* **11**, 2104–2121 (2016).
77. Andrews, S. FastQC. A quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).
78. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
79. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2020).
80. Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D. & Gurevich, A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**, i142–i150 (2018).
81. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
82. Marçais, G. et al. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
83. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

84. Rice, E. S. et al. Continuous chromosome-scale haplotypes assembled from a single interspecies F1 hybrid of yak and cattle. *Gigascience* **9**, giaa029 (2020).
85. Ghurye, J., Pop, M., Koren, S., Bickhart, D. & Chin, C.-S. Scaffolding of long read assemblies using long range contact information. *BMC Genomics* **18**, 527 (2017).
86. Ghurye, J. et al. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput. Biol.* **15**, e1007273 (2019).
87. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
88. Alonge, M. et al. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.* **20**, 224 (2019).
89. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
90. Robinson, J. T. et al. Juicebox.js provides a cloud-based visualization system for Hi-C data. *Cell Syst.* **6**, 256–258 (2018).
91. Seibt, K. M., Schmidt, T. & Heitkam, T. FlexiDot: highly customizable, ambiguity-aware dotplots for visual sequence analyses. *Bioinformatics* **34**, 3575–3577 (2018).
92. Horakova, A. H., Moseley, S. C., McLaughlin, C. R., Tremblay, D. C. & Chadwick, B. P. The macrosatellite *DXZ4* mediates CTCF-dependent long-range intrachromosomal interactions on the human inactive X chromosome. *Hum. Mol. Genet.* **21**, 4367–4377 (2012).
93. Chadwick, B. P. *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 (2008).
94. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
95. Maddison, W. P. & Maddison, D. R. Mesquite: a modular system for evolutionary analysis, v. 3.61. http://mesquiteproject.org (2019).
96. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
97. Harris, A. J., Foley, N. M., Williams, T. L. & Murphy, W. J. Tree house explorer: a novel genome browser for phylogenomics. *Mol. Biol. Evol.* **39**, msac130 (2022).
98. Murphy, W. J., Foley, N. M., Bredemeyer, K. R., Gatesy, J. & Springer, M. S. Phylogenomics and the genetic architecture of the placental mammal radiation. *Annu. Rev. Anim. Biosci.* **9**, 29–53 (2021).
99. O'Brien, S. J., Graphodatsky, A. S. & Perelman, P. L. *Atlas of Mammalian Chromosomes* (John Wiley & Sons, 2020).
100. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl Acad. Sci. USA* **117**, 9451–9457 (2020).
101. Vlahovic, I. et al. Global repeat map algorithm (GRM) reveals differences in α satellite number of tandem and higher order repeats (HORs) in human, Neanderthal and chimpanzee genomes—novel tandem repeat database. In *Proc. 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)* (IEEE, 2020).
102. Olson, D. & Wheeler, T. ULTRA: a model based tool. detect tandem repeats. *ACM BCB* **2018**, 37–46 (2018).
103. Davis, B. W. et al. A high-resolution cat radiation hybrid and integrated FISH mapping resource for phylogenomic studies across Felidae. *Genomics* **93**, 299–304 (2009).
104. Kent, J. W. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
105. Murphy, W. J. et al. A radiation hybrid map of the cat genome: implications for comparative mapping. *Genome Res.* **10**, 691–702 (2000).
106. Erdman, C. bcp: a package for performing a Bayesian analysis of change point problems. R package version 1.8.4. https://www.rdocumentation.org/packages/bcp/versions/1.8.4 (2007).
107. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
108. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0. 2013–2015. http://www.repeatmasker.org (2015).
109. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
110. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
111. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
112. Chen, K., Durand, D. & Farach-Colton, M. NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J. Comput. Biol.* **7**, 429–447 (2000).
113. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
114. García-Alcalde, F. et al. Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* **28**, 2678–2679 (2012).
115. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
116. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33 (2013).
117. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
118. Adrion, J. R., Galloway, J. G. & Kern, A. D. Predicting the landscape of recombination using deep learning. *Mol. Biol. Evol.* **37**, 1790–1808 (2020).
119. Kumar, S. & Subramanian, S. Mutation rates in mammalian genomes. *Proc. Natl Acad. Sci. USA* **99**, 803–808 (2002).
120. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
121. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
122. Ren, J. & Chaisson, M. J. P. lra: a long read aligner for sequences and contigs. *PLoS Comput. Biol.* **17**, e1009078 (2021).
123. Vollger, M. R. et al. Long-read sequence and assembly of segmental duplications. *Nat. Methods* **16**, 88–94 (2019).
124. Smolka, M. et al. Comprehensive structural variant detection: from mosaic to population-level. Preprint at *bioRxiv* https://doi.org/10.1101/2022.04.04.487055 (2022).
125. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
126. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
127. Ebler, J. et al. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat. Genet.* **54**, 518–525 (2022).
128. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS ONE* **11**, e0163962 (2016).
129. Dale, R. K., Pedersen, B. S. & Quinlan, A. R. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**, 3423–3424 (2011).
130. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

131. Cock, P. J. A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

Acknowledgements

We thank the High-Performance Research Computing Center at Texas A&M University for support. DNA library preparation and long-read sequencing were performed at the University of Maryland Institute for Genome Sciences (L. Tallon and L. Sadzewicz). Illumina sequencing was performed through the Texas A&M Institute for Genome Sciences & Society research core (A. Hillhouse). We thank R. Stanyon for the flow-sorted domestic cat chromosomes for FISH experiments. This research was supported by grants from the Morris Animal Foundation (grant D19FE-04 to W.J.M. and W.C.W.), the National Science Foundation (grants DEB-1753760 and DEB-2150664 to W.J.M.) and the National Institutes of Health (NIH; grant R01 GM59290 to M.A.B.). A.J.H. was funded, in part, by a training grant from the National Institute of General Medical Sciences, NIH (T32 GM135115). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

W.J.M. and W.C.W. were responsible for conceptualizing the project. K.R.B., L.H., A.J.H., G.H., N.M.F., C.L., R.C., J.M.S., E.R., B.W.D., T.R., L.A.L. and S.J.O. developed the methodology. K.R.B., L.H., A.J.H., N.M.F., G.H. and T.R. were involved in data visualization. Funding for the project was acquired by W.J.M. and W.C.W. W.J.M. and W.C.W. were responsible for project administration. Supervision of the project was provided by W.J.M., W.C.W. and M.A.B. The original draft of the manuscript was prepared by W.J.M., K.R.B., L.H., A.J.H., G.H., N.M.F. and W.C.W. All authors contributed to the investigation phase of the study and participated in the review and editing of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-023-01548-y.

Correspondence and requests for materials should be addressed to Wesley C. Warren or William J. Murphy.

Peer review information *Nature Genetics* thanks Michael Hiller and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): William Murphy

Last updated by author(s): Aug 9, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Genome sequence data was generated on Pacific Biosciences Sequel IIe and Illumina NovaSeq 6000 platforms using standard commercial software.
Data analysis	We used publicly available software/code for data analysis. We used the following programs and versions: FastQC v0.11.8, Trim Galore! v0.6.4, Canu v1.8, NextDenovo v2.2-beta.0, NextPolish v1.3.0, QUAST v5.0.2, BUSCO v4.0.6, mummer3.23, Merqury v1.3, Inspector v1.0.2, bwa mem v0.7.17, Samtools v1.9, BEDTools suite v2.30.0, R v.3.5.1, BLAST v2.9.0, RepeatMasker v4.0.9, seqTK subseq v1.3, SALSA v2.2, RagTag v1.0.1, Juicer v1.5.7, Juicebox v1.11.08, 3d-dna v.180922, Liftoff v1.4.2, Geneious Prime v2021.0.3, FlexiDot v1.06, Mafft Multiple Aligner v1.4.0, RAxML v8.2.11, Mesquite v3.61, FigTree v1.4.4, Mega-X v10.0.5, Tree House Explorer v1.0.2, RepeatMasker-4.1.2-p1, RepeatModeler 2 v2.0.2a, GenomeTools 1.6.2, LTR_Retriever v2.9.0, Ninja 1.10.2 , MAFFT 7.453, CD-HIT 4.8.1, hmmer3.3.2; cdhit4.8.1, ULTRA version 0.99.17, GRM, SEDEF, IQTREE v1.6.12, ReLERNNv1.0.0, PAV, minimap2 v2.24, Long-Read Aligner v1.3.2, PBSV v2.8.0, pbmm2 v1.9.0, Sniffles v2.0.7, SDA, SafFire, Pangenie v1.0.1, pybedtools v0.9.0, SciPy v1.7.3, pandas v1.4.0, biopython v1.79

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All genome assemblies are available in NCBI, under the following accession numbers: GCA_016509475.2, GCA_016509815.2, GCA_018350155.1, GCA_018350175.1, GCA_018350195.2, GCA_018350215.1. OR gene sequences are found at: https://figshare.com/s/68266360874d5078bdf5

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical method was used to predetermine sample size
Data exclusions	No data was excluded in our analyses.
Replication	All attempts to replicate results were successful
Randomization	The experiments were not randomized
Blinding	The investigators were not blinded to allocation during experiments and outcome assessment.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	The single haplotype genome assemblies were derived from DNA sequencing of primary skin fibroblast cell lines.
Authentication	All cell lines were karyotyped using standard methodology from early passages (p1-p3) and verified by comparison to known karyotypes and analyzed with PacBio long read technology to confirm species id.
Mycoplasma contamination	The cells were not tested for mycoplasma contamination
Commonly misidentified lines (See ICLAC register)	N/A