Full title: Quantifying structural variants in chromosomes using landmark-based disparity

Short title: Measuring structural variants using genomic landmarks

Authors and author information:

Ashwini V. Mohan[1,2]*, Anjali Goswami[2], and Jeffrey W. Streicher[2]*

1: Institut de Biologie, Université de Neuchâtel, Neuchâtel, Switzerland

2: Natural History Museum, London SW7 5BD, United Kingdom

* Authors for correspondence:

Ashwini V. Mohan (ORCID ID: 0000-0001-9505-0665)

Email: ashwini.venkatanarayana@unine.ch

Jeffrey W. Streicher (ORCID ID: 0000-0002-3738-4162)

Email: j.streicher@nhm.ac.uk

Abstract

Chromosomal architecture has played a key role in the evolution of biodiversity. Detecting structural variants (SVs) on chromosomes has informed the study of speciation, sex determination, adaptation, and some of the earliest divergences in the tree of life. Here we present a computationally non-intensive approach, based on geometric morphometrics, that uses conserved DNA sequences as landmarks to quantify structural disparities of focal chromosomes across multiple species, individuals, or cell types. Based on two approaches, we show that this 'geno-metric' method can be applied at micro- and macroevolutionary scales to discover and diagnose SVs. Using human X-linked genes and ultraconserved elements as landmarks, we provide empirical demonstrations with amniote sex chromosomes, the *Drosophila virilis* group, and placental mammal genomes. Landmark-based structural disparity analysis effectively identifies chromosomal rearrangements and has parallels with traditional morphometrics regarding chromosome size, landmark orientation and landmark availability. Using simulations, we show that structural disparity inferred from ultraconserved elements is correlated with overall levels of chromosome evolution; an attribute which is consistent with observed disparity between and within mammalian orders. We found that the disparity patterns of SVs have significant phylogenetic signal, giving them broad importance for studying evolutionary biology. Structural disparity analyses are a valuable addition to the comparative genomic toolkit in that they offer an intuitive, rapid mechanism for detecting SVs associated with single copy genetic landmarks and the potential to reveal broader patterns of chromosome evolution related to expansions, contractions, rearrangements and phylogeny.

Introduction

We are now firmly in the age of genomics with fully sequenced genomes commonplace for most organismal groups. Because of this, multi-species genomic comparisons are increasingly used in evolutionary and ecological research. This is even the case for organisms with relatively large nuclear genomes [1–3]. Excitingly, the comparison of whole genome sequences has offered valuable evolutionary insights into a range of topics including gene orthology [4], transposable element evolution [5], and adaptation to novel environments [6].

One important area of comparative genomics research is the study of chromosome architecture [7]. Detecting structural variants (SVs) of chromosomes is highly relevant to the evolution of biodiversity and disease [8]. From supergenes [9] to sex chromosomes [10], evolutionary shifts in chromosome structure are known to have a substantial influence on organismal phenotype. Recent research utilizing forward in time simulations has also indicated the important role of SVs in sustaining genomic adaptation [11]. Traditional methods for detecting SVs include chromosome banding and fluorescent in situ hybridization [12]. Several modern methods using genomic sequencing have also been developed including aberrant read alignment patterns, split reads, coverage changes, alignment of long-read sequences, and strand-seq [13–14]. Amongst sequencing-based methods, synteny assessment [15] is increasingly used to compare chromosome structure across multiple species with visual interpretation achieved through complex network diagrams, e.g. [16–19]. Most of these high-throughput DNA sequencing approaches for detecting SVs depend on comparisons to a single reference sequence and/or whole genome alignment, which can be subject to

reference-bias and time consuming, respectively [20]. To complement existing methods, here we present a method for detecting SVs across chromosomes from multiple species or individuals simultaneously that is reference-free, computationally non-intensive, and does not require whole genome alignment. Based on the principles of morphometrics, our method is applied to chromosomes from high-quality genome assemblies and facilitates the study of SVs in ordinated micro- and macroevolutionary contexts.

Within the study of comparative morphology, structural variation in homologous structures across species is commonly termed 'disparity' [21–22]. Measures of morphological disparity are routinely used to explore the evolutionary dynamics of complex phenotypes, e.g. [23–24]. Structural disparity analysis of chromosomes is a disparity-based method that uses conserved landmark sequences to capture genomic variation across different species in a single analytical framework. Our 'geno-metric' method is based on a simple premise: chromosome structure is morphology (i.e., a shape, form, or structure) and there are well-validated methods for capturing morphological disparity across multiple species [23–25]. Furthermore, genomes are filled with homologous sequences (=landmarks) making them amenable for such analyses (e.g. [26]). To develop and scrutinize landmark-based structural disparity analyses, we analyzed three empirical datasets. First, we quantified SVs associated with sex chromosome evolution across amniotes using human X-linked BUSCO (Benchmarking Universal Single-Copy Orthologs) genes as landmarks which represents a scenario when the structure of a particular genomic/chromosomal feature is of interest [sensu 18). Second, we compared SVs related to a speciation-associated chromosome in Drosophila [27] using ultraconserved elements (UCEs; [28]) as

landmarks which represents a scenario when chromosome-wide structure may be of interest. Third, we used genome sequences from placental mammals to develop a pipeline for discovering chromosomes that share homologous landmarks; a necessary step when chromosome homology is unclear or partial. We also used the placental mammal dataset to simulate SVs of chromosomes and determine the effects of indels and inversions on disparity measures. Finally, we tested all datasets for phylogenetic signals associated with disparity patterns of SVs.

Results

Pipeline development. We developed two approaches for conducting structural disparity analysis, [1] using generic feature format (gff3) annotation files as input and [2] a de novo method that allows users to landmark and analyze whole chromosome sequences from genome assemblies (Fig. 1). The gff3-formatted pipeline allows users to extract genes from the chromosome of a single species in the gff3 file and then identify orthologs in the other species contained within the file. This approach is useful when chromosome homology is known, and focal landmarks are largely confined to single chromosomes. The de novo method begins by acquiring chromosome sequences from NCBI and then reference mapping landmarks to the assemblies using bwa [29]. We tested other approaches for reference mapping including minimap2 using default settings [30] but found bwa resulted in the most mapped landmarks per chromosome (Fig. S1). After orthologous landmarks are mapped, landmarked chromosomes are either prepared for downstream analysis (when chromosomal homology is known), or clustered based on the presence/absence of landmarks via metric multi-dimensional scaling (when chromosomal homology is not known). The latter option allows for the

identification of chromosomes that share large numbers of landmarks, and thus likely large homologous genomic blocks. Following landmarking, both approaches have several steps in common including; removal of landmarks that are multi-copy or not shared across homologs, complementarity checks (5'-3' versus 3'-5'), and exportation of landmark position information into a spreadsheet. The spreadsheet has rows representing species or individuals (one for each species or individual included in the input assemblies) and columns are landmark positions on a homologous chromosome set. The spreadsheet is then used as input for the disparity inference stage. Depending on user preference, absolute landmark positions can be used, or positions can be transformed using landmark bounding and/or chromosome size correction. Unless otherwise indicated, all steps in the pipeline occur using base functions in the R statistical package [31]. As with other disparity analyses, we use ordination techniques to reduce the dimensionality of variation in landmark position wherein the spreadsheet of landmark positions is analyzed using principal component analysis (Fig. 1).

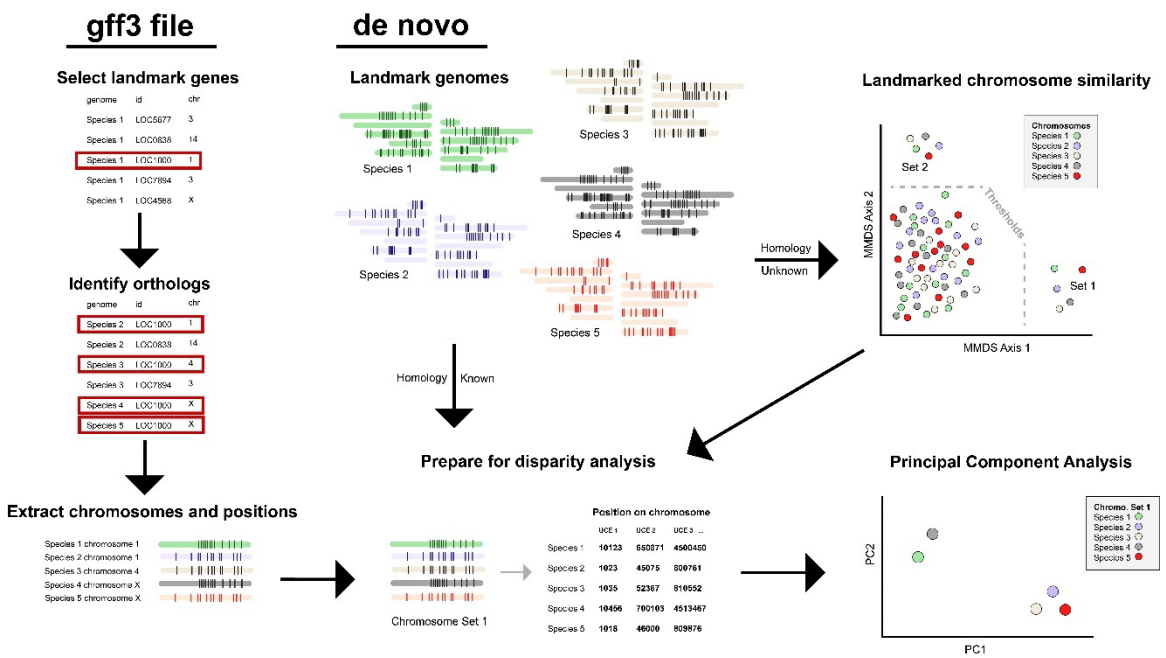


Figure 1. Schematic illustration of the landmarking pipeline for structural disparity analysis of chromosomes. When using generic feature format (gff3) annotation files, landmark genes are selected from a reference followed by identification of orthologs from other species in the file and extraction of chromosome information and landmark positions. In the de novo approach, the first step is reference mapping landmarks to chromosome-level genome assemblies of different species, individuals, or cell types. If the homology of chromosomes is known, then the chromosomes are further prepared for disparity analysis. If homology is unknown, a presence-absence matrix is used to identify chromosomes with similar landmark compositions using a gap-based threshold for similarity. Chromosome sets are prepared for disparity analysis by extracting chromosome position information for each landmark. The resulting data matrix is used for multivariate Principal Component Analysis to visualize disparity across species.

Structural disparity of human X-linked genes. Lovell et al. [18] examined 300 million years of vertebrate sex chromosome evolution using synteny-anchoring. We took their gff3-formatted annotation file and identified 53 human X-linked genes that could be used as landmarks on a set of single chromosomes for the focal taxa. We found that principal components (PCs) 1 and 2 explained 69.3% and 14.2% of structural variance, respectively. Using these first two axes, we found that major phylogenetic groups occupied distinct areas of geno-metric space (Fig. S2A). PC1 was significantly correlated with chromosome size (non-parametric Spearman's test, rho = 0.70, p = 0.003), but not PC2 (rho = -0.27, p = 0.3043). Mean structural disparity was highest amongst those groups where the landmarks were found on sex chromosomes (placental mammals, marsupial mammals and squamate reptiles) compared to those where these markers are located on autosomal chromosomes (birds, monotreme mammals). When the data were analyzed using the reverse complement orientation of the chromosomes, we found that some chromosomes with clustered landmarks

resulted in exaggerated PC scores for some taxa (e.g., chicken, *Gallus gallus*, Fig. S2B). We found that landmark bounding corrected for this methodological artifact regardless of the 5' to 3' orientation of the chromosomes (Fig. 2A, S2C, S2D). A size-corrected version of the analysis changed the patterns of disparity with birds being the most disparate group, followed closely by placental mammals (Fig. S2E). The size-corrected analysis was more robust to chromosome orientation with comparable results (Fig. S2F). Size-correction of the bounded version of the dataset generated indistinguishable results in the two directions of analysis (Fig. S2G, S2H). For comparison we generated a GENESPACE-style synteny link plot (Fig. 2B).

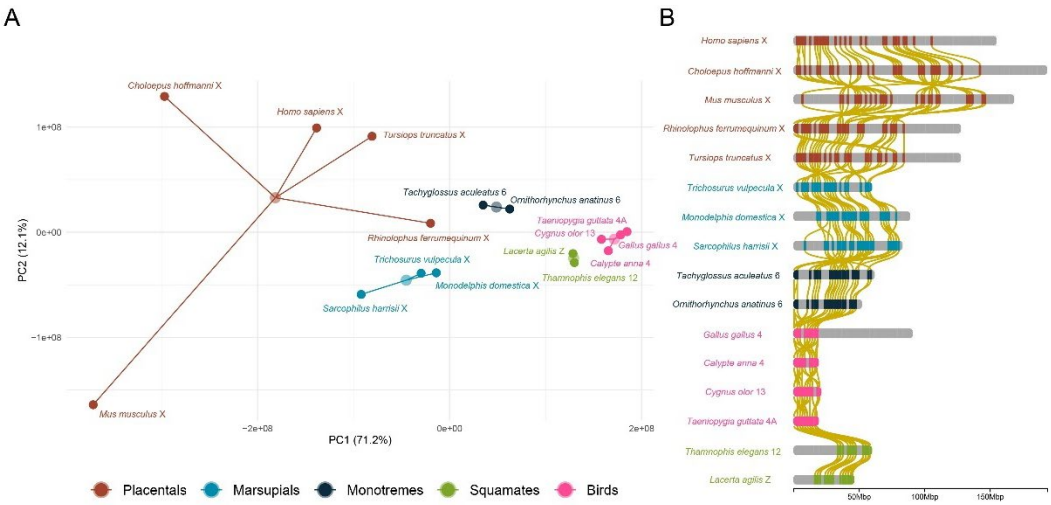


Figure 2. Structural disparity analysis of 53 human X-linked BUSCO gene landmarks in select amniote vertebrates studied by Lovell et al. (2022). Disparity analysis using bounded, absolute landmark positions reveals larger disparity in those groups where landmarks are found on the X chromosomes (placental mammals and marsupial mammals) than in those groups where landmarks are found on autosomes or Z chromosomes (A). This disparity is largely explained by chromosome size in that some of these groups maintain high synteny despite also having high structural disparity (B). Size-corrected disparity analyses capture different aspects of variation in landmark placement revealing birds as especially disparate after chromosome size is accounted for (Fig. S2E, F, G, H).

Structural disparity among closely related species of *Drosophila*. Poikela et al. [27] identified multiple chromosomal inversions that are associated with speciation patterns in the *Drosophila virilis* complex. We selected chromosome 5 for structural disparity analysis because there are known chromosomal inversions across the focal species [27]. We identified a set of 191 widely distributed dipteran UCEs [32] on chromosomes 5 for use as landmarks. Our structural disparity analysis revealed that three PC axes explained nearly 100% of the variance in landmark placement (PC1, 94%; PC2 5.2%; PC3, 0.07%). Patterns of disparity were highly polarized with PC1 separating *D. flavomontana* from the other species, PC2 separating *D. montana* from the other species, and PC3 scores separating *D. americana* + *D. novamexicana* from the two individuals of *D. virilis* (Fig. 3). Those individuals with highly conserved, syntenic UCE landmark placements occurred closest in geno-metric space. Notably, this was the case for the two individuals of *D. virilis* (Fig. 3A) and close relatives *D. americana* and *D. novamexicana* (Fig. 3B; [33]). In contrast, separation corresponded to distinct SVs. PC1 scores were largely explained by two blocks of UCE landmarks with substantially different placements in *D. flavomontana* compared to the other taxa (Fig. 3C). PC2 scores largely explained by a pericentric chromosomal inversion in *D. montana* relative to the other taxa (Fig. 3D). Importantly, when we compare the landmark orientations on *D. montana* and *D. flavomontana*, we see evidence (Fig. 3E) of the two blocks of UCE landmarks (PC1) and the pericentric inversion (PC2). PC3 scores were largely explained by a paracentric inversion that *D. virilis* has relative to both D. americana (Fig. 3F) and *D. novamexicana* (Fig. 3G). None of the three most explanatory axes were significantly

correlated with chromosome size (PC1, rho = 0.029, p = 1.000; PC2, rho = -0.371, p = 0.4972; PC3, rho = 0.600, p = 0.2417).

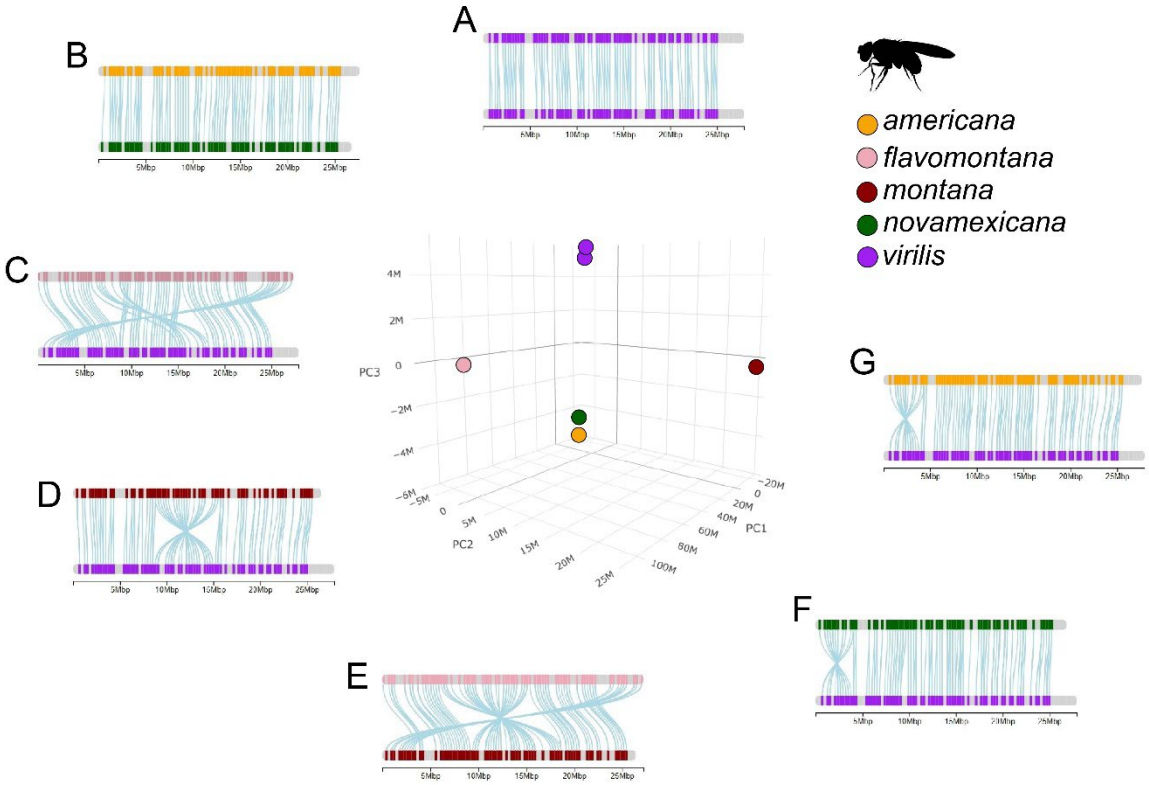


Figure 3. Structural disparity analysis using raw, absolute landmark positions of 191 ultraconserved elements (UCEs) on chromosome 5 of the *Drosophila virilis* group, different colors correspond to different species in the complex. Synteny link maps around the main plot demonstrate what separation in geno-metric space corresponds to, in terms of landmark placement on chromosome 5. When samples are close to one another in space they are highly syntenic (A, B). Different axes correspond to discrete differences in landmark placement including two translocations (C), a pericentric inversion (D, E) and paracentric inversion (E, F). An 'M' was arbitrarily added by plotly (91) to values on component axes because they are less than one. Synteny link maps were generated using chromoMap [35].

De novo homology discovery and structural disparity analysis using whole genome assemblies. Over long periods of evolutionary time, genome structure may change,

substantially eroding the homology of chromosomes. Despite this, large genomic regions often remain evolutionarily intact on single chromosomes (e.g., [16]) among which structural disparity can be explored. We used the chromosome clustering module of our de novo pipeline to identify several chromosome sets containing homologous tetrapod UCEs [34] from 26 genome assemblies of five placental mammal orders (Artiodactyla, Carnivora, Perissodactyla, Primates, and Rodentia; Table 1). The results of the 10-axis chromosome clustering revealed that score thresholding could be used to identify eight sets of mutually exclusive chromosomes that (a) were highly similar in their UCE compositions and (b) contained only a single chromosome per species in the dataset (Fig. 4A, Table 2). Chromosome similarity and threshold gaps decayed to after the second axis with not all sets containing 26 species and some instances of overlap with chromosomes belonging to other sets (e.g. *Giraffa tippelskirchi* chromosomes in sets 7 and 8, Fig 4A, Table 2). Axes 9 and 10 did not reveal any novel chromosome clusters associated with UCE landmark content (Table 2). Viewing axes as bivariate plots made some clustering patterns clearer, especially in axis 4 (Fig. 4B, C). Given (a) the clear separation of chromosomes sets 1 and 2, and (b) that protein content was significantly more similar within these chromosome sets than between them ($W = 6742$, $p < 0.0001$, nonparametric Wilcoxon tests, Fig. S3) we used these chromosomes as input for a structural disparity analyses based on SVs of their UCE landmarks.

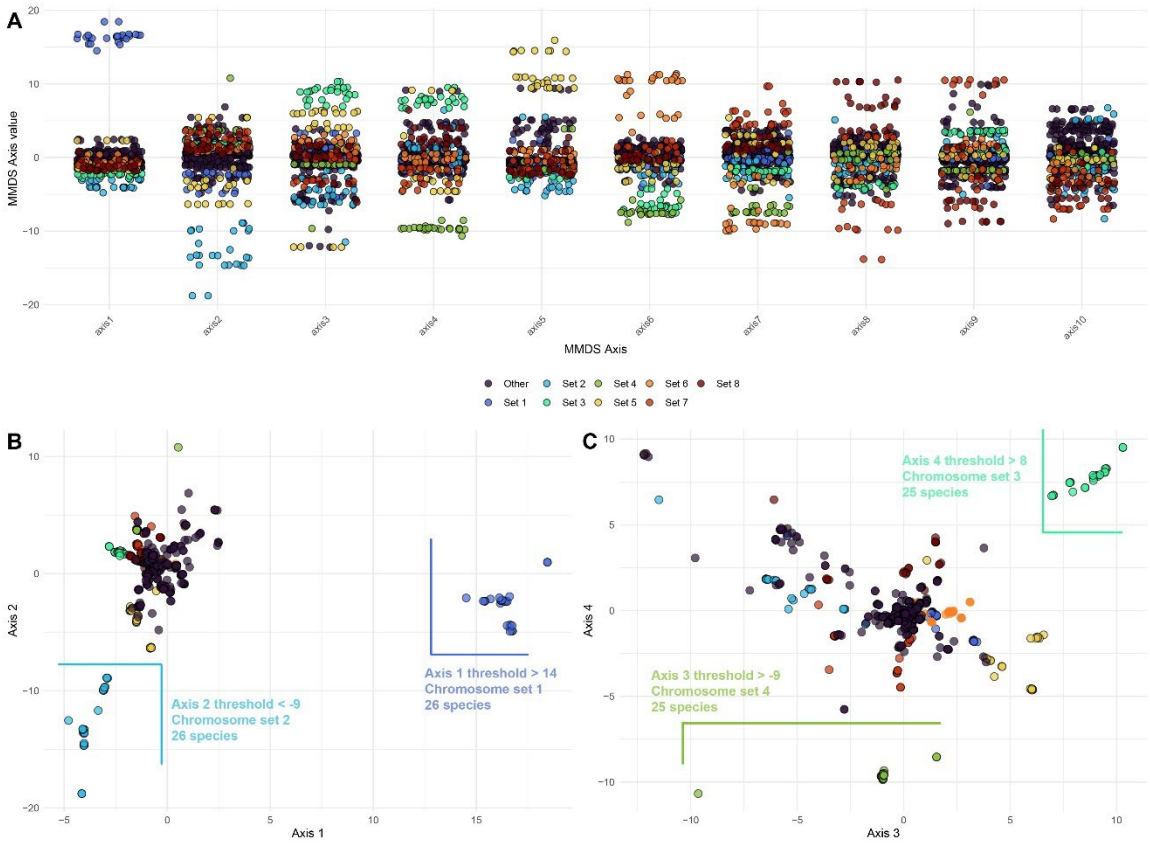


Figure 4. Metric multi-dimensional scaling (MMDS) of 567 landmarked placental mammal chromosomes based on patterns of ultraconserved element presence and absence. These chromosomes represent the genomes of 26 species (Table 1). MMDS axis thresholds are designated to establish 'chromosome sets' which contain a chromosome for most species included in the dataset. These chromosome sets likely contain mutually exclusive large homologous genomic blocks that are evolutionarily conserved and can be compared via structural disparity analysis. An initial assessment of 10 axes revealed eight sets of chromosomes that may be ideal for structural disparity analysis (A). Visualizing the clustering patterns in bivariate plots can assist with determining axis thresholds for defining the sets (B, C).

Table 1. Chromosome sets identified from genome assemblies used to conduct landmark-based estimates of structural disparity. Those chromosomes appearing in bold text were used in the simulation/validation section of our study. CN = chromosome number (= reference number used in genomic assemblies), bp = base pairs, UCE = ultraconserved elements. Following removal of an outlier landmark, the chromosome sets had 185 and 81 landmarks, respectively.

			Chromosome set 1 (185 UCEs)		Chromosome set 2 (81 UCEs)			
Species	Order	CN	Length	UCE landmarks	CN	Length	UCE landmarks	Reference
Bos indicus	Artiodactyla	2	140,680,885 bp	346	26	51,861,174 bp	235	[67]
Bos taurus	Artiodactyla	2	136,231,102 bp	353	26	51,992,305 bp	239	[68]
Bubalus bubalis	Artiodactyla	2	188,164,321 bp	453	23	52,095,120 bp	237	[69]
Capra aegagrus	Artiodactyla	2	128,023,632 bp	316	26	53,792,517 bp	233	[70]
Capra hircus	Artiodactyla	2	136,510,947 bp	356	26	51,421,553 bp	240	[71]
Giraffa tippelskirchi	Artiodactyla	3	190,444,849 bp	420	2	236,459,614 bp	562	[72]
Ovis aries	Artiodactyla	2	250,202,058 bp	569	22	51,512,491 bp	240	[73]
Felis catus	Carnivora	C1	221,611,373 bp	601	D2	88,083,857 bp	350	[74]
Neomonachus schauinslandi	Carnivora	3	192,837,575 bp	552	6	154,800,178 bp	565	[75]
Panthera tigris + P. leo	Carnivora	C1	219,880,392 bp	601	D2	86,999,983 bp	349	[76]
Ceratotherium simum	Perissodactyla	9	84,132,471 bp	322	6	90,918,853 bp	357	[77]
Equus asinus	Perissodactyla	4	92,916,174 bp	324	2	238,820,862 bp	688	[78]
Equus caballus	Perissodactyla	18	82,641,348 bp	320	1	188,260,577 bp	614	[79]

Gorilla gorilla	Primates	3	127,988,552 bp	368	11	127,481,205 bp	418	[80]
Macaca fascicularis	Primates	12	127,712,166 bp	360	9	129,235,353 bp	426	[81]
Macaca mulatta	Primates	12	130,043,856 bp	361	9	134,124,166 bp	427	[82]
Pan troglodytes	Primates	2B	128,755,405 bp	370	10	129,809,613 bp	421	[83]
Papio anubis	Primates	10	126,462,689 bp	358	11	125,913,696 bp	425	[84]
Cricetulus griseus	Rodentia	6	155,611,870 bp	515	3	282,827,514 bp	720	[85]
Mus caroli	Rodentia	2	171,410,997 bp	536	N/A	N/A	N/A	[86]
Mus musculus	Rodentia	2	181,755,017 bp	543	N/A	N/A	N/A	[87]
Mus pahari	Rodentia	3	164,141,553 bp	519	1	172,151,128 bp	531	[86]
Mus spretus	Rodentia	2	182,815,943 bp	537	N/A	N/A	N/A	[88]
Peromyscus maniculatus	Rodentia	4	154,712,973 bp	504	1	193,310,054 bp	533	[89]
Rattus norvegicus	Rodentia	3	169,034,231 bp	517	1	260,522,016 bp	640	[90]

Table 2. Axis thresholding based on metric multidimensional scaling (MMDS) of 567 placental mammal chromosomes from 26 species. Total number of species are unique chromosomes found to cluster on a given axis that are mutually exclusive when compared to clustering patterns of other axes. Threshold value is the smallest axis value identified in each group that is used to define the different clusters. Threshold gap value is the amount of space between the chromosome set and other chromosomes. An asterisk indicates two chromosomes sets where Macaca mulatta and/or

Giraffa tippelskirchi chromosomes from other sets were within the threshold range. See Figure 4 for additional information.

MMDS Axis	Number of species	Axis threshold	Threshold gap
1	26	14.52	12.03
2	26	-8.90	2.52
3	25	6.96	0.38
4	24	-8.54	2.78
5	26	9.44	0.03
6	26	3.40	0.72
7	26* (*Giraffa* set 5, *Macaca* other)	4.21	0.01
8	24* (*Giraffa* set 2)	3.60	< 0.01
9	0	0	0
10	0	0	0

Landmarking of UCES on the two chromosome sets resulted in 186 (set 1) and 82 (set 2) landmarks, respectively. Using absolute landmark positions (Figs. S4A, S5A), we found that most of the variation in both chromosome sets was explained by PC1. In chromosome set 1, PC1 described 98.9% of the variation and in chromosome set 2, PC1 described 99.4% of the variation. We discovered that in both chromosome sets, PC1 was significantly correlated with chromosome size (set 1, rho = -0.65, $p < 0.0001$; set 2, rho = 0.76, $p < 0.0001$) when using absolute size. However, this was not the case for PC2 (set 1, rho = 0.19, $p = 0.3308$; set 2, rho = 0.37, $p = 0.066$). Bounding the landmark positions increased the amount of variation explained by PC1 (chromosome set 1, 99.2%, Fig. S4C; chromosome set 2, 99.9%, Fig. S5C), and produced unexpected

patterns with several notable outliers across the first two PCs (e.g. Cricetulus griseus, Fig. S5C). Size-corrected analyses of chromosome set 1 resulted in each mammal order occupying a unique region of geno-metric across PCs 1 and 2 (Fig. S4E), but this was not the case in chromosome set 2 where orders largely overlapped (Fig. S5E). Size-correction using bounded landmarks resulted in less variation explained by PC1 for both chromosome sets (chromosome et 1, 89.2%, Fig. S4G; chromosome set 2, 91.6%, Fig. S5G), and similar outlier patterns that were observed in the raw bounded position analysis. Mapping of component loadings onto chromosome plots revealed a single landmark (UCE 6948) that was heavily loaded on PC1 and PC2 of both chromosome sets (Fig.S6). Removing this outlier landmark changed the inferred structural disparity patterns across the PCAs with mammalian orders mostly occupying distinctive regions of geno-metric space in both chromosome set 1 (Fig. S4B, D, F, H) and, apart from a single size-corrected analysis (Fig. S5F), chromosome set 2 (Fig. S5B, D, H).

Following the removal of the outlier landmark, we found that PC1 explained most variation across the analyses. When using raw, absolute landmark positions PC1 explained 99.9% of the variance for both chromosome sets (Fig. S7A, S8A). Unlike the analyses including the outlier landmark, using bounded landmarks reduced the amount of variation explained on PC1 (95.3%, chromosome set 1, Fig. S7B; 98.6% chromosome set 2, Fig. S8B) compared to the raw positions. PC2 scores in the bounded analyses explained 2.0% (chromosome set 1) and 0.8% (chromosome set 2). Notably, the bounded landmark PC1 and PC2 separated all mammalian orders into distinct regions of geno-metric space. Size-correction via chromosome size resulted in different disparity patterns in both datasets with higher disparity within Artiodactyla in the chromosome set 1 analysis (Fig. S7E) and substantial overlap across orders in the

289 chromosome set 2 analysis (Fig. S8E). The size-corrected, bounded analyses featured
290 tighter clustering patterns within mammalian orders than the raw, size-corrected
291 analyses in both chromosome sets (Figs. S7G, S8G). As observed in the X-linked BUSCO
292 analyses, we found that using raw landmark positions on the reverse complemented
293 version of the chromosome resulted in outlier taxa (*Macaca mulatta*, *Mus musculus*,
294 chromosome set 1, Fig. S7B; *Equus asininus*, *E. caballus*, *Cricetulus griseus*, *Mus*
295 *pahari*, *Peromyscus maniculatus*, *Rattus norvegicus*, chromosome set 2, Fig. S8B).
296 However, unlike the X-linked BUSCO analyses, these outliers were also observed in the
297 size-corrected analyses (Fig. S7F, S8F). Reverse complemented versions of the raw and
298 size-corrected bounded analyses were more consistent with their equivalent original
299 analyses (Fig. S7D, E, S8D, E). As observed in the X-linked BUSCO analyses, the amount
300 of variance explained by PCs 1 and 2 in the size-corrected, bounded analyses was
301 identical for both directions of analysis.

302 Based on our results, we selected the raw, bounded analysis with outlier landmark
303 removed for further interpretation and analysis (Fig. 5). Interestingly, although they
304 explained a small proportion of the overall structural variation, PC3 (0.5%, chromosome
305 set 1, 0.5% chromosome set 2) and PC4 (0.1% chromosome set 1, 0.09% chromosome
306 set 2) still clustered mammalian orders in largely distinct regions of geno-metric space
307 for both chromosome sets (Fig. S9). However, *Capra aegragus* was an outlier on PC3 of
308 chromosome set 1 (Fig.S9A). Synteny analysis of landmarks using chromoMap [35] to
309 visualize structural differences captured by UCE landmarks revealed that certain
310 chromosomal regions are more conserved as indicated by low PC loading values of
311 landmarks into PC1 (blue links, Fig. 6), compared to other regions where landmarks
312 show higher differences in their placement (maroon links, Fig. 6). Unlike the *Drosophila*

example where landmark inversions explained most structural disparity (Fig. 3), insertions and deletion events between the UCE landmarks account for the structural disparity observed in the two placental mammal chromosomes sets (Fig. 6).

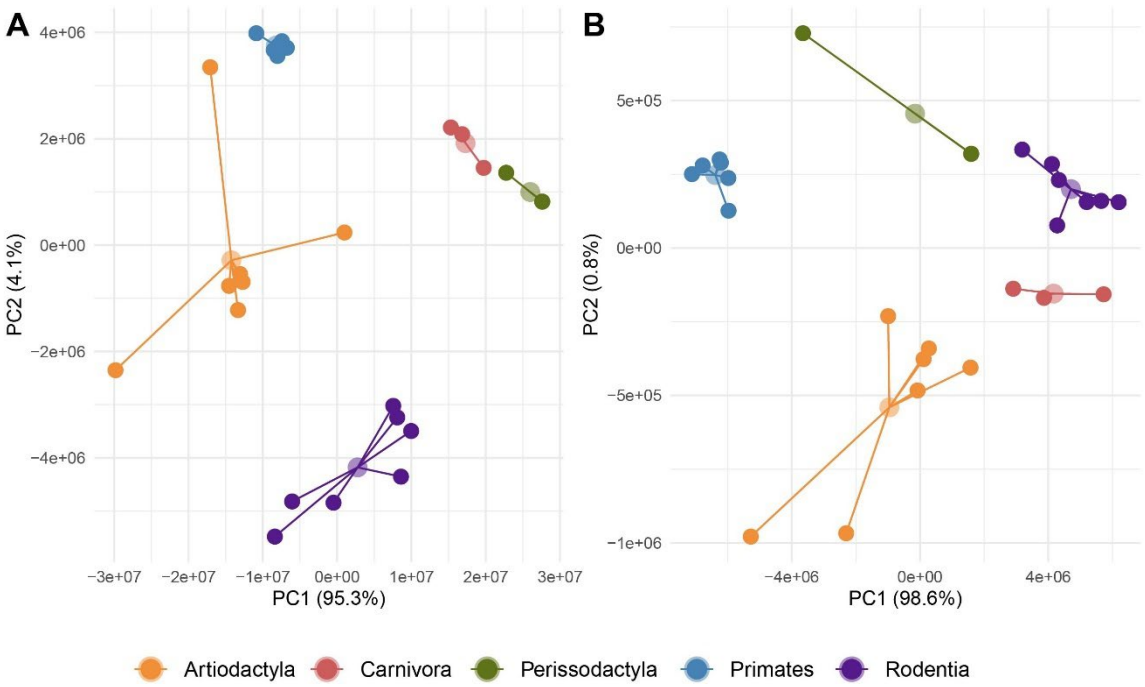


Figure 5. Ordination results from principal component (PC) analysis of UCE landmark positions in two sets of chromosomes from 26 species of placental mammals using bounded raw positions after the removal of an outlier landmark (see text). Representatives for five orders (Artiodactyla, Carnivora, Perissodactyla, Primates and Rodents) were included (Table1). The depicted structural variation is inferred from 185 (chromosome set 1, A) and 81 (chromosome set 2, B) landmarks, respectively. In both datasets, the first two PCs explained > 99% of the variation. Comparison of PCs 3 and 4 for both chromosome sets is provided in Figure S9.

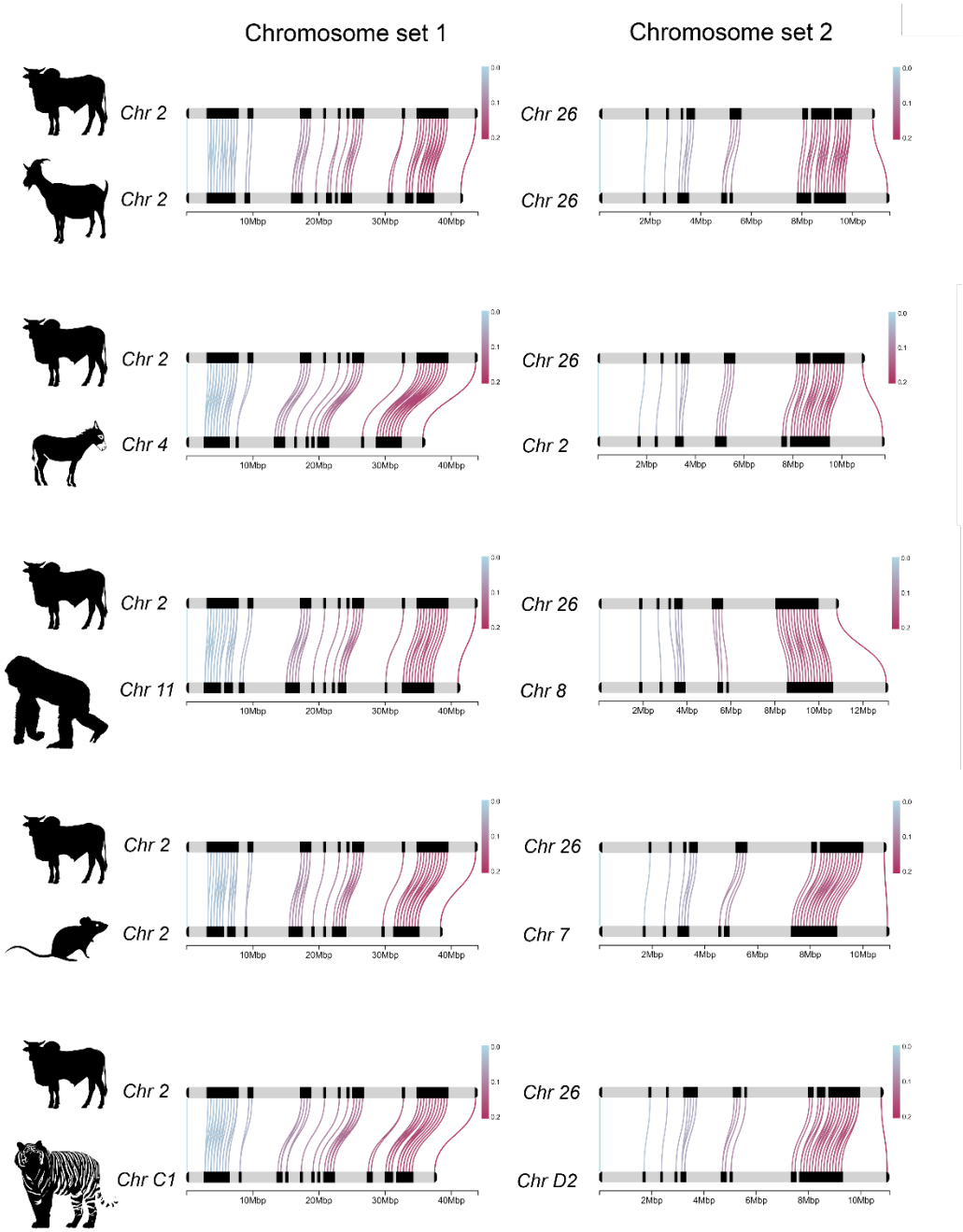


Figure 6. Select synteny link maps between *Bos indicus* and other placental mammals using bounded, absolute landmarking. Principal component loadings for various landmarks are indicated by the color of the links. The chromosome number (Chr) is indicated next to each species' silhouette. The left and right columns are results from chromosome set 1 and set 2, respectively. Comparisons are (from top to bottom) between *Bos indicus* to *Capra aegagrus* (Artiodactyls), *Bos indicus* and *Equus asinus* (Perissodactyla), *Bos indicus* and *Gorilla gorilla* (Primates), *Bos indicus* and *Mus musculus* (Rodentia), and *Bos indicus* and the liger (*Panthera tigris* + *P. leo*, Carnivora). Synteny link maps were generated using chromoMap [35].

Structural disparity increases through time. In the simulation study, > 99% of the variation in landmark placement was explained by PC1 with PC2 explaining 0.2% of the variation. As predicted, pairwise increase in disparity (ID) increased with the amount of simulated evolution confirming that, at least under neutral (random) conditions, disparity measures should correspond to relative amounts of genomic evolution (Fig. 7A-D). We compared the variance of log-transformed ID scores for indel versus inversion simulations using two-way ANOVAs. We found that on both axes, inversions caused a significantly higher increase in disparity than indels (PC1, $F = 2.88$, $p = 0.0251$; PC2, $F = 3.34$, $p = 0.0121$; Fig. S10). Using empirical data, we found that inter-order disparity was significantly higher than intra-order disparity for both PC1 (chromosome set 1, $W = 63997$, $p < 0.0001$; chromosome set 2, $W = 67457$, $p < 0.0001$, Fig. 7E, F) and PC2 (chromosome set 1, $W = 67997$, $p < 0.0001$; chromosome set 2, $W = 61201$, $p < 0.0001$, Fig. 7G, H). We found that protein-coding content tended to be more similar within mammalian orders than between them, however, this was not a significant result ($W = 499.5$, $p = 0.0722$; nonparametric Wilcoxon tests; SI Appendix, Fig. S11).

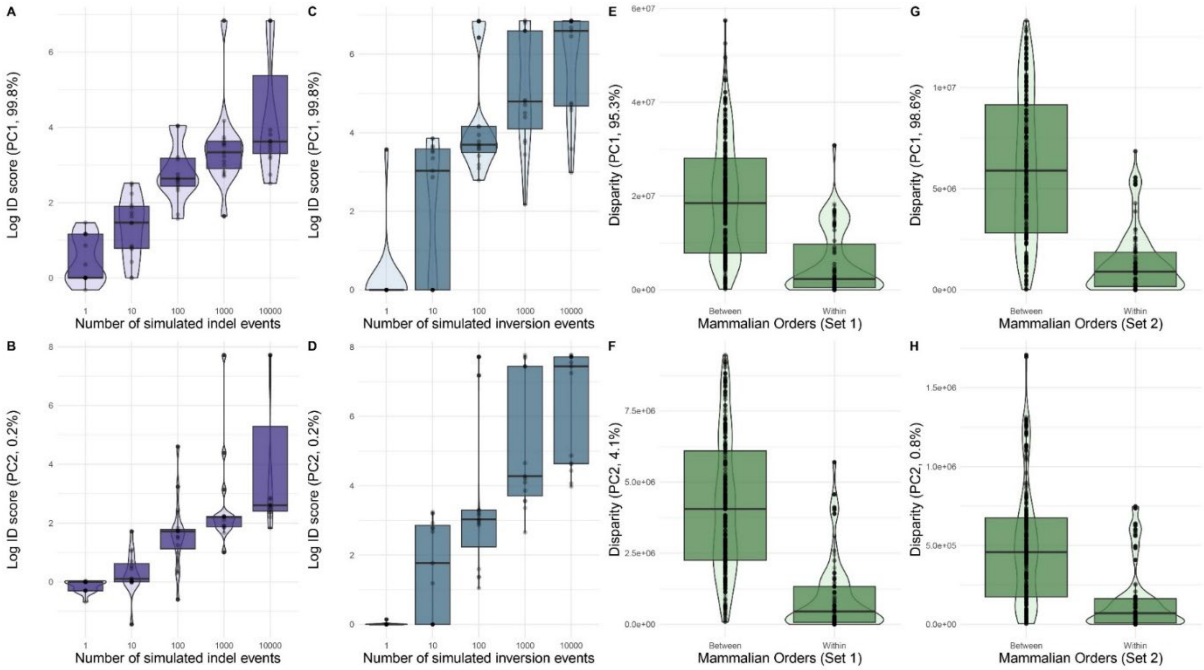
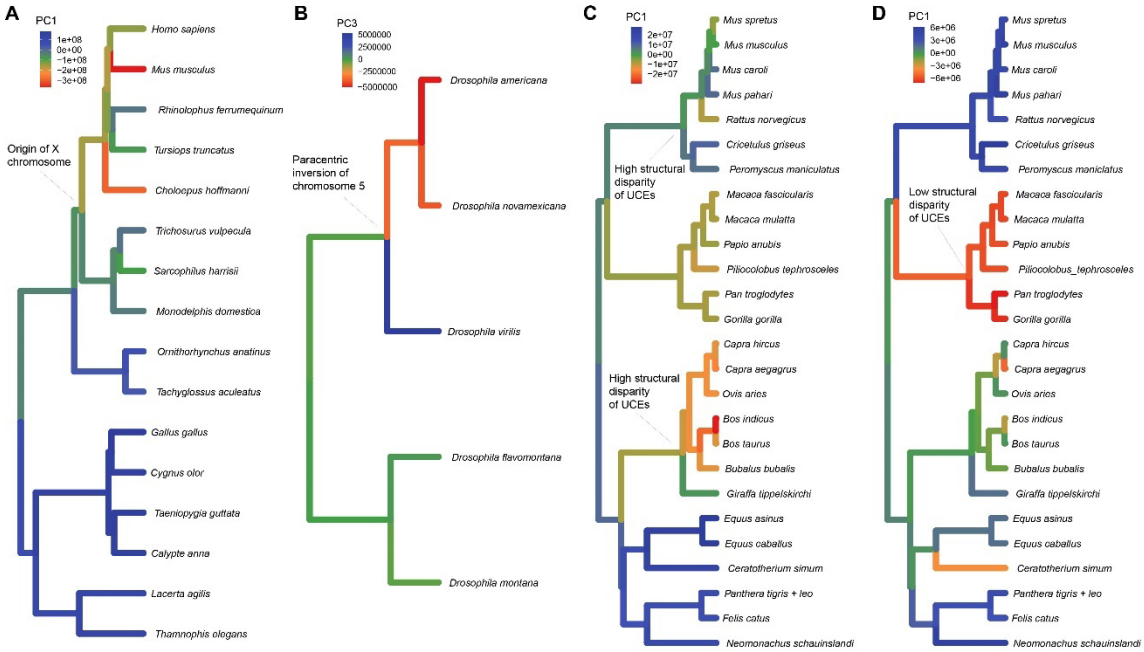


Figure 7. Log-transformed increase in disparity (ID) scores inferred from principal component (PC) analysis of ancestral and descendent insertion-deletion simulations for PC1 (A) and PC2 (B). Log-transformed ID scores inferred from inversion simulations for PC1 (C) and PC2 (D). Insertion-deletion simulation results are depicted in purple and inversion simulation results are depicted in blue. Observed empirical disparity scores for within and between mammalian orders for PC1 (E) and PC2 (F) of chromosome set 1 and PC1 (G) and PC2 (H) of chromosome set 2 are depicted in green.

Phylogenetic signal of structural variants. We found that the structural disparity captured by the most explanatory axes in all datasets had significant phylogenetic signal in 58% of the test cases (14/24 tests, Table 3). This included some cases where PC3 (describing < 1% of the overall variation) had significant phylogenetic signal (e.g. X-linked BUSCO genes, placental mammal UCE chromosome sets). Values of K and λ were also high (indicative of putative phylogenetic signal) in several non-significant tests. Ancestral state estimations of disparity revealed an increase in structural disparity related with the evolution of an X chromosome in the ancestor of placental and marsupial mammals (Fig. 8A), the evolution of a pericentric inversion on

366 chromosome 5 in the *Drosophila virilis* group (Fig. 8B), and phylogenetically associated

367 patterns of structural disparity increase or decrease in placental mammals (Fig. 8C, D).



368

369 **Figure 8.** Ancestral state estimation of structural variants (SV) using select principal component (PC)

370 scores from disparity analyses conducted herein. Structural variants detected using disparity analysis

371 often have phylogenetic signal (Table 2) and can be used to quantitatively study the rate of SV

372 evolution. Examples from this study include human X-linked BUSCO gene disparity of amniote

373 vertebrates where structural disparity increases following the evolution of the X chromosome (PC1,

374 A), the evolution of a speciation-associated paracentric inversion in chromosome 5 of the *Drosophila*

375 *virilis* group that caused rearrangements in the position of UCE landmarks (PC3, B), and the

376 identification of placental mammal orders with the least (chromosome set 1, PC1, C) and most

377 (chromosome set 2, PC1, D) conserved UCE structure, respectively. Phylogenies were obtained from

378 Kumar et al. [63].

379 **Table 3.** Results of phylogenetic signal tests using K [64] and λ [65] statistics for three most

380 explanatory principal components (PCs) of structural variant disparity from the empirical analyses

381 conducted herein. P-values are based on 1,000 randomizations (K) and likelihood ratio tests (λ).

382 Significant tests are indicated with bold text.

Principal component	Variance explained (%)	Phylogenetic signal	
		K	λ
X-linked BUSCO landmarks from amniotes: PC1	71.2	**1.3418 (p = 0.001)**	**1.160 (p < 0.001)**
X-linked BUSCO landmarks from amniotes: PC2	12.1	0.4237 (p = 0.501)	< 0.001 (p = 1.0)
X-linked BUSCO landmarks from amniotes: PC3	7.5	**0.6591 (p = 0.049)**	0.825 (p = 0.454)
UCE landmarks from *Drosophila*: PC1	94.0	0.6935 (p = 0.568)	**1.151 (p = 0.013)**
UCE landmarks from *Drosophila*: PC2	5.2	**0.9949 (p = 0.027)**	**1.151 (p = 0.004)**
UCE landmarks from *Drosophila*: PC3	0.07	0.6831 (p = 0.195)	< 0.001 (p = 1.0)
UCE landmarks placental mammals (set 1): PC1	95.3	**0.2173 (p = 0.022)**	**0.866 (p < 0.001)**
UCE landmarks placental mammals (set 1): PC2	4.1	**0.2600 (p = 0.005)**	**0.818 (p < 0.001)**
UCE landmarks placental mammals (set 1): PC3	0.5	0.0643 (p = 0.717)	**0.476 (p = 0.004)**
UCE landmarks placental mammals (set 2): PC1	98.6	**0.2677 (p = 0.009)**	**0.849 (p < 0.001)**
UCE landmarks placental mammals (set 2): PC2	0.8	0.1104 (p = 0.294)	**0.733 (p < 0.001)**
UCE landmarks placental mammals (set 2): PC3	0.5	0.0573 (p=0.708)	0.659 (p = 0.235)

383

Discussion

Using a morphometrics-inspired approach we identified SVs associated with chromosome inversions, insertion-deletion events (=relative size change), and phylogeny. These qualities mirror the utility of morphological disparity analysis as applied to complex phenotypic structures. In a phenotypic context, disparity can be descriptive or serve as a proxy for phylogenetic affinities and ecology [22]. For example, disparity is widely used to understand adaptive radiation in the paleontological record [36]. Our simulation and empirical results suggest that, like classic morphological disparity, structural disparity of chromosomes increases through time (Fig. 7). In a genomic context, we are optimistic that disparity can play a similar role in understanding how chromosome architecture varies among species and can be used for quantitative comparisons to study organismal evolution and adaptation. Below we discuss the results of our empirical case studies, suggest guidelines for best practice when conducting structural disparity analysis, and present several future directions for applications and methods development.

Structural disparity and the evolution of chromosome architecture. Disparity analyses of SVs enable a broad understanding of how similar (or dissimilar) chromosome structures are across different species based on the distribution of conserved landmark sequences like BUSCO genes and UCEs. This analytical quality is particularly advantageous for studying large-scale shifts in chromosome organization across organisms relevant to the study of supergenes and homologous synteny blocks (sensu [16]). In our study we found that following the evolution of a key mammalian sex chromosome, structural disparity of 53 BUSCO genes increased compared to other

407 amniotes (Fig. 2A, Fig. 8A). This disparity increase appears to be associated with an
408 overall increase in chromosome size in those species where the landmark genes are
409 located on an X chromosome. Due to the cessation of regular recombination, size
410 increases and decreases in vertebrate sex chromosomes are well-documented [37].
411 When we scaled the positions by chromosome size, we observed that placental and
412 marsupial mammals were no longer the most disparate group (Fig. S2E, G), further
413 supporting chromosome size increase as the key driver of structural disparity among
414 BUSCO genes in the mammalian X chromosome. Following size-correction, structural
415 disparity analysis identified high disparity among the bird chromosomes analyzed,
416 which from viewing the synteny plot (Fig. 2B) appears to be related to several inversions
417 (within birds) associated with the landmarks. The ability for structural disparity analyses
418 to capture SVs associated with inversions in closely related species or populations was
419 also observed in our study of chromosome 5 from the *Drosophila virilis* group based on
420 191 UCE landmarks (Fig. 3). In this simple example, we observed that the first three
421 principal components all corresponded to distinctive changes previously linked to
422 speciation [27]. Although not all axes had statistically significant phylogenetic signal
423 (Table 3), small sample sizes may have contributed to this outcome (see 38) as even
424 those axes without significant signal (e.g. PC3) were consistent with phylogenetically
425 associated changes (Fig. 8B).

426 The de novo study of structural disparity using tetrapod UCEs with placental mammals
427 revealed order-specific disparity patterns in PCs 1 and 2 of both chromosome sets (Fig.
428 5). In addition to revealing clear phylogenetic patterns in UCE-related structural
429 disparity, these analyses also revealed that disparity patterns are order-specific across
430 different chromosomes (Fig. 8C, D). Specifically, using the chull function in the R

431 package ‘sp’ [39] we found that convex hull values for each mammalian order on PCs 1
432 and 2 were perfectly correlated across chromosomes set 1 and set 2 (Pearson’s
433 correlation coefficient = 0.996, p = 0.0003). This suggests that order-specific structural
434 disparity associated with UCEs may be similar across different genomic regions. Given
435 the importance of UCEs in development and gene regulation [40] these types of
436 genome-wide patterns are important to explore as additional genomic data become
437 available.

438 Best practices for structural disparity analysis. Based on comparative analyses of the X-
439 linked BUSCO genes from amniote vertebrates and UCEs from placental mammals, we
440 can recommend a set of best practices for structural disparity that will allow users to
441 avoid several analytical artifacts that we encountered. First, examining component
442 loadings for each landmark is important for identifying outlier landmarks that may
443 obscure more generalized signal of other landmarks. In our study of placental
444 mammals, we found that a single outlier (UCE 6948) explained most variation in the first
445 two PCs (Fig. S6). The presence of this UCE in both datasets demonstrates that it is not
446 single-copy and justifies removing these types of spurious landmarks prior to disparity
447 analysis. Second, using raw, absolute positions may lead to outlier taxa if the landmarks
448 are clustered at one end of a large chromosome. This phenomenon was observed in
449 both the X-linked BUSCO gene dataset and the placental mammals (Fig. S2B, S7B,
450 S8B). In some cases, scaling the positions based on chromosome size may resolve this
451 issue (Fig. S2F), whereas in others it does not (Fig. S7F, S8F). We found that the use of
452 bound landmark positions resolves this analytical artifact associated with the
453 directionality of the chromosome sequence. If size is an important parameter for
454 understanding structural disparity, we recommend using bound, absolute positions

455 (Fig. S2C, S2D, S7C, S7D, S8C, S8D). If patterns like inversions and landmark
456 translocations are of primary interest, we recommend using size-corrected and bound
457 landmark positions (Fig. S2G, S2H, S7G, S7H, S8G, S8H). Finally, if using the de novo
458 approach and clustering chromosomes with MMDS to identify homologous landmark
459 regions for analysis we recommend viewing various axes as bivariate plots (Fig. 4B, 4C)
460 where clustering may be more obvious than gap threshold sizes would suggest on the
461 individual axes (Table 2).

462 A complementary tool for studying structural changes in chromosomes. Dimensionality
463 reduction is already in wide use in comparative genomics (e.g., uniform manifold
464 approximation and projection, [41]). Structural disparity analysis is therefore a natural
465 companion method for assessing structural differences across the chromosomes of
466 different species. It complements existing methods like WGA and synteny analysis by
467 capturing a different dimension of how chromosomes vary across species. Synteny
468 analysis is a powerful tool for identifying similarities in gene order and arrangement.
469 While synteny analysis does capture spatial variation (unlike WGA), it is known to be
470 very sensitive to parameter settings [42]. Furthermore, some methods of synteny
471 inference are performed on post-WGA datasets meaning that structural information
472 about the whole chromosomes has been removed a priori [15]. In contrast, landmark-
473 based estimates of genome structure offer at least two qualities that syntenic analyses
474 lack (Fig. S12). First, minor chromosomal rearrangements will be interpreted as breaks
475 in synteny and ignored. In other words, inversions within syntenic regions not involving
476 all the contiguous markers will break synteny. Second, large gap regions between
477 contiguous orthologs can also break synteny (depending on parameter settings). The
478 latter example highlights another important difference; disparity analyses will capture

479 distance information between landmarks whereas synteny only provides landmark
480 (ortholog) order information.

481 Applications beyond diploid genomes and evolutionary biology. Apart from diploid
482 organisms such as vertebrates and insects, the structural disparity pipeline could also
483 be implemented on polyploid organisms when chromosome phasing has been
484 performed. Utilizing the different phased genome assemblies, the pipeline can identify
485 structural differences between haplotigs of the genome. In addition, the pipeline may
486 be useful for measuring structural evolution of pangenomes without the need for a
487 reference input genome. Instead, a class/category of genetic markers could be used as
488 structural landmarks. As a visualization tool, structural disparity analysis may
489 complement the routinely used pangenome graph building methods (e.g.[43]). Lastly, a
490 potential use of structural disparity analysis relates to cancer genomics. The genomes
491 of somatic cancer cells undergo rapid structural evolution and SVs in cancer cells are
492 very specific to the type of cancer [44]. SVs in cancer cells have an impact on gene
493 expression through modifying structural accessibility, DNA methylation (and more) in
494 the cis-regulatory regions of the genome [44], therefore the ability to identify genomic
495 rearrangement is valuable (especially when using long-read data).

496 Avenues to expand the method towards genomic disparity. Structural disparity analyses
497 could be developed for whole-genome comparisons. At present, the pipeline performs
498 chromosome to chromosome comparisons, limiting comparisons across genomic
499 regions that have undergone translocation, fission, or fusion events. With greater
500 awareness of the frequency of these events, weighting parameters could be defined to
501 measure global genomic disparity levels. While we limited our validation analyses to

using BUSCOs and UCEs as landmarks and a handful of (largely homologous) chromosomes, landmarks could be any conserved genetic sequences which can be accurately mapped. Variation in the placement of conserved sequences can be used to study colocalization of transcribed genes (e.g. BUSCO [45]), cis-regulatory elements, enhancers, promoting regions, and more. The ethos of the approach is also scalable to three-dimensional genome structure once that information is more broadly available for many species (see [46, 47]). The extensive literature on geometric morphometrics offers substantial guidance for how 3D comparisons may be performed in the future [25]. Landmark-based estimates of genomic disparity provide a rapid mechanism for comparisons of genome structure across species and populations, and we anticipate they will expedite discovery and study of novel structural patterns within comparative genomics.

Methods

Bioinformatics and identification of landmarks. The bioinformatics and computational analyses were performed on Crop Diversity HPC, described by Percival-Alwyn et al. [48]. The structural disparity pipelines are modular and can be modified for a variety of other uses. Three step-by-step tutorials along with custom R and shell scripts are available on GitHub (https://github.com/nhm-herpetology/genomic-disparity/blob/main/README.md). Structural disparity analysis can be conducted using a variety of genomic landmarks assuming they are single copy and orthologous (e.g. BUSCO genes [45], UCEs [32,34]). In this study, we used human X-linked BUSCO genes and UCEs as landmarks. We selected human X-linked genes because many of them are found in homologous blocks on a single chromosome across amniote vertebrate

525 animals [18] and UCEs because they are highly conserved, lack repeats, and widely
526 distributed in genomes [34, 49]. UCEs have also been implicated as relevant to the
527 structure of key developmental and transcriptomic pathways [49, 50], making variation
528 in their relative chromosomal placements of biological interest.

529 Sensitivity analyses related to chromosome size, complementarity and landmark
530 boundaries. Data transformations associated with object size are a well-studied
531 component of morphometrics [51]. In our datasets where structural disparity was
532 correlated with chromosome size, we conducted a series of analyses to establish best
533 practice for structural disparity analysis. We analyzed these datasets in four ways using:
534 (1) raw positions, (2) bounded, raw positions, (3) chromosome size-corrected positions,
535 and (4) bounded, size-corrected positions. We repeated each analysis using the reverse
536 complement version of the four datasets for a total of eight versions of the analysis.
537 Boundary setting was achieved for each landmarked chromosome using a two-step
538 procedure: (1) re-positioning all the landmarks relative to the first landmark and (2)
539 using the position of the last landmark to establish the size of the genomic region being
540 analyzed. Absolute size correction was performed with R functions that divided the total
541 chromosome size by each landmark position. Bounded size correction was achieved by
542 dividing the size of the bounded genomic region by each landmark position with R
543 functions.

544 BUSCO genes and amniote vertebrates. Using the BUSCO gene annotation file from
545 Lovell et al. [18], we extracted all the genes annotated in the human (*Homo sapiens*) X
546 chromosome. We utilized this list of x-linked genes and identified homologous
547 chromosomes in other species based on the proportion of the human x-linked genes

548 mapped on them. The other species included in the study were four placental
549 mammals (*Mus musculus*, *Rhinolophus ferrumequinum*, *Tursiops truncatus*, and
550 *Choloepus hoffmanni*), three marsupial mammals (*Sarcophilus harrisii*, *Monodelphis*
551 *domestica*, and *Trichosurus vulpecula*), two monotreme mammals (*Tachyglossus*
552 *aculeatus* and *Ornithorhynchus anatinus*), four birds (*Gallus gallus*, *Cygnus olor*,
553 *Calypte anna*, and *Taeniopygia guttata*) and two squamate reptiles (*Thamnophis*
554 *elegans* and *Lacerta agilis*). When there was more than a single chromosome that
555 shared human x-linked genes (e.g. *Tachyglossus, Ornithorhynchus*, and *Trichosurus*),
556 the chromosome with the higher proportion of genes was used to maximize the number
557 of landmarks available for structural disparity analysis. Lovell et al. (2022) also analyzed
558 the budgie (*Melopsittacus undulatus*) genome, however, because our pipeline did not
559 identify focal landmarks on a single chromosome, we did not include this species in the
560 structural disparity analysis. Using the gff3 file approach (Fig. 1), we identified human x-
561 linked BUSCO genes that were homologous across all the species' chromosomes.
562 Detailed instructions for identifying and extracting these genes are available in Step 1 of
563 the GitHub tutorial. Next, chromosomes are prepared for downstream analysis using a
564 pipeline which requires the following R statistical software [31] packages: rehshape2
565 [52], matrixStats [53], tidyr [54], dplyr [55], and ggplot2 [56]). From 838 genes we
566 identified 53 that were present in all focal species making them suitable for use as
567 landmarks in a structural disparity analysis. Detailed instructions for identifying and
568 extracting landmarks are available in Step 2 of the GitHub tutorial. We also conducted
569 data matrix manipulation and principal component analysis in R, with code available in
570 Step 3 of the GitHub tutorial. For size-corrections, we obtained chromosome sizes from
571 NCBI. We tested for correlations between chromosome size and disparity using

572 Spearman's rho nonparametric tests in R. To create a synteny plot for comparative
573 purposes, we used the Chromomap package in R [35]. A list of the 53 human X-linked
574 BUSCO genes used as landmarks can be found in Appendix S1.

575 Drosophila structural variants, speciation and UCEs. Based on the findings of Poikela et
576 al. [27], we landmarked chromosome 5 for *D. americana*, *D. flavomontana*, *D. montana*,
577 *D. novamexicana*, and two individuals of *D. virilis*. This chromosome has known
578 inversions and translocations and may have contributed to speciation in the group. We
579 downloaded chromosome 5 from the *Drosophila virilis* group either via NCBI Entrez
580 Direct UNIX E-utilities (2) or directly from the literature (1). The following chromosomes
581 sequence were used: *D. americana* (NCBI accession no: CM061086.1), *D.*
582 *flavomontana* (1), *D. montana* (1), *D. novamexicana* (CM061080.1), and two individuals
583 of *D. virilis* (CM017608.2 and CM061075.1). Detailed instructions for downloading
584 chromosome sequences are available in Step 1 of the GitHub tutorial. For reference
585 mapping of ultraconserved elements (UCEs), we utilized a published set of
586 ultraconserved elements (UCEs) from dipterans [3]. We indexed the chromosome
587 assemblies and reference mapped UCEs using default settings in BWA-MEM [29] and
588 SAMtools [57]. Custom shell commands are used to loop the reference mapping and
589 create a spreadsheet that contains presence and absence information for the UCEs on
590 each chromosome 5 variant included in the analyses. Detailed instructions on this
591 reference mapping approach are available in Step 2 of the associated GitHub tutorial.
592 Next, chromosomes were prepared for downstream analysis using an R pipeline as
593 described above. The pipeline includes three steps (i) remove mapped UCEs that the
594 chromosomes do not share, (ii) complementarity checks to ensure all chromosomes
595 have comparable orientations (5' to 3'), and (iii) merging redundant UCE probes that

target the same UCE locus. The pipeline resulted in the identification of 191 homologous UCE landmarks for chromosome 5 of the *Drosophila virilis* group. Detailed instructions for the preparation pipeline are available in Step 3 of the GitHub tutorial. In the final step of the disparity analysis, we conducted a principal component analysis (PCA) and visualized the results. The input file for the PCA is available on GitHub as 'Drosophila_PCA.csv'. We used R base functions along with dplyr to format the data matrix and conduct the PCA. Visualization was conducted using ggplot2 [56] and cowplot [58]. Detailed instructions for this final step are available in Step 4 of the GitHub tutorial.

De novo landmark discovery and placental mammal genomes. To apply our disparity pipeline to a group where homology among chromosomes was unknown, we used 26 nuclear genomes from five orders of placental mammals (Table 1). Because placental mammal orders have diverged over millions of years [59] the homology of their chromosomes is fragmentary, making them a good candidate for exploring the extent to which disparity analysis can be applied to single genomic regions in divergent groups. We downloaded mammal genome assemblies using a custom script (chromosome_download.sh) that we developed to work with a standard input file (chromosome_download.config). Setting up a batch download for a subset of the mammal taxa (26 species) is described in detail in Step 1 of the GitHub tutorial. For reference mapping, we utilized a published set of tetrapod UCEs [32]. We indexed and reference-mapped UCEs on the 1,004 mammal chromosomes using default settings in BWA-MEM [29] and SAMtools [57]. Reference mapping was conducted using a custom shell script that loops through the chromosome sequences (landmark_mapping.sh). We tested the utility of minimap2 [30] for mapping conserved regions to the genome

620 assemblies utilizing default settings on chromosome set 2 of the mammalian genomes.
621 We then compared these results to those obtained using BWA MEM. We observed that
622 BWA MEM mapped higher numbers of UCEs than minimap2 (Fig. S1). This difference in
623 mapping quality could be explained by the short length of the UCE sequences used, at
624 about 180 bp length [32] because BWA MEM can achieve higher accuracy with short-
625 read mapping and produces less false negatives than minimap2 [30]. However,
626 minimap2 may be preferable if larger sequences were used as landmarks, and that
627 approach could then be easily combined with our post-mapping pipeline steps. In
628 addition to providing a UCE presence-absence output file for downstream analysis in
629 the pipeline, landmark_mapping.sh outputs the number of (UCE) landmarks mapped
630 and length information for each chromosome. Detailed instructions for refence
631 mapping are available in Step 2 of the GitHub tutorial.

632 Interestingly, we found that the number of UCEs mapped onto each mammal
633 chromosome was positively and significantly correlated with chromosome size ($r =$
634 0.75, $p < 0.05$; SI Appendix, Fig. S2). To identify sets of chromosomes that share
635 homologous landmarks, we applied metric multidimensional scaling (MMDS) on the
636 presence-absence matrix generated by landmark_mapping.sh. Using MMDS scores
637 across 10 axes we identified mutually exclusive sets of chromosomes that consisted of
638 only one chromosome from a species. These chromosome sets were identified by
639 selecting a threshold value on the relevant axes of the MMDS plot. Detailed instructions
640 for clustering chromosomes and thresholding are available in Step 3 of the GitHub
641 tutorial. To determine if protein content was similar on chromosomes sets (implying
642 that they share functional, protein-coding homologs in addition to UCE landmarks, we
643 used protein annotations from the Ensembl database (release 110; [60]), to generate

Jaccard similarity coefficients for comparing protein compositions on the two focal chromosome sets across a subset of placental species where annotations were available (N = 14). After extracting the protein-coding markers' identities from the Ensembl annotation files of each species (for the specific chromosomes in set 1 and set 2 separately), we utilized a custom defined Jaccard's similarity index which removed case sensitivity in the names of the markers (available on GitHub page as 'Jaccard_proteins.R'). This index estimated pairwise similarities between each chromosome in set 1 and set 2 respectively, as well as for all chromosome pairs in both the sets. The latter was done to test if chromosomes within a set show higher similarity in protein-coding markers than chromosomes between the two sets. Following the identification of chromosome sets, we prepared them for PCA by (i) removing UCEs that did not occur on all chromosomes, (ii) checking the directional orientation (5' to 3'), and (iii) merging redundant UCE probes. Unlike the highly conserved chromosomes from the Drosophila example, determining the directionality of mammal chromosomes was not always obvious. In other words, sometimes we observed mixed forward and reverse mappings of UCE landmarks, suggesting that only particular segments of the landmarked region were sequenced in the opposite direction of most taxa. This is likely explained by within-chromosome recombination and to account for it we used a majority rule criteria so that we only changed chromosome orientations if >50% of landmarks were mapped in the opposite direction of the reference chromosomes. Detailed instructions for the chromosome preparation pipeline are available in Step 4 of the GitHub tutorial. Finally, we conducted a PCA on each of the chromosome sets and visualized the results which are detailed in Step 5 of the GitHub tutorial. The input files for the PCA analysis are available on GitHub as 'Set_1_mammals_PCA.csv' and

668 ‘Set_2_mammals_PCA.csv’, respectively. As with the Drosophila example, applying the
669 disparity pipeline to placental mammal genomes required the following R packages:
670 rehshape2, matrixStats, dplyr, ggplot2 and cowplot.

671 Simulation and other tests for increased disparity through time. We simulated two
672 common evolutionary shifts in chromosome structure that occur as species and
673 populations diverge: insertion-deletion (indel) events and chromosomal inversions. We
674 performed these simulations using chromosome set 1 from three species subsampled
675 from the larger dataset, *Capra aegagrus*, *Mus pahari* and *Neomonachus schauinslandi*
676 (Table 1). We selected these taxa based on phylogenetic diversity (representing
677 Artiodactyla, Carnivora and Rodentia) and being outliers in some analyses (*Capra*
678 *aegragrus*, Fig. S9). We simulated evolutionary events using SimuG [61]. We designed
679 the experiment logarithmically, simulating 1, 10, 100, 1,000, and 10,000 random indels
680 or inversions on the ancestral chromosome of the five focal species. We ran five
681 replicated simulations for each of the five focal species. To assess the change in
682 disparity between ancestral chromosomes and their simulated descendants, we used a
683 simple pairwise metric of increase in disparity (ID):

684 $$ID = | PC^{A} – PC^{B} |$$

685 Where PC^{A} is the principal component score of the ancestral chromosome, PC^{B} is the
686 principal component score of the descendent chromosome, and ID is calculated by
687 taking the absolute difference of these values. As such, ID is an intuitive measurement
688 of how much landmark placement changed during the simulations, where a value of 0
689 indicates no change and larger values correspond to increasingly disparate landmark
690 placements. We predicted that if genomic disparity reflects the degree of evolutionary

change of architecture between two genomes, there should be a positive correlation between ID and the number of simulated events. While analysis of individual PCs may be problematic when data have high effective dimensionality [62], here the dominance of PC1 and the overwhelming correlation with chromosome size on that primary axis, justifies separate analysis of these components. We compared the original chromosome (set 1) sequence from three species of mammals, referred to as the 'ancestral chromosome', to simulated 'descendent chromosomes' which we made by simulating random indel or inversion events (see main text for more detail). Following the disparity pipeline described above, we mapped UCEs onto the simulated chromosomes, extracted homologous landmarks and carried out a PCA. PCA also included ancestral chromosomes from each species. We calculated the disparity between the ancestral chromosomes and simulated descendent chromosomes using the equation described in the main manuscript. Disparity was calculated using the logarithm of the absolute difference between the PC values of the ancestral chromosome and each simulated chromosome. Furthermore, we tested for differences between the disparity measures of indel-simulated and inversion-simulated chromosomes using two-way ANOVAs implemented in R.

To provide empirical context for the simulation results, we tested for the signature of disparity increase in our placental mammal datasets using nonparametric Wilcoxon tests in R. If UCE-inferred disparity increases with evolutionary time, we expected inter-order disparity to be significantly higher than intra-order disparity for the most explanatory PCs because there has been more evolutionary divergence between mammalian orders than within them. We also generated Jaccard similarity coefficients for comparing protein compositions between and within mammalian orders.

Tests for phylogenetic signal and ancestral state estimation. Using the three most explanatory structural disparity PCs, we tested for phylogenetic signal in the (1) amniote vertebrate, (2) *Drosophila virilis* group, and (3) placental mammal (both chromosome sets 1 and 2) datasets. Phylogenies for each dataset were obtained from Kumar et al. [63]. We used two common metrics for phylogenetic signal, K [64] and λ [65]. We implemented test using the phytools package [66] in R. We assessed the significance of the tests using 1,000 randomizations (K) and likelihood ratio tests of the observed data to that estimated under Brownian motion (λ). To place structural disparity of genomic regions in a phylogenetic context, we estimated putative ancestral disparity values we used the 'fastAnc' function in phytools for select PCs in each dataset.

Acknowledgments

We thank members of the Goswami Lab and NHM Herpetology Group for many helpful comments on experimental design and methods development. The authors acknowledge ChatGPT for assistance with some R coding (https://openai.com/index/chatgpt/). We thank Kate Thomas for an introduction to several R packages. Several delegates at the 2023 Society for Integrative and Comparative Biology (Austin, Texas, USA) and 2023 Evolution (Albuquerque, New Mexico, USA) and Young Investigators meeting 2024 (Bhopal, India) conferences provided valuable suggestions on earlier version of this work and putative applications for structural disparity analysis. We thank Jason Head and members of the Department of Zoology at the University of Cambridge for helpful suggestions related to size correction. We thank Christophe Liedtke for discussions on dimensionality reduction in genomics. We are indebted to Brant Faircloth and two anonymous reviewers for

comments that greatly improved the quality of this manuscript. We are grateful to Iain Milne for bioinformatics support. The authors acknowledge Research Computing at the James Hutton Institute for providing computational resources and technical support for the “UK’s Crop Diversity Bioinformatics HPC” (BBSRC grants BB/S019669/1 and BB/X019683/1), use of which has contributed to the results reported within this paper. This study was part of a project funded by the European Commission through Marie Skłodowska-Curie Actions to A.V.M. (Grant Agreement ID: 101027832). This work also received support from the Genomics Research Theme at the Natural History Museum, London.

References

1. Genereux DP, Zody MC, Shen W, McGaugh SE, Stemple DO, Roberts AC, et al. 2020. A comparative genomics multitool for scientific discovery and conservation. Nature 587:240–245. doi:10.1038/s41586-020-2874-7

2. Feng S, Stiller JW, Deng Y, Armstrong J, Burt DW, Kang H, et al. 2020. Dense sampling of bird diversity increases power of comparative genomics. Nature 587:252–257. doi:10.1038/s41586-020-2873-8

3. Osmanski AB, Maples B, Blahnik J, Griesmann L, Willis WM, Schmitz RJ, et al. 2023. Insights into mammalian TE diversity through the curation of 248 genome assemblies. Science 380:eabn1430. doi:10.1126/science.abn1430

4. Rubin GM, Yandell MD, Wortman JR, Clayton DF, Meeker S, Filipski AJ, et al. 2000. Comparative genomics of the eukaryotes. Science 287:2204–2215. doi:10.1126/science.287.5461.2204

5. Zhang X, Wessler SR. 2004. Genome-wide comparative analysis of the transposable elements in the related species Arabidopsis thaliana and Brassica oleracea. Proc Natl Acad Sci U S A 101:5589–5594. doi:10.1073/pnas.0401243101

6. Yuan Y, Liu X, Wang J, Sun Z, Smith AB, Peterson DG, et al. 2021. Comparative genomics provides insights into the aquatic adaptations of mammals. Proc Natl Acad Sci U S A 118:e2106080118. doi:10.1073/pnas.2106080118

7. Lynch M, Walsh B. 2007. The origins of genome architecture. Sunderland (MA): Sinauer Associates.

8. Krupina K, Goginashvili A, Cleveland DW. 2024. Scrambling the genome in cancer: causes and consequences of complex chromosome rearrangements. Nat Rev Genet 25:196–210. doi:10.1038/s41576-023-00574-6

9. Hu W, Xie J, Song X, Chen S, Ling H, Li Z, et al. 2020. Nondestructive 3D image analysis pipeline to extract rice grain traits using X-ray computed tomography. Plant Phenomics 2020:3414926. doi:10.34133/2020/3414926

10. Graves JAM. 2016. Evolution of vertebrate sex chromosomes and dosage compensation. Nat Rev Genet 17:33–46. doi:10.1038/nrg.2015.2

11. Banse P, et al. Forward-in-time simulation of chromosomal rearrangements: The invisible backbone that sustains long-term adaptation. Mol Ecol [Internet]. n.d.;(n/a):[page-unspecified].

12. Chaves R, Adega F, Santos S, Guedes-Pinto H, Heslop Harrison JS. 2002. [Title missing]. Cytogenet Cell Genet. 96:113–116. doi:10.1159/000063020

13. Balachandran P, Beck CR. 2020. Structural variant identification and characterization. Chromosome Res. 28:31–47. doi:10.1007/s10577-019-09623-z

14. De Coster W, Van Broeckhoven C. 2019. Newest methods for detecting structural variations. Trends Biotechnol. 37:973–982.

15. Krasheninnikova K, et al. 2020. halSynteny: a fast, easy-to-use conserved synteny block construction method for multiple whole-genome alignments. GigaScience. 9:giaa047. doi:10.1093/gigascience/giaa047

16. Larkin DM, et al. 2009. Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. Genome Res. 19:770–777. doi:10.1101/gr.087407.108

17. Chaney L, Sharp AR, Evans CR, Udall JA. 2016. Genome mapping in plant comparative genomics. Trends Plant Sci. 21:770–780. doi:10.1016/j.tplants.2016.05.003

18. Lovell JT, et al. 2022. GENESPACE: syntenic pan-genome annotations for eukaryotes. bioRxiv. 2022.03.09.483468. doi:10.1101/2022.03.09.483468

19. Quigley S, Damas J, Larkin DM, Farré M. 2023. syntenyPlotteR: a user-friendly R package to visualize genome synteny, ideal for both experienced and novice bioinformaticians. Bioinformatics Adv. 3:vbad161. doi:10.1093/bioadv/vbad161

20. Armstrong J, Hickey G, Diekhans M, et al. 2020. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature. 587:246–251. doi:10.1038/s41586-020-2871-y

21. Foote M. 1997. The evolution of morphological diversity. Annu Rev Ecol Evol Syst 28:129–152. doi:10.1146/annurev.ecolsys.28.1.129 philpapers.org+1onlinelibrary.wiley.com+1

22. Guillerme T, Cooper N, Martin RD, Wills MA. 2020. Disparities in the analysis of morphological disparity. Biol Lett 16:20200199. doi:10.1098/rsbl.2020.0199

23. Fabre AC, Botero-Castro F, Delsuc F, Douzery EJ, Goswami A. 2020. Metamorphosis shapes cranial diversity and rate of evolution in salamanders. Nat Ecol Evol 4:1129–1140. doi:10.1038/s41559-020-1200-4

24. Briggs DEG, Fortey RA, Wills MA. 1992. Morphological disparity in the Cambrian. Science 256:1670–1673. doi:10.1126/science.256.5064.1670

25. Zelditch ML, Lundrigan BL, Garland T Jr. 2004. Developmental regulation of skull morphology. I. Ontogenetic dynamics of variance. Evol Dev 6:194–206. doi:10.1111/j.1525-142X.2004.04025.x

26. Christmas MJ, et al. 2023. Evolutionary constraint and innovation across hundreds of placental mammals. Science 380:eabn3943. doi:10.1126/science.abn3943

27. Poikela N, Laetsch DR, Hoikkala V, Lohse K, Kankare M. 2024. Chromosomal inversions and the demography of speciation in Drosophila montana and Drosophila flavomontana. Genome Biol Evol 16:evae024. doi:10.1093/gbe/evae024

28. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, et al. 2004. Ultraconserved elements in the human genome. Science 304:1321–1325. doi:10.1126/science.1098119

29. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760. doi:10.1093/bioinformatics/btp324

30. Li H. 2018. Minimap2: pairwise alignment for nucleic acid sequences. Bioinformatics 34:3094–3100. doi:10.1093/bioinformatics/bty191

31. R Core Team. 2021. R: A language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. Available from: https://www.R-project.org/ faircloth-lab.orgbibsonomy.org+10ropensci.org+10scirp.org+10

32. Faircloth BC. 2017. Identifying conserved genomic elements and designing universal bait sets to enrich them. Methods Ecol Evol 8:1103–1112. doi:10.1111/2041-210X.12754 academic.oup.com+9faircloth-lab.org+9pmc.ncbi.nlm.nih.gov+9

33. Yusuf LH, Tyukmaeva V, Hoikkala A, Ritchie MG. 2022. Divergence and introgression among the virilis group of Drosophila. Evol Lett 6:537–551.

34. Faircloth BC, Branstetter MG, White ND, Brady SG, Fisher BL, Joubert MF, et al. 2012. Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales. Syst Biol 61:717–726. doi:10.1093/sysbio/sys004 arxiv.org+6faircloth-lab.org+6pmc.ncbi.nlm.nih.gov+6

35. Anand L, Rodriguez Lopez CM. 2022. ChromoMap: an R package for interactive visualization of multi omics data and annotation of chromosomes. BMC Bioinformatics 23:33. doi:10.1186/s12859-021-04516-6

36. Ruta M, Angielczyk KD, Fröbisch J, Benton MJ. 2013. Decoupling of morphological disparity and taxic diversity during the adaptive radiation of anomodont therapsids. Proc R Soc B 280:20131071. doi:10.1098/rspb.2013.1071

37. Schartl M, Schmid M, Nanda I. 2016. Dynamics of vertebrate sex chromosome evolution: from equal size to giants and dwarfs. Chromosoma 125:553–571. doi:10.1007/s00412-015-0569-y

38. Münkemüller T, et al. 2012. How to measure and test phylogenetic signal. Methods Ecol Evol 3:743–756. doi:10.1111/j.2041-210X.2012.00196.x

39. Bivand R, Pebesma E, Gómez Rubio V. 2013. Applied Spatial Data Analysis with R, 2nd ed. New York: Springer.

40. Cummins M, Watson C, Edwards RJ, Mattick JS. 2024. The evolution of ultraconserved elements in vertebrates. Mol Biol Evol 41:1724–1734. doi:10.1093/molbev/msae146

41. Dorrity MW, Saunders LM, Queitsch C, Fields S, Trapnell C. 2020. Dimensionality reduction by UMAP to visualize physical and genetic interactions. Nat Commun 11:1537. doi:10.1038/s41467-020-15351-4 doaj.org+9nature.com+9pubmed.ncbi.nlm.nih.gov+9

42. Liu D, Hunt M, Tsai IJ. 2018. Inferring synteny between genome assemblies: a systematic evaluation. BMC Bioinformatics 19:26. doi:10.1186/s12859-018-2008-1

43. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Paten B, et al. 2024. Building pangenome graphs. Nat Methods 21:1–5. doi:10.1038/s41592-024-01450-1

44. Dubois F, Sidiropoulos N, Weischenfeldt J, Beroukhim R. 2022. Structural variations in cancer and the 3D genome. Nat Rev Cancer 22:533–546. doi:10.1038/s41571-022-00698-0

45. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO Update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol 38:4647–4654. doi:10.1093/molbev/msab199

46. Oluwadare O, Highsmith M, Cheng J. 2019. An overview of methods for reconstructing 3 D chromosome and genome structures from Hi C data. Biol Proced Online 21:7. doi:10.1186/s12575-019-0095-0

47. Lappala A, DeMaere MZ, Dansereau C, Villani TS, Lieberman Aiden E, Dekker J, et al. 2021. Four dimensional chromosome reconstruction elucidates the spatiotemporal reorganization of the mammalian X chromosome. Proc Natl Acad Sci U S A 118:e2107092118. doi:10.1073/pnas.2107092118

48. Percival Alwyn L, Smith J, Patel K, Ahmed A, Zhang Y, Vickers M, et al. 2024. UKCropDiversity HPC: A collaborative high performance computing resource approach for sustainable agriculture and biodiversity conservation. Plants People Planet. doi:10.1002/ppp3.10607

49. Sandelin A, Klasson L, Medstrand P, Engström U, Fredman D, Lenhard B. 2004. Arrays of ultraconserved non coding regions span the loci of key developmental genes in vertebrate genomes. BMC Genomics 5:99. doi:10.1186/1471-2164-5-99

50. Leclair NK, Anczuków O. 2022. ‘Poisoning’ of the transcriptome by ultraconserved elements. Nat Rev Mol Cell Biol 23:777. doi:10.1038/s41580-022-00527-1 europepmc.org+2pubmed.ncbi.nlm.nih.gov+2researchgate.net+2

51. Jungers WL, Falsetti AB, Wall CE. 1990. Shape, relative size, and size-adjustments in morphometrics. J Hum Evol 19:333–348. doi:10.1016/0047-2484(90)90043-W

52. Wickham H. 2007. Reshaping data with the reshape package. J Stat Softw 21:1–20. Available from: http://www.jstatsoft.org/v21/i12/

53. Müller P, Eddelbuettel D, Bates D. 2025. matrixStats: Functions for Highly Optimized Computing of Matrix Statistics. Available from: https://cran.r-project.org/web/packages/matrixStats/matrixStats.pdf

54. Wickham H, Vaughan D, Girlich M. 2024. tidyr: Tidy messy data. R package version 1.3.1. Available from: https://github.com/tidyverse/tidyr, https://tidyr.tidyverse.org

55. Wickham H, François R, Henry L, Müller K, Vaughan D. 2023. dplyr: A grammar of data manipulation. R package version 1.1.4. Available from: https://github.com/tidyverse/dplyr

56. Wickham H. 2016. ggplot2: Elegant graphics for data analysis. New York: Springer-Verlag. Available from: https://ggplot2.tidyverse.org

57. Danecek P, Bonfield JK, Liddle J, Marshall J, et al. 2021. Twelve years of SAMtools and BCFtools. GigaScience 10:giab008. doi:10.1093/gigascience/giab008

58. Wilke C. 2024. cowplot: Streamlined plot theme and plot annotations for 'ggplot2'. R package version. Available from: https://cran.r-project.org/package=cowplot

59. Foley NM, et al. 2023. A genomic timescale for placental mammal evolution. Science 380:1239–1245. doi:10.1126/science.abl8189

60. Martin FJ, et al. 2023. Ensembl 2023. Nucleic Acids Res 51:D933–D941. doi:10.1093/nar/gkac1040

61. Yue JX, Liti G. 2019. simuG: a general-purpose genome simulator. Bioinformatics 35:4442–4444. doi:10.1093/bioinformatics/btz366

62. Uyeda JC, Caetano DS, Pennell MW. 2015. Comparative analysis of principal components can be misleading. Syst Biol 64:677–689. doi:10.1093/sysbio/syv019

63. Kumar S, Stecher G, Suleski M, Hedges SB. 2022. TimeTree 5: an expanded resource for species divergence times. Mol Biol Evol 39:msac174. doi:10.1093/molbev/msac174

64. Blomberg SP, Garland T Jr, Ives AR. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57:717–745. doi:10.1111/j.0014-3820.2003.tb00285.x

65. Pagel M. 1999. Inferring the historical patterns of biological evolution. Nature 401:877–884. doi:10.1038/44766

66. Revell LJ. 2024. phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). PeerJ 12:e16505. doi:10.7717/peerj.16505

927 67. Canavez FC, Luche DD, Stothard P, Leite KR, Sousa-Canavez JM, Plastow G, et al.
928 2012. Genome sequence and assembly of *Bos indicus*. J Hered 103:342–348.
929 doi:10.1093/jhered/esr150

930 68. Rosen BD, Bickhart DM, Koren S, Schnabel RD, Hall R, Zimin A, et al. 2020. De
931 novo assembly of the cattle reference genome with single-molecule sequencing.
932 Gigascience 9:giaa021. doi:10.1093/gigascience/giaa021

933 69. Ananthasayanam S, Kothandaraman H, Nayee N, Saha S, Baghel DS,
934 Gopalakrishnan K, et al. 2019. First near complete haplotype phased genome assembly
935 of River buffalo (*Bubalus bubalis*). bioRxiv 618785. doi:10.1101/618785

936 70. Dong Y, Zhang X, Xie M, Arefnezhad B, Wang Z, Wang W, et al. 2015. Reference
937 genome of wild goat (*Capra aegagrus*) and sequencing of goat breeds provide insight
938 into genic basis of goat domestication. BMC Genomics 16:1–11. doi:10.1186/s12864-
939 015-1925-6

940 71. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, et al. 2017. Single-
941 molecule sequencing and chromatin conformation capture enable de novo reference
942 assembly of the domestic goat genome. Nat Genet 49:643–650. doi:10.1038/ng.3802

943 72. Farré M, Li Q, Darolti I, Zhou Y, Damas J, Proskuryakova AA, et al. 2019. An
944 integrated chromosome-scale genome assembly of the Masai giraffe (*Giraffa*
945 *camelopardalis tippelskirchi*). GigaScience 8:giz090. doi:10.1093/gigascience/giz090

946 73. Davenport KM, Bickhart DM, Worley K, Murali SC, Salavati M, Clark EL, et al.
947 2022. An improved ovine reference genome assembly to facilitate in-depth functional

948 annotation of the sheep genome. Gigascience 11:giab096.
949 doi:10.1093/gigascience/giab096

950 74. Bredemeyer KR, Harris AJ, Li G, Zhao L, Foley NM, Roelke-Parker M, et al. 2021.
951 Ultracontinuous single haplotype genome assemblies for the domestic cat (*Felis catus*)
952 and Asian leopard cat (*Prionailurus bengalensis*). J Hered 112:165–173.
953 doi:10.1093/jhered/esaa047

954 75. Mohr DW, Naguib A, Weisenfeld NI, Kumar V, Shah P, Church DM, et al. 2017.
955 Improved de novo genome assembly: linked-read sequencing combined with optical
956 mapping produce a high-quality mammalian genome at relatively low cost. bioRxiv
957 128348. doi:10.1101/128348

958 76. Bredemeyer KR, Murphy WJ. 2021. Unpublished Liger assemblies. Veterinary
959 Integrative Biosciences, Texas A&M University, College Station, TX. Submitted 11 Jan
960 2021.

961 77. Wang G, Brändl B, Rohrandt C, Hong K, Pang A, Lee J, et al. 2021. Chromosome-
962 level genome assembly of the functionally extinct northern white rhinoceros
963 (*Ceratotherium simum cottoni*). bioRxiv. doi:10.1101/2021.12.01.470802

964 78. Wang C, Li H, Guo Y, Huang J, Sun Y, Min J, et al. 2020. Donkey genomes provide
965 new insights into domestication and selection for coat color. Nat Commun 11:6014.
966 doi:10.1038/s41467-020-19702-6

967 79. Kalbfleisch TS, Rice ES, DePriest MS Jr, Walenz BP, Hestand MS, Vermeesch JR,
968 et al. 2018. Improved reference genome for the domestic horse increases assembly
969 contiguity and composition. Commun Biol 1:197. doi:10.1038/s42003-018-0209-z

970 80. Gordon D, Murali SC, Ventura M, Mercuri L, Eichler EE. De novo PacBio assembly
971 of a female gorilla (Kamilah). Genome Sciences, University of Washington, Seattle, WA,
972 USA. Unpublished.

973 81. Jayakumar V, Nishimura O, Kadota M, Hirose N, Sano H, Murakawa Y, et al. 2021.
974 Chromosomal-scale de novo genome assemblies of cynomolgus macaque and
975 common marmoset. Sci Data 8:159. doi:10.1038/s41597-021-00911-0

976 82. Dutcher S, Fulton R, Lindsay T. The rhesus macaque (*Macaca mulatta*) genome
977 assembly. The McDonnell Genome Institute, Washington University School of Medicine;
978 University of Washington, Seattle, WA. Unpublished.

979 83. Kronenberg ZN, Fiddes IT, Gordon D, Murali S, Cantsilieris S, Meyerson OS, et al.
980 2018. High-resolution comparative analysis of great ape genomes. Science
981 360:eaar6343. doi:10.1126/science.aar6343

982 84. Batra SS, Levy-Sakin M, Robinson J, Guillory J, Durinck S, Vilgalys TP, et al. 2020.
983 Accurate assembly of the olive baboon (*Papio anubis*) genome using long-read and Hi-C
984 data. Gigascience 9:giaa134. doi:10.1093/gigascience/giaa134

985 85. Rupp O, MacDonald ML, Li S, Dhiman H, Polson S, Griep S, et al. 2018. A
986 reference genome of the Chinese hamster based on a hybrid assembly strategy.
987 Biotechnol Bioeng 115:2087–2100. doi:10.1002/bit.26809

988 86. Roller M, Navarro FCP, Fiddes I, Streeter I, Feig C, Martin-Galvez D, et al. Repeat
989 associated mechanisms of genome evolution and function revealed by the *Mus caroli*
990 and *Mus pahari* genomes. Available from: http://bura.brunel.ac.uk/handle/2438/16068

87. Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, et al. 2009. Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol 7:e1000112. doi:10.1371/journal.pbio.1000112

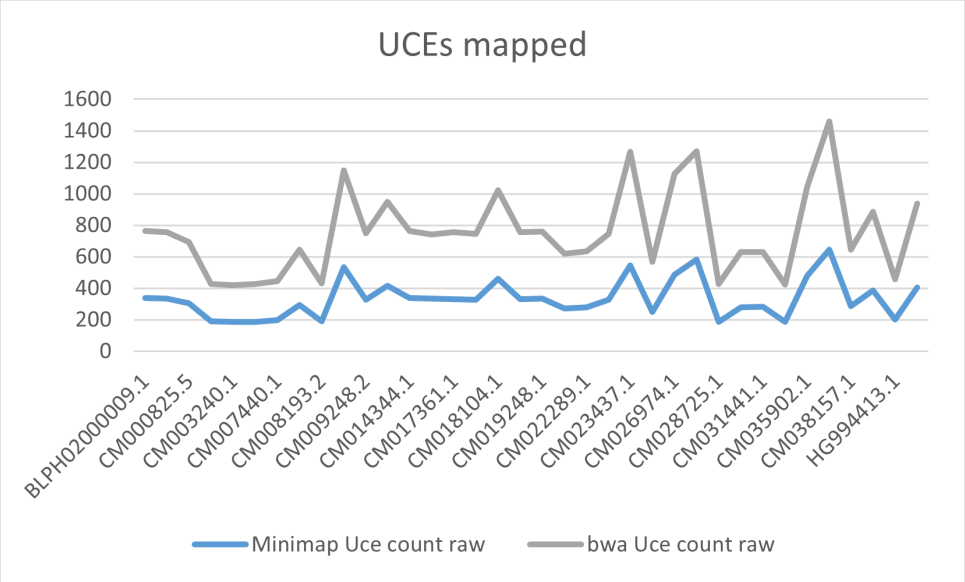
88. Lilue J, Doran AG, Fiddes IT, Abrudan M, Armstrong J, Bennett R, et al. 2018. Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. Nat Genet 50:1574–1583. doi:10.1038/s41588-018-0223-8

89. Lassance JM, Hoekstra HE. Improved assembly of the deer mouse (Peromyscus maniculatus) genome. Organismic and Evolutionary Biology / Molecular and Cellular Biology, Harvard University / Howard Hughes Medical Institute, Cambridge, MA, USA. Unpublished.

90. Howe K, Dwinell M, Shimoyama M, Corton C, Betteridge E, Dove A, et al. 2021. The genome sequence of the Norway rat, *Rattus norvegicus* Berkenhout 1769. Wellcome Open Res 6:60. doi:10.12688/wellcomeopenres.17242.1

91. Sievert C. 2020. Interactive web-based data visualization with R, plotly, and shiny. Chapman and Hall/CRC. ISBN 9781138331457. Available from: https://plotly-r.com

1012 **Supplementary Information**



1013

1014 **Figure S1.** Number of ultraconserved elements (UCEs) mapped utilizing two different mapping
1015 software: minimap2 (blue) and bwa-mem (grey). Demonstrated using a subset of chromosome
1016 assemblies picked from the 'chromosome set 2' of the placental mammals UCE dataset discussed in
1017 the main text.

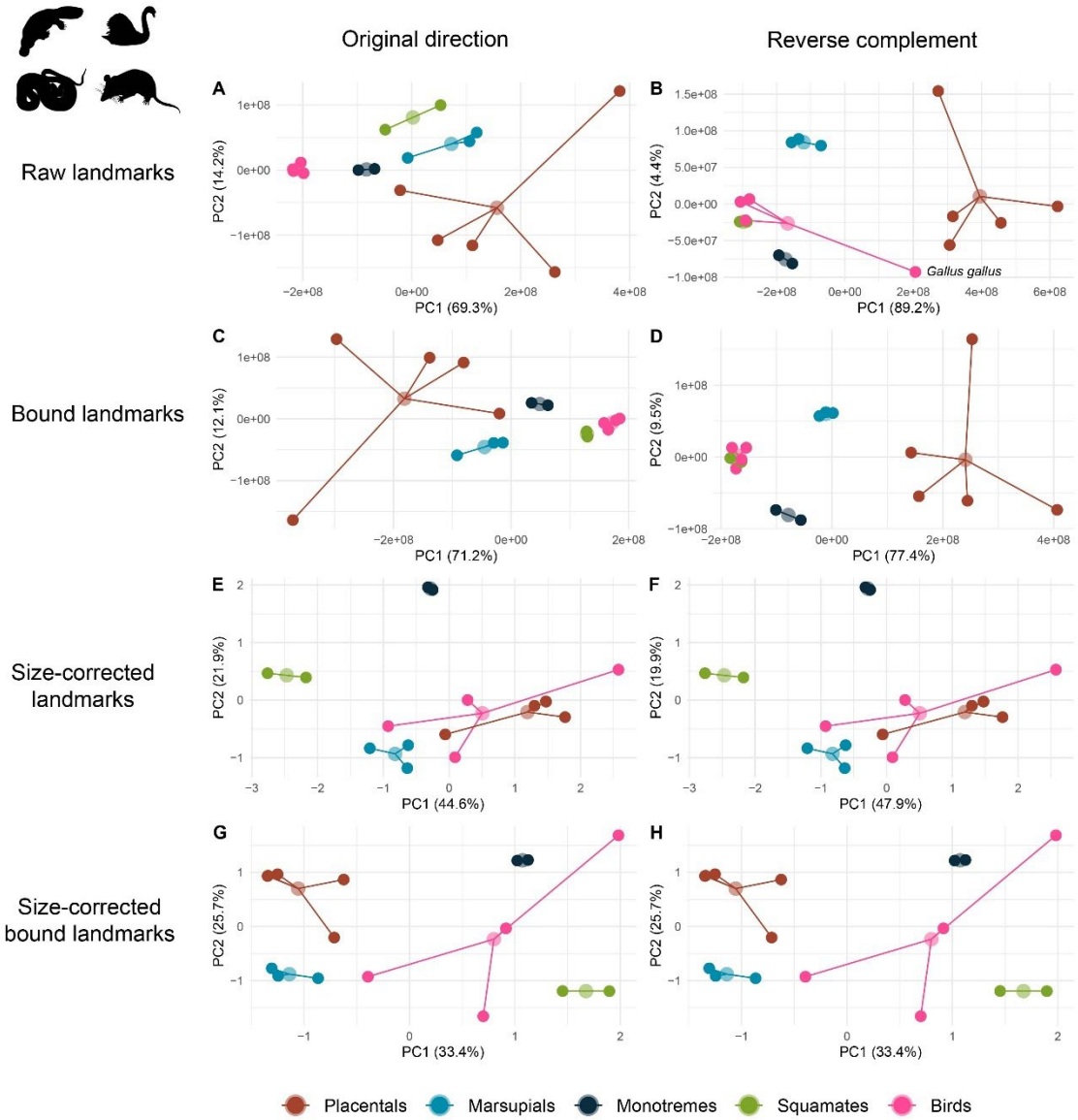
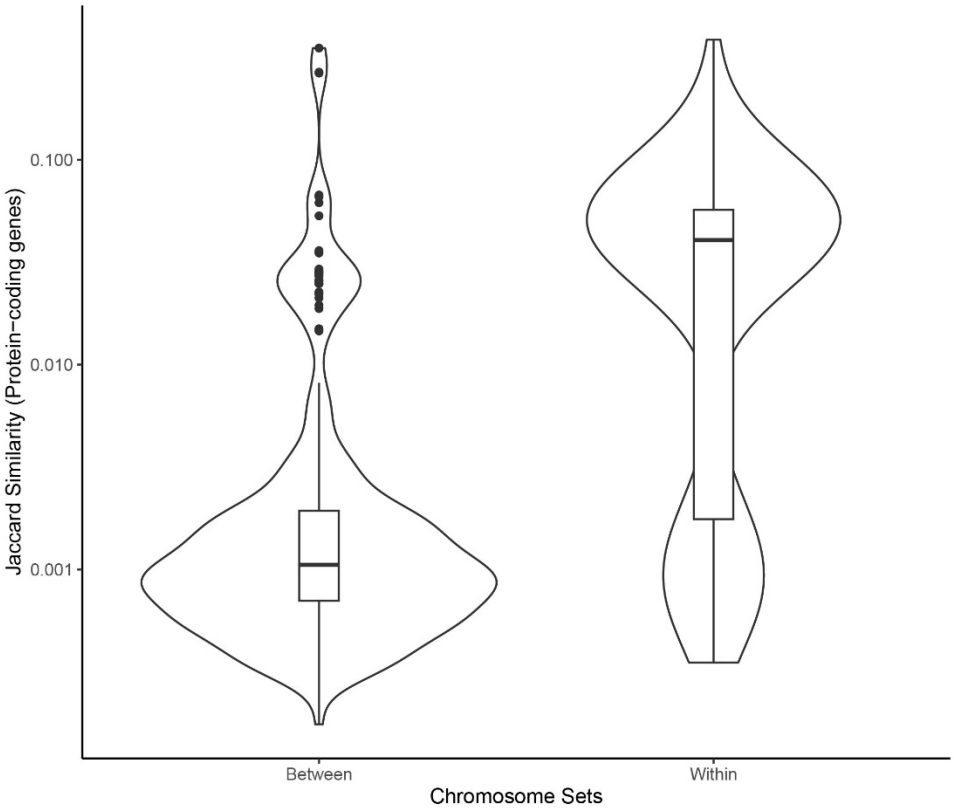


Figure S2. Principal component (PC) sensitivity analyses of 53 X-linked BUSCO landmarks in amniote vertebrates taken from Lovell et al. (2022). The first row are PCAs conducted using raw, absolute positions for the original direction (A) and reverse complement (B), respectively. The second row are PCAs conducted using bounded, absolute positions for the original direction (C) and the reverse complement (D). The third row are PCAs conducted using size-correction with chromosome size for the original direction (E) and the reverse complement (F). The fourth row of plots are PCAs conducted using size-correction using bounded chromosome sizes and positions for the original direction (G) and reverse complement (H). The chicken (*Gallus gallus*) is highlighted in plot B to illustrate an analytical artifact associated with clustered landmarks on large chromosomes. Organism silhouettes downloaded from https://www.phylopic.org/.



1030
1031 Figure S3. Jaccard similarity coefficients for within and between protein-coding gene content in the
1032 two focal chromosome sets used in the study.

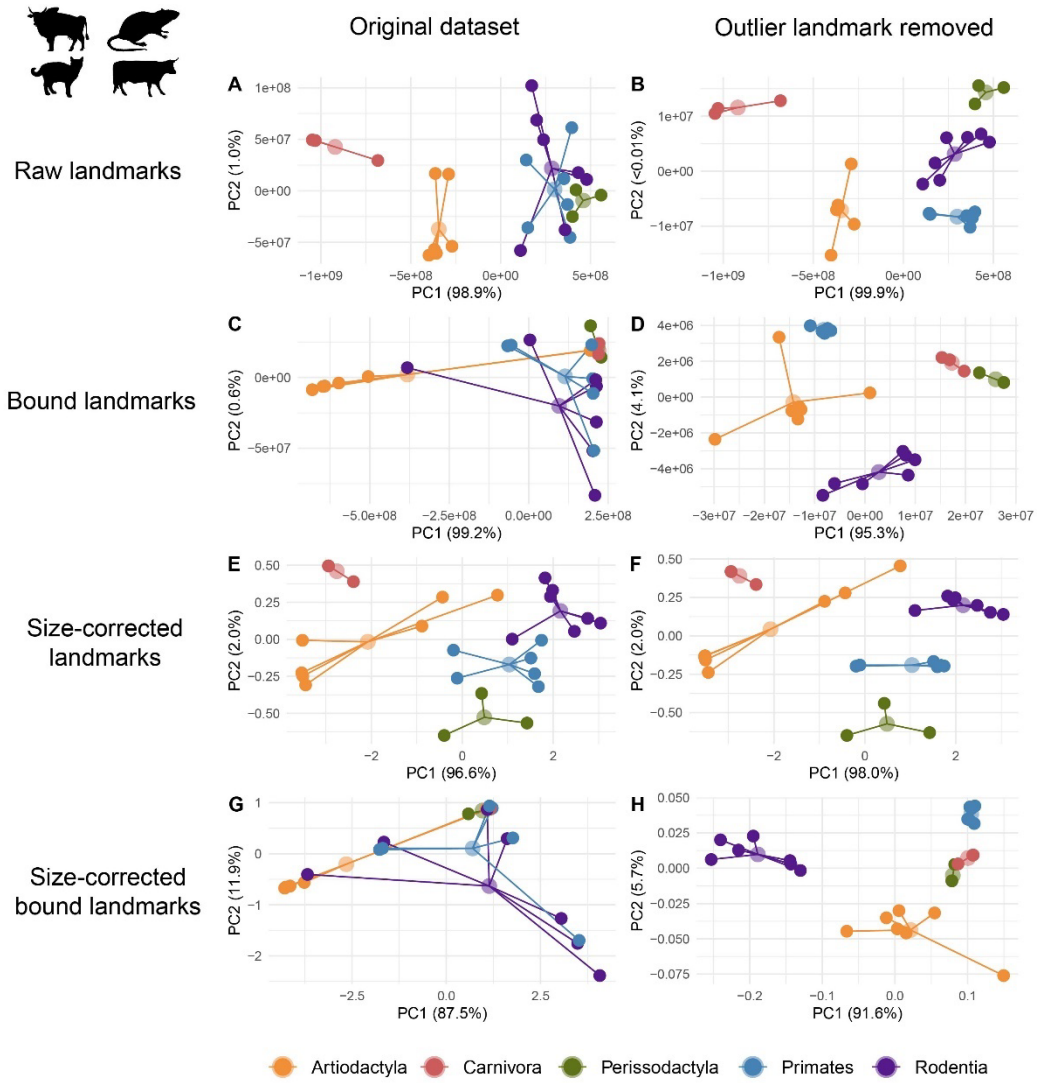


Figure S4. Principal component analyses (PCAs) of ultraconserved element (UCE) landmarks in placental mammals for chromosome set 1 before and after the removal of an outlier landmark (UCE 6948). The first row of plots are PCAs conducted using raw, absolute positions for the original direction (A) and reverse complement (B), respectively. The second row of plots are PCAs conducted using bounded, absolute positions for the original direction (C) and the reverse complement (D). The third row of plots are PCAs conducted using size-correction with chromosome size for the original direction (E) and the reverse complement (F). The fourth row of plots are PCAs conducted using size-correction using bounded chromosome sizes and positions for the original direction (G) and reverse complement (H). The first column of plots are PCAs conducted with all 186 UCE landmarks and the second column of plots are PCAs based on 185 UCE landmarks (outlier removed).

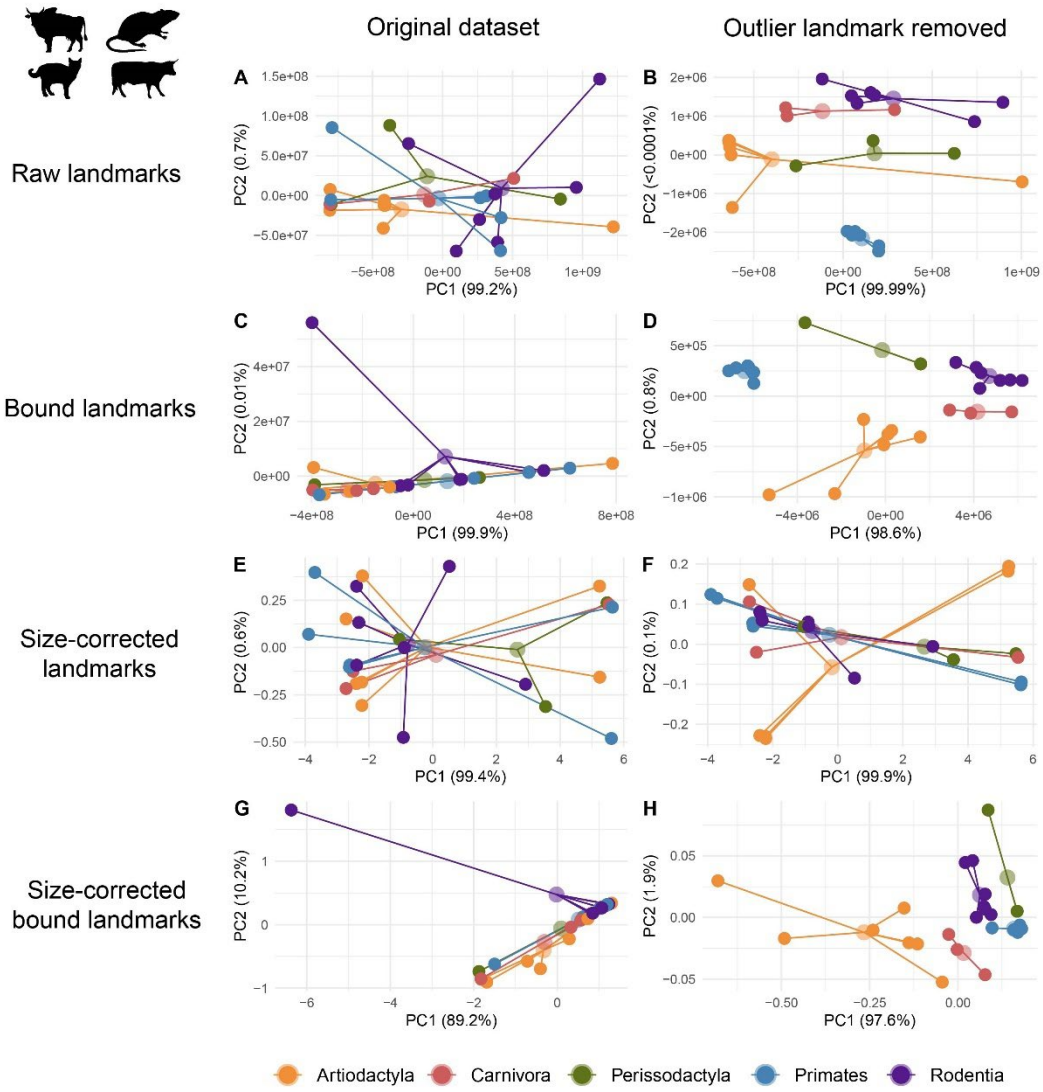
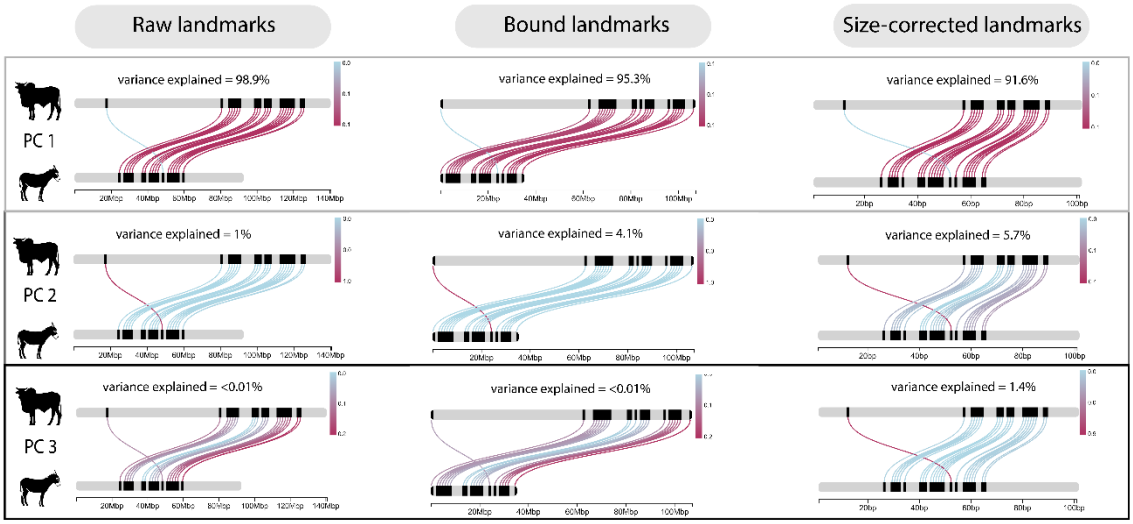


Figure S5. Principal component analyses (PCAs) of ultraconserved element (UCE) landmarks in placental mammals for chromosome set 2 before and after the removal of an outlier landmark (UCE 6948). The first row of plots are PCAs conducted using raw, absolute positions for the original direction (A) and reverse complement (B), respectively. The second row of plots are PCAs conducted using bounded, absolute positions for the original direction (C) and the reverse complement (D). The third row of plots are PCAs conducted using size-correction with chromosome size for the original direction (E) and the reverse complement (F). The fourth row of plots are PCAs conducted using size-correction using bounded chromosome sizes and positions for the original direction (G) and reverse

complement (H). The first column of plots are PCAs conducted with all 82 UCE landmarks and the second column of plots are PCAs based on 81 UCE landmarks (outlier removed).



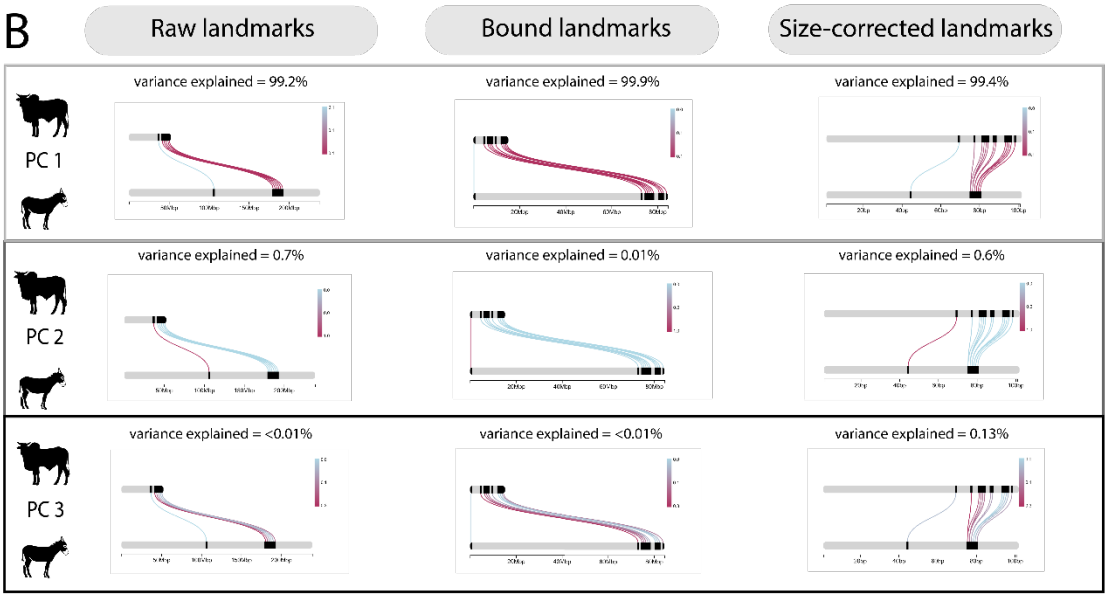


Figure S6. ChromoMap plots visualizing principal component analysis (PCA) loadings of landmarks on PC1, PC2, and PC3 in the raw landmark PCA, bounded landmark PCA, and size corrected landmark PCA of chromosome set 1 (A) and set 2 (B) from the placental mammal visualized using links between *Bos indicus* and *Equus asinus*. Respective variances explained by each principal component are listed in the plot. A single landmark (UCE 6948) was found to be uniquely weighed on the various PCs identifying it as an outlier landmark.

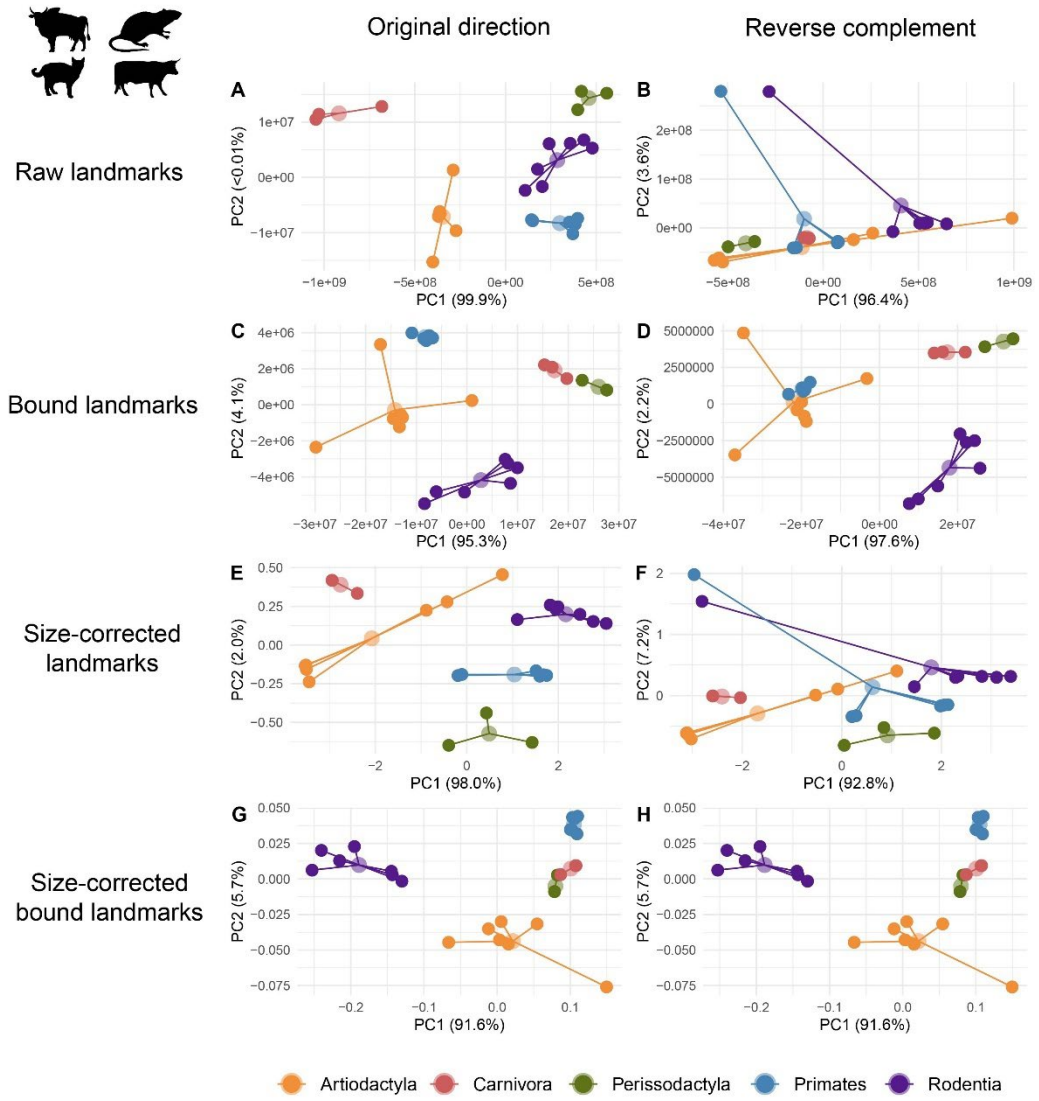


Figure S7. Principal component (PC) sensitivity analyses of outlier removed UCE landmarks in placental mammals chromosome set 1. The first row are PCAs conducted using raw, absolute positions for the original direction (A) and reverse complement (B), respectively. The second row are PCAs conducted using bounded, absolute positions for the original direction (C) and the reverse complement (D). The third row are PCAs conducted using size-correction with chromosome size for the original direction (E) and the reverse complement (F). The fourth row are PCAs conducted using size-correction using bounded chromosome sizes and positions for the original direction (G) and reverse complement (H).

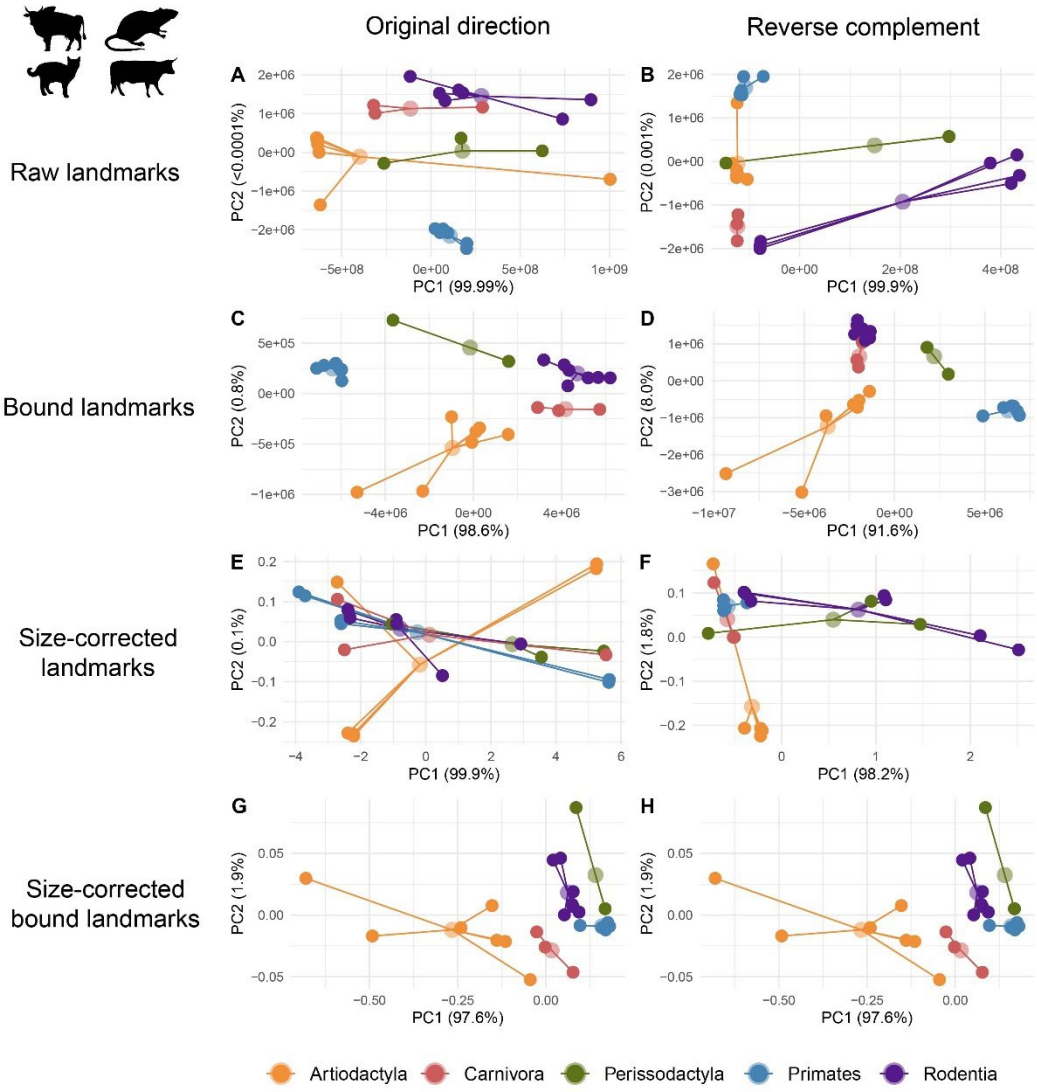


Figure S8. Principal component (PC) sensitivity analyses of outlier removed UCE landmarks in placental mammals chromosome set 2. The first row are PCAs conducted using raw, absolute positions for the original direction (A) and reverse complement (B), respectively. The second row are PCAs conducted using bounded, absolute positions for the original direction (C) and the reverse complement (D). The third row are PCAs conducted using size-correction with chromosome size for the original direction (E) and the reverse complement (F). The fourth row are PCAs conducted using size-correction using bounded chromosome sizes and positions for the original direction (G) and reverse complement (H).

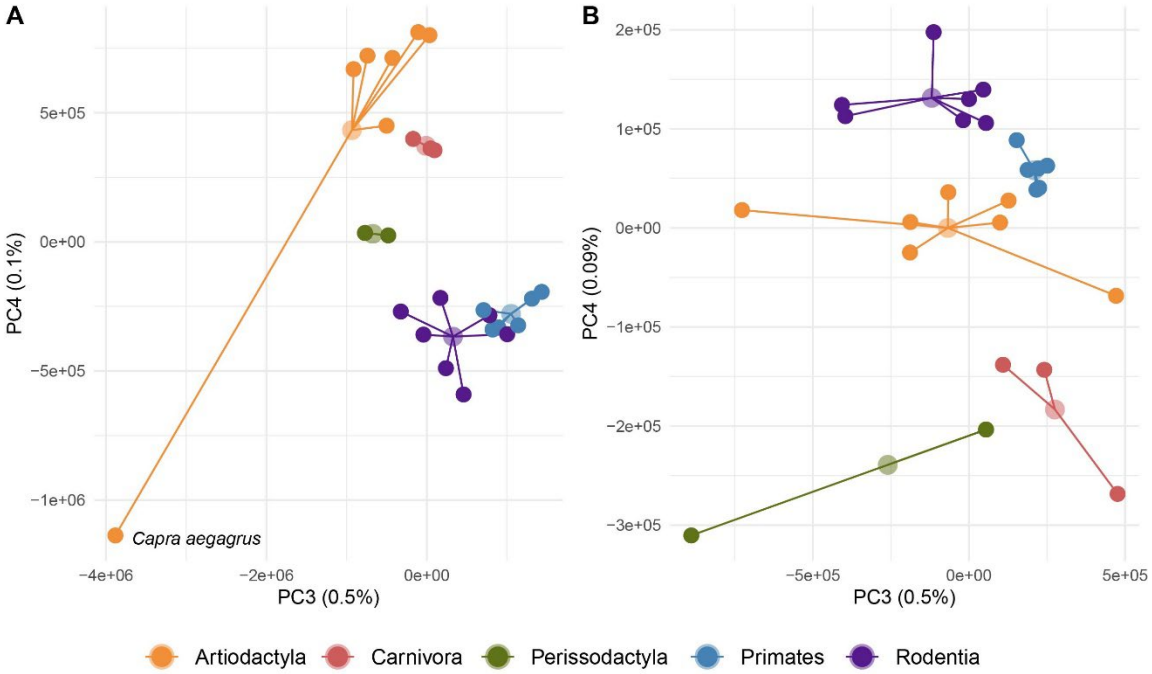


Figure S9. Ordination results from principal component (PC) analysis of UCE landmark positions in two sets of chromosomes from 26 species of placental mammals using bounded raw positions after the removal of an outlier landmark (see text). Representatives for five orders (Artiodactyla, Carnivora, Perissodactyla, Primates and Rodents) were included (Table1). The depicted structural variation is inferred from PCs 3 and 4 of 185 (chromosome set 1, A) and 81 (chromosome set 2, B) landmarks, respectively. In both datasets, the first two PCs explained > 99% of the variation (see **Figure 5**, main text). While explaining < 1% of the overall variation, PCs 3 and 4 mostly separated mammalian orders in geno-metric space. *Carpa aegragus* was a notable outlier in chromosome set 1.

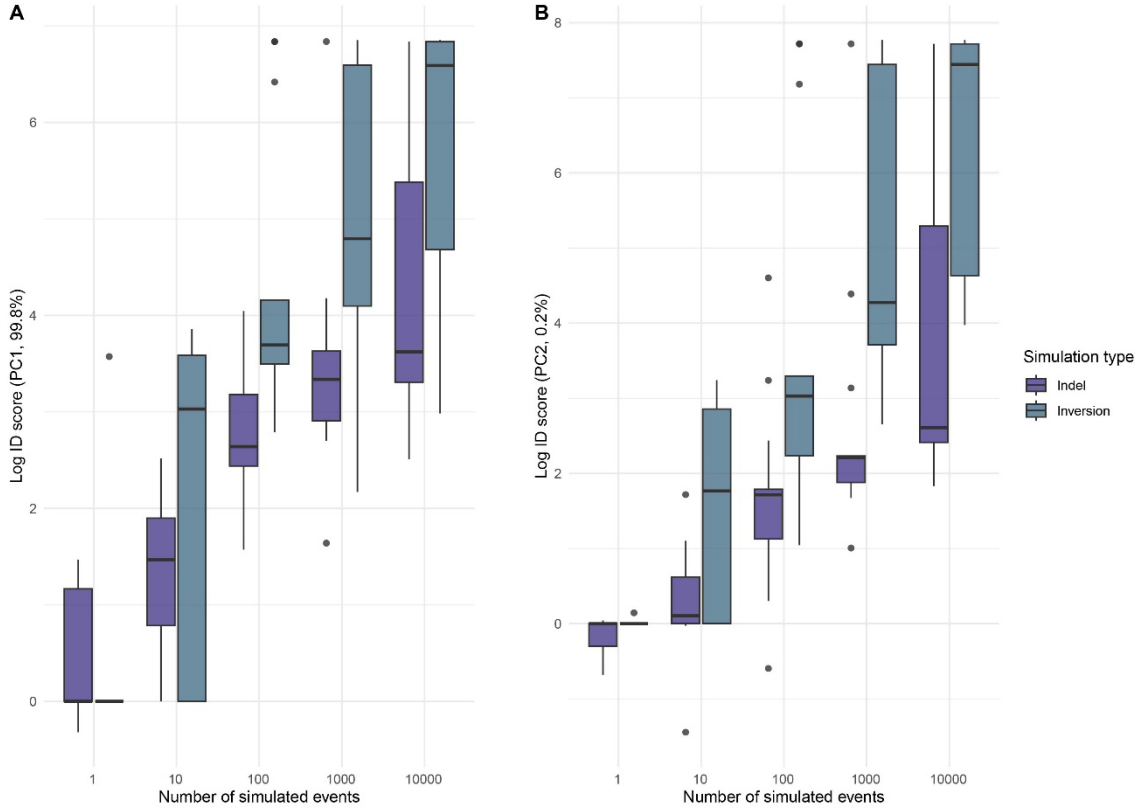


Figure S10. Boxplots comparing increase in disparity (ID) between indel and inversion simulations for principal component (PC)1 (A) and PC2 (B), respectively. Note that inversions cause significantly greater ID than indels.

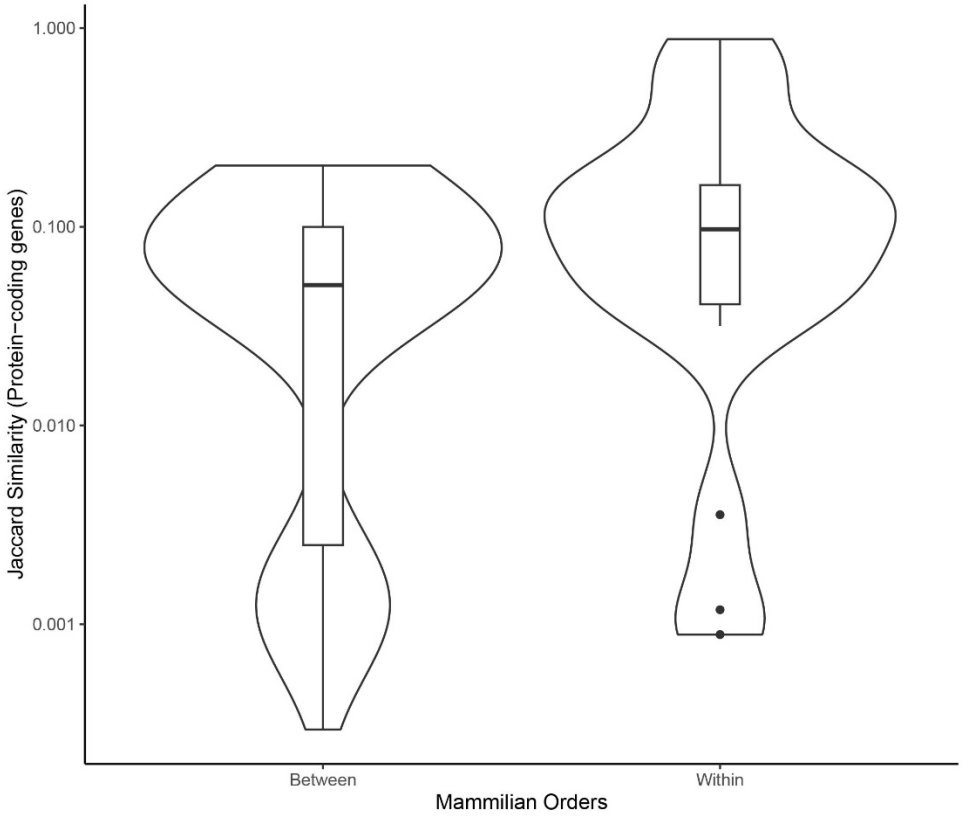


Figure S11. Jaccard similarity coefficients for within and between protein-coding gene content in mammalian orders used in the study.

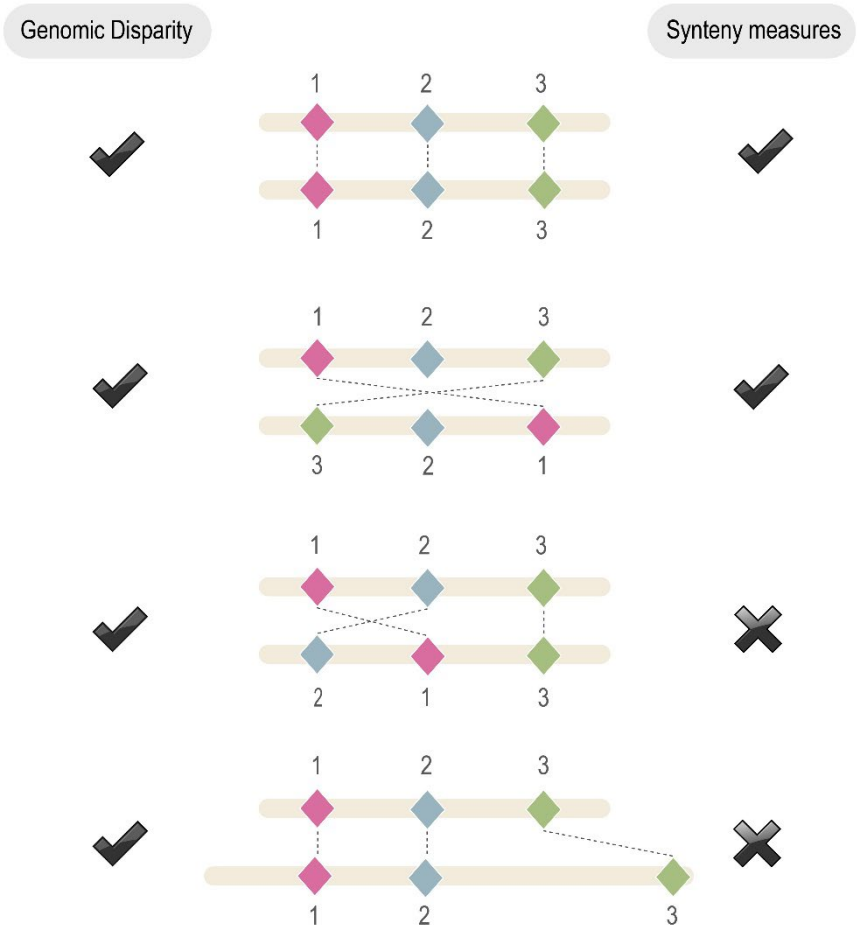


Figure S12. Schematic diagram illustrating sensitivities of disparity analysis and synteny analysis to different landmark scenarios. Scenarios are described from top to bottom of the figure. First, both analytical approaches will report highly conserved landmark/ortholog order, indicating when two species have highly conserved genomic architecture. Second, both approaches will report large inversions among landmark positions. Third, depending on parameter settings syntenic analysis may ignore small landmark rearrangements whereas disparity analysis captures this variation. Finally, synteny analysis ignores spatial variation in landmark placement and ‘synteny’ can be broken if chromosomal gaps are large enough whereas disparity analysis captures this variation in landmark placement.

Appendix S1. List of 53 BUSCO genes identified as landmarks for the human X-linked study of amniote vertebrates.

eda, dlg3, snx12, gjb1, zmym3, nono, shroom4, amot, lrch2, pls3, plp1, gla, btk, drp2, sytl4, srpx2, tspan6, dach2, apool, rps6ka6, lpar4, magt1, fgf16, zdhhc15, uprt, nsdhl, vma21, cd99l2, mtmr1, mtm1, aff2, fmr1, slitrk4, atp11c, cd40lg, vgll1, htatsf1, gpc3, rap2c, frmd7, bcorl1, utp14a, smarca1, stag2, cul4b, lamp2, tmem255a,