nature ecology & evolution

Article

https://doi.org/10.1038/s41559-022-01974-x

Eighty million years of rapid evolution of the primate Y chromosome

Received: 1 July 2022

Accepted: 15 December 2022

Published online: 2 June 2023

Check for updates

Yang Zhou[1], Xiaoyu Zhan[1], Jiazheng Jin[1], Long Zhou[2,3], Juraj Bergman[4,5], Xuemei Li[1,6], Marjolaine Marie C. Rousselle[5], Meritxell Riera Belles[5], Lan Zhao[7], Miaoquan Fang[1], Jiawei Chen[1], Qi Fang[1], Lukas Kuderna[8], Tomas Marques-Bonet[8,9,10,11], Haruka Kitayama[12], Takashi Hayakawa[13,14], Yong-Gang Yao[15,16,17,18,19], Huanming Yang[1,6,20,21,22], David N. Cooper[23], Xiaoguang Qi[7], Dong-Dong Wu[17,18,19,24], Mikkel Heide Schierup[5] & Guojie Zhang[2,3,24,25]

The Y chromosome usually plays a critical role in determining male sex and comprises sequence classes that have experienced unique evolutionary trajectories. Here we generated 19 new primate sex chromosome assemblies, analysed them with 10 existing assemblies and report rapid evolution of the Y chromosome across primates. The pseudoautosomal boundary has shifted at least six times during primate evolution, leading to the formation of a Simiiformes-specific evolutionary stratum and to the independent start of young strata in Catarrhini and Platyrrhini. Different primate lineages experienced different rates of gene loss and structural and chromatin change on their Y chromosomes. Selection on several Y-linked genes has contributed to the evolution of male developmental traits across the primates. Additionally, lineage-specific expansions of ampliconic regions have further increased the diversification of the structure and gene composition of the Y chromosome. Overall, our comprehensive analysis has broadened our knowledge of the evolution of the primate Y chromosome.

Sex chromosomes usually carry the master genes that determine sex and play prominent roles in many evolutionary processes, including genomic conflict, adaptation and speciation[1–3]. This is reflected in a very different genomic evolution of X and Y chromosomes as compared with autosomes[4,5]. The sex chromosomes of therian mammals share a common origin from a pair of autosomes around 180 million years ago (MYA)[6,7] and have undergone subsequent episodes of recombination suppression, leading to massive structural and sequence divergence between the X and Y (refs. [8,9]). The Y chromosome has been subject to unique selective forces. It contains the pseudoautosomal region (PAR), which is shared with the X chromosome and a non-recombining male-specific region, which has experienced extensive gene loss compared with other genomic regions while also accumulating repetitive sequences[3,8]. All eutherian sex chromosomes share three 'evolutionary strata' corresponding to cessation of recombination at different evolutionary timepoints[9,10]. Recent sequence analysis of the human and other mammalian Y chromosomes has suggested that some primate species, including humans, contain two additional strata that appeared during primate evolution[7,10,11]. These studies also revealed a preponderance of large palindromes on the Y chromosome that homogenize through the continuous action of intrachromosomal gene conversion[8,12–14], as well as a high turnover of genes in multiple, very similar copies, termed ampliconic genes (AGs). However, owing to the highly repetitive nature of the Y chromosome and the inherent difficulty in assembling it[15], high-quality Y chromosomes have until now been published only for ten primate species, with considerable phylogenetic bias towards the great apes[8,13,16–19].

In this Article, to explore primate Y chromosome evolution more broadly, we analysed Y chromosomes across a more diverse set of

A full list of affiliations appears at the end of the paper. e-mail: guojiezhang@zju.edu.cn

primate lineages. We generated the X-linked and Y-linked sequence assemblies for 19 previously uncharacterized primate species using long-read sequencing, thereby increasing the number of available Y chromosomes from 10 to 29 primate species. Our new dataset includes species from all the major primate lineages (Fig. 1 and Supplementary Data 1), including 2 prosimian species from the Lorisidae and Daubentoniidae, 5 New World monkeys (NWMs) from the Atelidae, Callitrichidae and Cebidae, 9 apes from the Hylobatidae and Hominidae, and 13 Old World monkeys (OWMs), that is, Cercopithecidae, covering both subfamilies Cercopithecinae and Colobinae. The divergence times of these species range from 1.9 million years (Myr) between species from the same genus (for example, between *Papio hamadryas* and *Papio anubis*), to 81.8 Myr between the prosimians and simians[20]. With this unique dataset, we were able to identify the conserved and lineage-specific patterns of change in different regions of the Y chromosome during primate diversification.

Results

Structural variation of Y chromosomes across the primates

The sex chromosome dataset is based on the male individuals of a larger set of de novo assembled genomes (Supplementary Data 1)[20]. Briefly, long-read genome sequencing technologies, including PacBio and Nanopore, together with short-read sequencing, were used to create and polish primary genome assemblies[20]. In five species (*Colobus guereza*, *Hylobates pileatus*, *Nasalis larvatus*, *Nycticebus pygmaeus* and *Saguinus midas*), Hi-C technology was further used to scaffold the assembled contigs.

On the basis of the depth difference between male and female resequencing reads, we first identified the X- and Y-linked sequences from the long-read assembly (Fig. 1a and Methods). As the Y chromosome is usually enriched in large segmental duplications and repeats, assemblers are prone to collapse these sequences in the resulting assembly[15]. We addressed this by performing additional de-collapsing of the identified Y-linked sequences from coverage information to obtain more complete and accurate Y chromosome assemblies (Supplementary Notes and Supplementary Data 2). Finally, we also performed manual examination for sex-linked sequences in each species using Hi-C data, 10X-linked reads and long-read mappings (Methods). The X chromosome assemblies are all close to being complete (on average 97% coverage of the human X chromosome), but due to their rapid evolution, Y chromosome completeness is more difficult to evaluate. We used published male karyotype images for 25 species to estimate the lengths of their Y chromosomes (Methods). Of the 15 newly sequenced species where such karyotype images were available, the Y assemblies for 10 species were within 53.7–100% of these length estimates, suggesting that a major fraction of the euchromatin region of Y chromosomes has been assembled (Supplementary Fig. 1 and Supplementary Data 2).

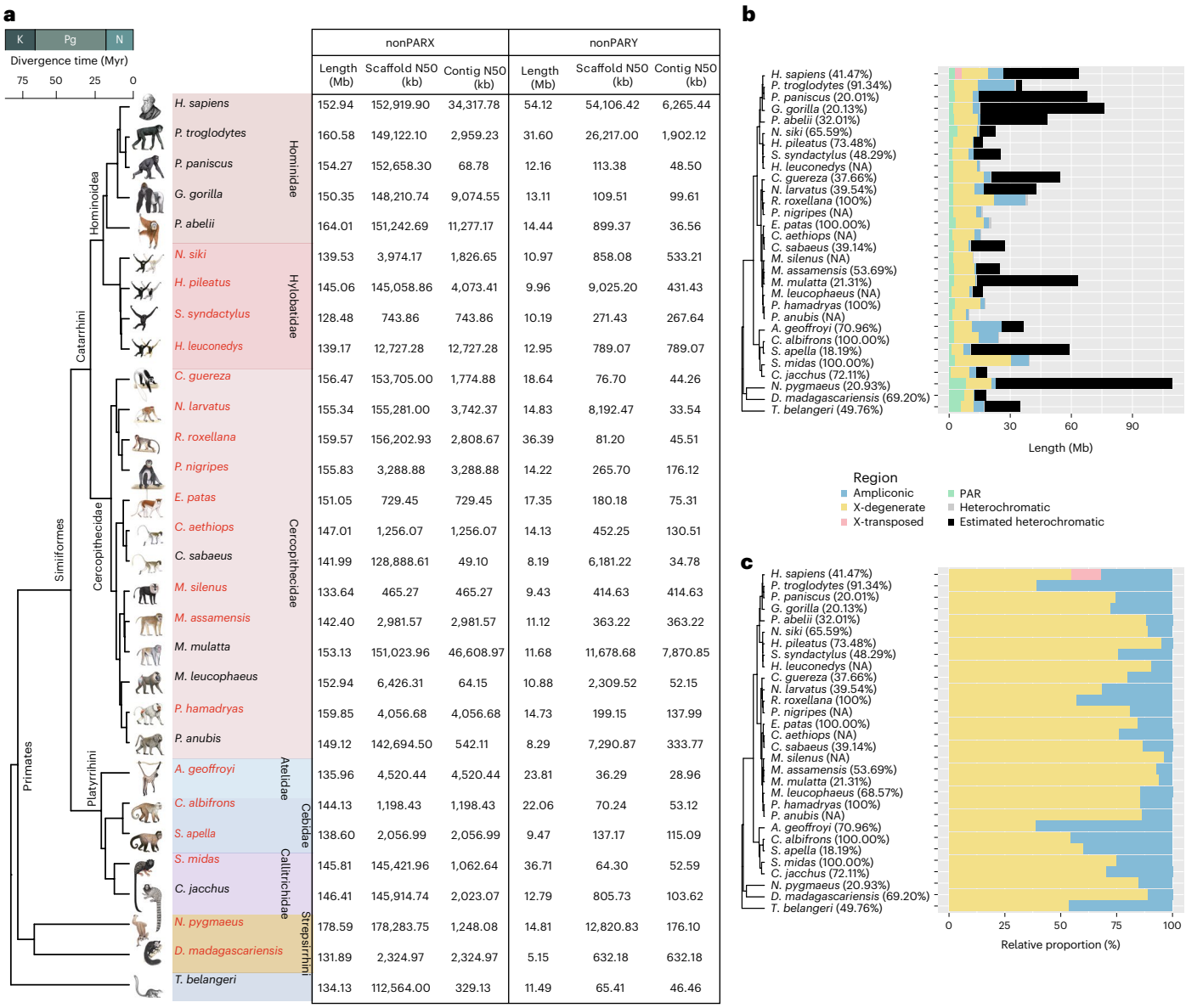
In contrast to the X chromosome, the Y chromosome length varies widely across mammalian species[16,21,22]. Here, by estimating the size of the Y chromosome based on the karyotype image data collected from a number of studies, we extend this observation to a broader range of primates. The sizes of the Y chromosomes from karyotype images varies by more than sixfold from 16.24 Mb (*Papio hamadryas*) to 109.66 Mb (*Nycticebus pygmaeus*) (mean 39.48 Mb, standard deviation 22.91, Supplementary Data 2). The human Y chromosome consists of four major male-specific sequence classes, viz. the X-degenerate, X-transposed, ampliconic and heterochromatic regions[8]. Of these, the X-degenerate region mainly contains the single-copy genes derived from the proto sex chromosomes and which has undergone X/Y divergence due to recombination suppression. The X-transposed region is highly identical to its X counterpart and is thought to have originated via an X-to-Y transposition after the human–chimpanzee split[8]. The ampliconic regions feature copies of highly identical (up to 99.9% identity) sequences, which are often arranged in a palindromic structure. The fourth class, the heterochromatic region, contains sequences that are compacted during cell division and highly enriched in satellite sequences[23]. Previous studies with few primate species showed that the length and composition of Y chromosomes varied across species[16,17]. Here we provided a broader view with data from more lineages and found that the size variation among the primate Y chromosome is mainly due to variation in the sizes of the heterochromatic and ampliconic regions (standard deviation 23.23 and 4.91 in the heterochromatic and ampliconic regions, respectively, compared with standard deviation 4.63 in the X-degenerate region; Fig. 1b,c and Supplementary Data 3). As has been previously reported in studies that focused on the great apes and other mammals[16,21,22], the length of the X-degenerate region is broadly similar across lineages. Of note, two NWMs, *Saguinus midas* and *Cebus albifrons*, with complete Y chromosome assemblies of similar size as estimated from karyotype, have short heterochromatic regions but very large ampliconic regions accounting for 25.00% and 45.24% of their Y chromosomes, respectively. Indeed, NWMs tend to have larger ampliconic regions compared with OWMs and apes (Supplementary Fig. 2). We confirm that the large X-transposed region is unique to human[8]. However, a smaller number of different X-transposed genes (Supplementary Fig. 3) are found in seven diverse species across the primate phylogeny. The wide phylogenetic distribution and the uniqueness of these genes suggest that X transposition has occurred independently several times during primate evolutions ('Evolutionary strata of primate sex chromosomes').

PAB shifts during the primate sex chromosome evolution

It is well established that recombination suppression between the sex chromosomes gradually evolves over time, leading to a reduction in PAR size and the emergence of new evolutionary strata[7,10,24]. The localization of the PAR boundary (PAB) is therefore crucial for elucidating strata evolution in primates. On the basis of the difference in sequencing coverage between male and female individuals (Supplementary Figs. 4–7), we defined the PAB in each species and found large variation in PAR length, particularly between the Strepsirrhini and Simiiformes. The Strepsirrhini exhibit the longest PAR of 8.28 Mb (Fig. 2 and Supplementary Data 3), while the common marmoset (*Callithrix jacchus*) exhibits the shortest PAR at 0.98 Mb, covering only 0.67% of its X chromosome[25]. The PAB is generally conserved among the Catarrhini (located in the third intron of the *XG* gene), and the boundary position varies only slightly across the Platyrrhini.

We further inferred, by parsimony, the PAB in the ancestral nodes of the primate phylogeny and the minimum number of independent subsequent reductions (Fig. 2). Compared with the outgroup species, treeshrew (*Tupaia belangeri*)[26,27], recombination suppression events have shortened the PAR in the most recent common ancestor (MRCA) of primates leading to two PAR genes in the treeshrew being located in the sex-differentiated region (SDR) of primates. The PAR was further reduced in the Simiiformes, with the relocation of the PAB between *PRKX* (PAR gene) and *NLGN4X* (SDR gene), thus generating a new evolutionary stratum spanning at least nine genes that is shared across the Simiiformes. The PAR has continually reduced in size during the diversification of the Simiiformes, resulting in the independent emergence of new evolutionary strata in several lineages. The reduction of the PAR is most striking in the common marmoset where a large segment of the ancestral PAR sequence on the Y chromosome became translocated into the SDR of this species[25]. Sexually antagonistic selection is one proposed evolutionary force that favours recombination suppression, that is, the reduction of PAR size[28]. Previous small-scale studies suggested that the different level of sex antagonism may be associated with the variation of PAR in primates[29]. However, we found no correlation between PAR length and primate life history traits, such as male-to-female body mass ratio (Supplementary Fig. 8). Indeed, we find no evidence supporting the conclusion that differences in sexual dimorphism between these two clades contributed to the varied PAR length in primates.

https://doi.org/10.1038/s41559-022-01974-x



	nonPARX			nonPARY		
	Length (Mb)	Scaffold N50 (kb)	Contig N50 (kb)	Length (Mb)	Scaffold N50 (kb)	Contig N50 (kb)
H. sapiens	152.94	152,919.90	34,317.78	54.12	54,106.42	6,265.44
P. troglodytes	160.58	149,122.10	2,959.23	31.60	26,217.00	1,902.12
P. paniscus	154.27	152,658.30	68.78	12.16	113.38	48.50
G. gorilla	150.35	148,210.74	9,074.55	13.11	109.51	99.61
P. abelii	164.01	151,242.69	11,277.17	14.44	899.37	36.56
N. siki	139.53	3,974.17	1,826.65	10.97	858.08	533.21
H. pileatus	145.06	145,058.86	4,073.41	9.96	9,025.20	431.43
S. syndactylus	128.48	743.86	743.86	10.19	271.43	267.64
H. leuconedys	139.17	12,727.28	12,727.28	12.95	789.07	789.07
C. guereza	156.47	153,705.00	1,774.88	18.64	76.70	44.26
N. larvatus	155.34	155,281.00	3,742.37	14.83	8,192.47	33.54
R. roxellana	159.57	156,202.93	2,808.67	36.39	81.20	45.51
P. nigripes	155.83	3,288.88	3,288.88	14.22	265.70	176.12
E. patas	151.05	729.45	729.45	17.35	180.18	75.31
C. aethiops	147.01	1,256.07	1,256.07	14.13	452.25	130.51
C. sabaeus	141.99	128,888.61	49.10	8.19	6,181.22	34.78
M. silenus	133.64	465.27	465.27	9.43	414.63	414.63
M. assamensis	142.40	2,981.57	2,981.57	11.12	363.22	363.22
M. mulatta	153.13	151,023.96	46,608.97	11.68	11,678.68	7,870.85
M. leucophaeus	152.94	6,426.31	64.15	10.88	2,309.52	52.15
P. hamadryas	159.85	4,056.68	4,056.68	14.73	199.15	137.99
P. anubis	149.12	142,694.50	542.11	8.29	7,290.87	333.77
A. geoffroyi	135.96	4,520.44	4,520.44	23.81	36.29	28.96
C. albifrons	144.13	1,198.43	1,198.43	22.06	70.24	53.12
S. apella	138.60	2,056.99	2,056.99	9.47	137.17	115.09
S. midas	145.81	145,421.96	1,062.64	36.71	64.30	52.59
C. jacchus	146.41	145,914.74	2,023.07	12.79	805.73	103.62
N. pygmaeus	178.59	178,283.75	1,248.08	14.81	12,820.83	176.10
D. madagascariensis	131.89	2,324.97	2,324.97	5.15	632.18	632.18
T. belangeri	134.13	112,564.00	329.13	11.49	65.41	46.46

Fig. 1 | Overview of the X and Y assemblies used in this study. a Metrics of the nonPARX and nonPARY assembly. **b**,**c**, Length (**b**) and relative proportions (**c**) of X-degenerate, X-transposed, ampliconic and heterochromatic regions in the Y chromosomal sequences from each species. The estimated completeness of the Y chromosome is given in parentheses after each species name. The extent of the heterochromatic region is estimated from the unassembled length of the Y chromosome. Relative proportions are calculated as the ratio of each type of region to the total euchromatic regions. Primate pictures are copyrighted by Stephen D. Nash/IUCN/SSC Primate Specialist Group, and are used with their permission in this study.

Across all 30 species we investigated, we found that the PAB shifted at least six times in the ancestral nodes or individual lineages (Fig. 2). The position of the PAB within the third intron of the *XG* gene is conserved in extant OWMs and apes, as including humans, great apes and the rhesus macaque[16,30] (Supplementary Figs. 5 and 6). By contrast, the PAB has shifted more frequently across the NWMs. The boundary was inferred to be located in a region between *CD99* and *XG* in the Platyrrhini, and the PAR length in the Callitrichidae lineage is further reduced (Fig. 2 and Supplementary Figs. 4 and 7).

We also investigated the evolution of the human-specific PAR2, which harbours three protein-coding genes (*SPRY3*, *VAMP7* and *IL9R*). A previous study proposed that the *IL9R* had been translocated to the X chromosome in the primate common ancestor[31], but we found the gene to be autosomal in both the treeshrew and the Strepsirrhini (Fig. 2), suggesting that the translocation occurred in the Simiiformes MRCA. *VAMP7* and *IL9R* are X-specific in all non-human primate species, except in the green monkey *Chlorocebus sabaeus*. In the green monkey, this region spans about 0.14 Mb and exhibits similar male and female normalized sequencing depths (Supplementary Fig. 9a), suggesting that it is located within a PAR. The *Chlorocebus sabaeus* PAR2 is much shorter than the ~0.33 Mb PAR2 of humans and does not include the *SPRY3* gene. Furthermore, in *Chlorocebus aethiops* and *Pan troglodytes*, which are closely related to *Chlorocebus sabaeus* and human, respectively, the orthologous region of PAR2 is X-specific, showing differentiated male and female sequencing depths (Supplementary Fig. 9b,c). These various pieces of findings suggest an independent evolution of PAR2 in *Chlorocebus sabaeus*.

https://doi.org/10.1038/s41559-022-01974-x

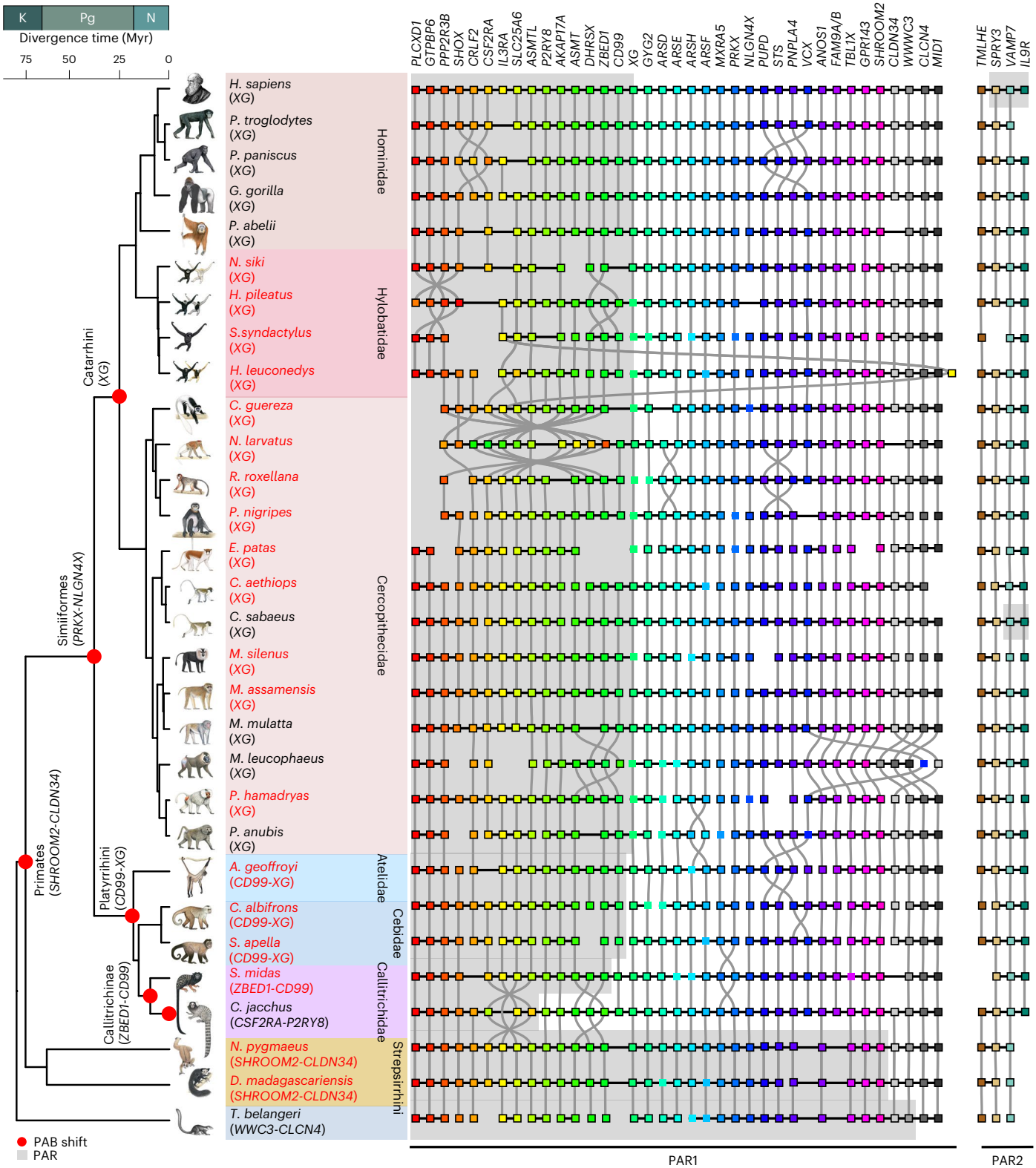


Fig. 2 | The evolution of the primate PAR. PABs of PAR1 are given in parentheses below each species name, in the format of 'PAR gene - nonPARX gene', except in the Catarrhini where the PABs are within *XG*. Genes that are fragmentally assembled in the genome are shown as boxes without a black border. The evolutionary nodes with inferred PAB shifts are labelled with red dots. PAR2 is found only in *Homo sapiens* and *Chlorocebus sabaeus*. PAR genes are denoted by grey shading. Genes on different scaffolds were arranged according to their synteny with the human X chromosome. Genes are linked by black lines in each species if they were assembled in the same scaffold. Orthologous gene relationships across species are linked by grey lines, while different orthologues are encoded with different colours. Only genes known to be present in >25 species are shown. Multi-copy genes are presented as one box. Geological epoch: K, Cretaceous, Pg, Paleogene, N, Neogene. Species name colour: black, published assembly, red, assembly produced in this study. Primate pictures are copyrighted by Stephen D. Nash/IUCN/SSC Primate Specialist Group, and are used with their permission in this study.

Evolutionary strata of primate sex chromosomes

Combining the gametologue phylogeny and the pairwise synonymous substitution rate (*dS*) of the gametologue pairs in each species, we confirmed that all primates share the three ancestral strata (S1–S3), which evolved in the common ancestor of the Eutheria as previous reported[7,10,11] (Fig. 3a and Supplementary Data 4). In contrast to the prosimians, which maintain the ancestral state of the Eutheria, the PAB shift in the Simiiformes MRCA led to the emergence of S4, which is unique to the clade, whereas species in the Platyrrhini and Catarrhini have each independently evolved another new stratum (S5) (Fig. 3a,b). The common marmoset is the only species with an additional stratum, S6, due to the recent expansion of the SDR[25]. Both the non-synonymous and synonymous substitution rates between X and Y gametologues increase with stratum age (Fig. 3c and Supplementary Fig. 10), but the *dN*/*dS* ratio is at its highest for the youngest strata, indicating less constrained evolution for these genes, as was expected owing to their relative recent origin (Supplementary Fig. 10). The GC content of the corresponding sequence follows a decreasing trajectory in Y-linked gametologues with stratum age, particularly at the third codon positions, probably due to older strata having evolved for a longer time in the absence of recombination-associated GC-biased gene conversion (Supplementary Fig. 11).

Among the 29 primates, we inferred 20 ancestral gametologue pairs in S1–S3 including three (*MBTPS2X/Y*, *TAB3X/Y* and *BCORX/Y*) newly discovered in this study[7,11,16,19]. However, it should be noted that these new gametologue pairs are present in only one or two species. Briefly, we discovered *BCORX/Y* in *Sapajus apella* (*dS* = 0.4521) and *Daubentonia madagascariensis* (*dS* = 0.2968), and *MBTPS2X/Y* and *TAB3X/Y* in *D. madagascariensis* (*dS* = 0.3300 and *dS* = 0.3114). The pairwise *dS* values of these X/Y gametologues are similar to those S3 X/Y gametologues (for example, *EIF1AX/Y*) that are intact in most primates (Supplementary Data 4). The gametologue phylogeny also confirms their ancestral origin as being derived from the proto sex chromosome (Supplementary Fig. 12). At least 15 of these ancient gametologue pairs have been retained in over half of the extant primates under study (Fig. 3a). These surviving Y-linked genes have been suggested to play important roles in relation to male viability, and are dosage sensitive with a higher likelihood of haploinsufficiency[11]. One of the most conserved X/Y gametologue pairs was the assumed first therian X/Y gametologue pair, *SOX3/SRY*, both of which are highly conserved in therians[7,11]. As a member of the SOX family, SOX3 interacts with other transcription factors via the high-mobility-group (HMG) domain and participates in gonad and neural development[32,33]. Its Y counterpart *SRY* is the therian sex-determining gene[34,35]. We found that all species retain their *SRY* as expected; however, the X gametologue, *SOX3*, has become pseudogenized in *Saguinus midas* (Fig. 3a). We confirmed with short reads that four indels in *Saguinus midas* SOX3 resulted in a change in the open reading frame, leading to the fragmentation of the HMG domain and the loss of the SOXp domain (Supplementary Fig. 13). It would be worthwhile to study the functional consequence of the loss of *SOX3* in this species and whether another gene may have been recruited to complement its pseudogenization of *SOX3*.

Our data suggest that primate S4 evolved specifically in the Simiiformes but was absent in the Strepsirrhini. We detected three X/Y gametologue pairs in this stratum that have been retained in at least one of the extant species examined, including *VCX/Y*, which is present in human, chimpanzee and two *Hylobatid* species, *Nomascus siki* and *Symphalangus syndactylus*. Previous studies have suggested that *VCY* was acquired in the human–chimpanzee common ancestor due to gene conversion with its X counterpart *VCX*[21,36,37]. We confirmed this conclusion, but our findings suggest an independent origin for *VCY* in the *Hylobatidae*. In contrast to the gametologue pair in human and chimpanzee, the *VCX/Y* pair in the *Hylobatidae* has a relatively high pairwise *dS* value similar to those in other S4 gametologues (Supplementary Data 4), suggesting its more ancient origin and absence of gene conversion.

Previous studies have suggested that a relatively young stratum, S5, evolved in the common ancestor of the Catarrhini[7,16]. Our data provide evidence for the convergent evolution of a similar stratum in the Platyrrhini lineage. While these two strata span the same genomic region and contain up to four X/Y gametologue pairs, they have originated independently after the lineages split 42.8 MYA[20] (Fig. 3a), according to the gametologue phylogenetic topology (Fig. 3b and Supplementary Fig. 14). In the *PRKX/Y* tree, X/Y gametologues cluster in groups of Catarrhini and Platyrrhini rather than by chromosomal origin (Fig. 3b). Human *PRKX* plays a role in kidney development, angiogenesis and other developmental processes[38]. Only partial *PRKX* genes are evident in the genome of five primate species (Fig. 3a), indicating potential independent loss of this gene in these lineages. Additional development-related transcriptomic data need to be investigated to learn more about the underlying consequences. In addition, we found that all the other S5 X/Y gametologue pairs, *MXRA5X/Y, ARSD*, *ARSE*, *ARSF*, *ARSH* and *GYG2*, cluster by groups of Hominoidea, Cercopithecidae and Platyrrhini (Fig. 3b and Supplementary Fig. 14). Although gene conversion is known to obscure gene phylogeny construction[39,40], our finding that all of the S5 genes (*MXRA5X/Y* to *GYG2*) show similar phylogenetic signals suggests that these X/Y divergences occurred independently across the three clades, in contrast to the situation with *PRKX/Y* divergence. In addition, we discovered nine new insertions of X/Y gametologue pairs into S1–S3 strata that were present in one or a few species. These gametologues have low levels of pairwise *dS* (Fig. 3a and Supplementary Data 4) and their X and Y counterparts cluster together (Supplementary Fig. 15), suggesting their recent duplication from X to Y. Out of the six X-transposed Y genes for which transcriptome data were available, four of them were found to exhibit a testis-specific expression pattern (Supplementary Fig. 16).

Lastly, we used gametologue diversity to quantify the relative change in the per base pair genomic mutation rate for each species over the past 180 Myr since the formation of the first stratum. Specifically, we compared the observed *dS* values of each stratum with our expectation dS_{exp} that is based upon mutation rates that have been estimated by parent–offspring sequencing studies (Supplementary Data 5). The ratio dS/dS_{exp} quantifies the change that the genome-wide mutation rate experienced at different timepoints defined by stratum formation, relative to the extant mutation rate. Notably, dS/dS_{exp} decreased across S1–S3 (Fig. 3d), in line with previous evidence of a mutation rate slowdown in primates relative to the Simiiformes MRCA[41], especially for the apes where the discrepancy between *dS* and dS_{exp} is at its highest[42]. On average, a ~3.3-fold higher mutation rate than that estimated by parent–offspring sequencing studies would be required to achieve the *dS* level observed for S1. In similar vein, a 0.9–2.9-fold difference in mutation rates would be needed to account for the observed *dS* of S2–S5 since their formation.

Genomic re-arrangement in Simiiformes-specific strata

Structural variations between the X and Y chromosomes have been proposed to contribute to the recombination suppression between them. On the basis of the observation that pairwise *dS* values of the X/Y gametologue clusters vary greatly among strata along the X chromosomes[10], previous studies hypothesized that the Y chromosome has experienced stepwise recombination suppression region by region[10,28]. The oldest stratum S1 is suggested to have evolved at the MRCA of all therians ~181 MYA; eutherian S2/3 formed later, beginning ~117 MYA[7]. With X/Y chromosomes available from more species, we are able to extend the investigation more broadly to additional taxa, and to assess the genomic re-arrangements of the X and Y chromosomes in different primate species. The primate X chromosomes have a high degree of conserved synteny in relation to the outgroup and have rarely experienced large re-arrangements over the past 81.76 Myr;

https://doi.org/10.1038/s41559-022-01974-x

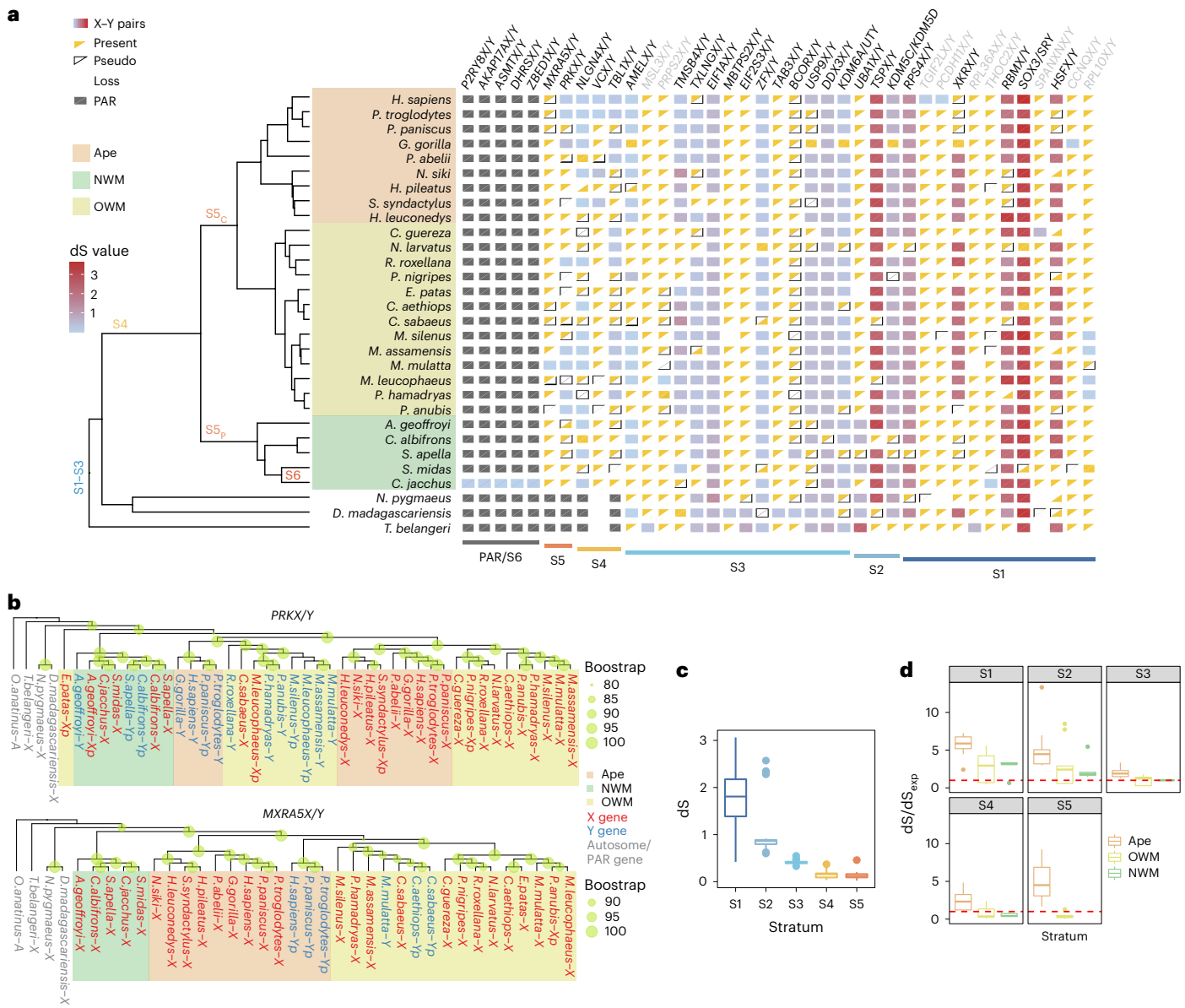


Fig. 3 | Evolutionary strata of primates. a, X/Y gametologues and the evolutionary strata in which they are located. The gametologue pair box is colour coded according to the X/Y pairwise *dS* value if both X-linked and Y-linked gametologues were intact. If only the X or Y gametologue was available, the upper triangle represents the presence of the X-linked gene while the lower triangle represents the presence of the Y-linked gene. Yellow: gene with intact open reading frame, blank: lost gene, white with black outline: pseudogene, grey: PAR gene. Gametologues have intact copies in both the X and Y but without *dS* values because they were missing in the assemblies and have been recovered from raw sequencing reads and RNA-seq data. The approximate placement of strata on branches is based on the topology of individual gene trees and *dS* values, after consideration of their placement in previous studies[7,10,11]. $S5_C$ and $S5_P$ indicate the stratum 5 convergently evolved in Catarrhini and Platyrrhini, respectively. Gametologue pair names colour: black, derived from the proto sex chromosome, grey: Y newly transposed from X. **b**, Phylogenetic trees of two examples of S5 X/Y gametologues, *PRKX/Y* and *MXRA5X/Y*. Gene names are colour coded according to chromosomes (red, X; blue, Y; grey, PAR). Leaf node background is colour coded according to lineage (yellow, OWM; red, apes; green, NWM). Pseudogenes are also included and are marked with a 'p' suffix. **c**, Distributions of *dS* values calculated using the KaKs_Calculator for concatenated stratum sequences of each species. $n = 30, 30, 30, 19$ and 8 in the order from S1 to S5 (treeshrew and prosimians are included in S1–S3). Box plots show median, quartiles (boxes) and range (whiskers). **d**, Mutation rate evolution across different strata. The distribution of the dS/dS_{exp} ratio for each stratum and species across different timepoints defined by strata formation events. The *dS* values were calculated from the X/Y alignment of concatenated CDSs in a given stratum. The dS_{exp} values represent the expected rates of synonymous substitution based on mutation rates estimated by parent–offspring sequencing studies (Supplementary Data 5). The red dashed line represents the dS/dS_{exp} ratio of 1, expected in the case of no mutation rate change. In S1–S3, $n = 9, 13$ and 5 for ape, OWM and NWM, respectively. In S4, $n = 7, 9$ and 3 for ape, OWM and NWM, respectively. In S5, $n = 3$ and 5 for ape and OWM, respectively. Box plots show median, quartiles (boxes) and range (whiskers).

one notable exception was an ~14 Mb inversion on the long arm of the pygmy slow loris (*Nycticebus pygmaeus*) X chromosome, which was further confirmed with Hi-C map data (Supplementary Fig. 17). The highly conserved X chromosome that resembles the primate ancestral stage therefore provides a background for the investigation of structural alterations to the Y chromosomes following the formation of the Simiiformes-specific strata, S4 and S5. Owing to their comparatively recent formation, most of the alignable regions between the

X and Y chromosomes are found in these two strata (Supplementary Fig. 18). Utilizing the Hi-C scaffolded or chromosome-scale Y assemblies from eight primate species covering the NWMs, OWMs and apes, along with the human X as the outgroup, we detected 21 conserved syntenic blocks under 50 kb resolution across species within these two regions and the PAR, and found a high frequency of inversions in these regions that had originated during their diversification (Fig. 4a). For instance, S4 was likely to have originated via an inversion[9] at Simiiformes MRCA (Fig. 4a). Further, S5 of Catarrhini first experienced four inversions that shuffled the sequence order of both S4 and S5, and then acquired several lineage-specific changes in each extant species (Fig. 4a). We found that OWMs, especially *Rhinopithecus roxellana*, largely maintained the Catarrhini ancestral Y structure after the evolutionary emergence of S5. By contrast, the Y chromosomes of the Hominoidea experienced more structural changes resulting in more diverse Y structures in S4 and S5 among the extant species. Similar results were found from the reconstruction under 30 kb resolution (Supplementary Fig. 19).

The extensive genomic re-arrangement of the Y chromosomes motivated us to investigate whether the chromatin configuration of the homologous regions between the two sex chromosomes was altered after recombination suppression[3]. Here, on the Hi-C interaction map[43], we estimated the chromatin configuration differences between the X/Y homologous regions of the new SDR, that is, S4 and S5. Of the five species for which we could detect sufficiently long homologous sequences between the X and Y, four showed significant changes in chromatin configuration after transformation from the PAR to SDR, when compared with the control distribution represented by the difference in chromatin configurations between autosomal alleles (FDR-corrected P value <0.01; Fig. 4b,c, Supplementary Fig. 20 and Supplementary Data 6). Such chromatin contact change on the sex chromosomes was also observed in the recently emerged stratum in the Z/W system of the emu[44]. Nevertheless, this pattern does not appear to be present in the strata that have evolved very recently in the eel X/Y chromosomes[45], which still maintain high similarity between X and Y homologous sequences and might be too young to have established such a difference in chromatin configuration. By contrast, when measuring the chromatin configuration difference between Simiiformes and Strepsirrhini (or treeshrew) along the X chromosome, we did not find a significant difference between the new SDR and the ancestral SDR (Wilcoxon rank-sum test, P value >0.05; Supplementary Fig. 21 and Supplementary Data 7 and 8). Additionally, it would appear that both the new SDR and the ancient SDR on the X chromosome maintain the same level of topologically associating domain (TAD) boundary conservation across the Simiiformes (Fig. 4d, Supplementary Fig. 22 and Supplementary Data 9). Specifically, about 50% of TAD boundaries are conserved between Simiiformes and treeshrew in both the new SDR (S4 and S5) and the ancestral SDR (S1–S3). Between Simiiformes and pygmy slow loris, about 60% of TAD boundaries are conserved in both the new and the ancestral SDR. These results suggest that the configuration difference between the X and Y in new strata may be largely attributable to alterations on the Y chromosome, probably after rather than before the recombination suppression. It also supports the association between epigenetic modification changes and heterochromatinization during Y chromosome evolution[3].

Evolutionary dynamics of genes on the Y chromosome

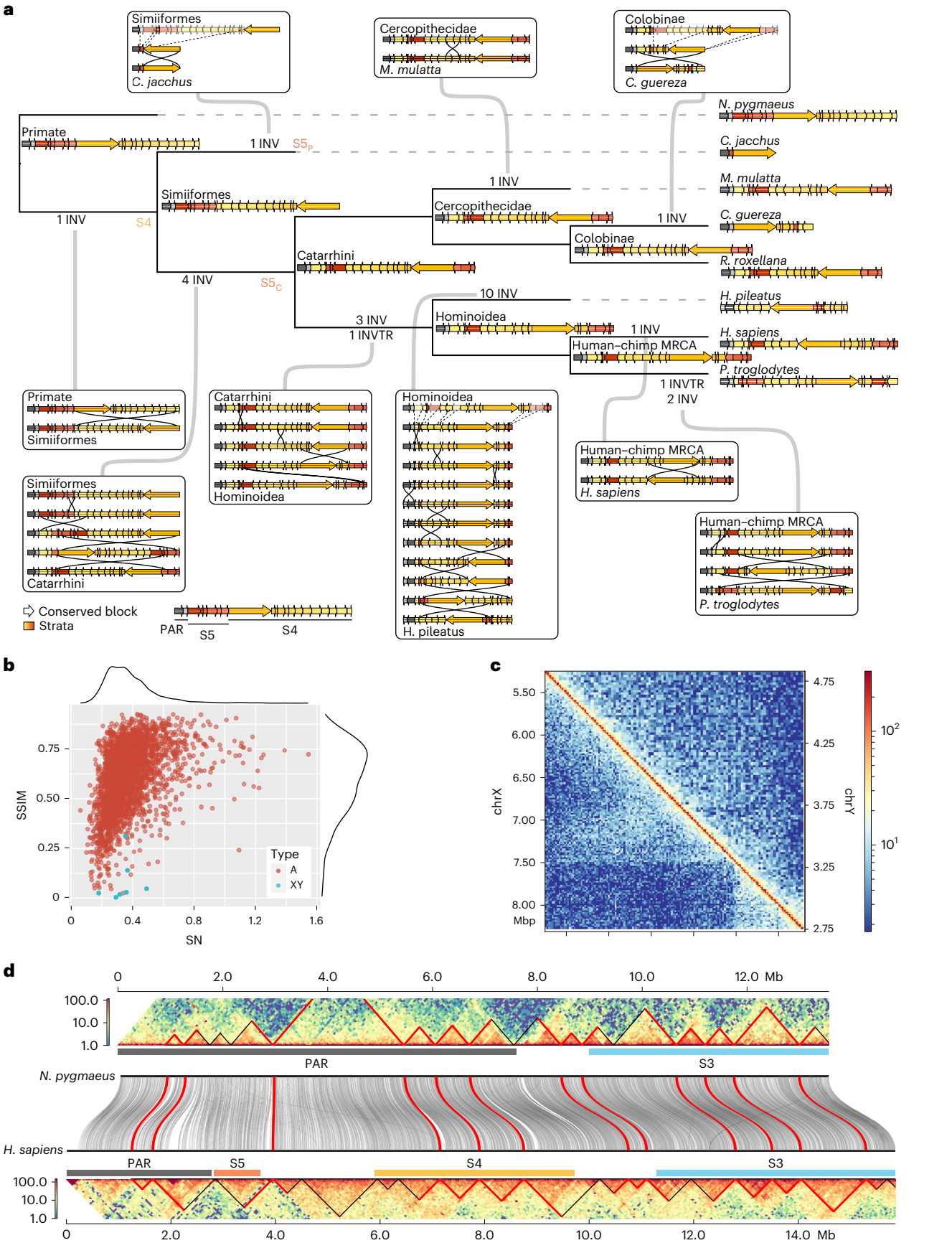
In contrast to the X chromosome, which has maintained most of the ancestral gene properties of the proto sex chromosome, the Y chromosome has experienced dramatic gene loss facilitated by recombination inhibition, and therefore possesses markedly different gene properties between extant primate species. It has been shown that degeneration proceeds quite rapidly on newly formed Y chromosomes, but the rate decreases in evolutionarily older Y chromosomes[11,16]. This degeneration model was confirmed by the different preservation pattern of genes between the therian ancient strata and the Simiiformes-specific strata. While most Y-linked genes in S1–S3 were lost before the Simiiformes radiation, the remaining 20 genes were preserved in most primates over the last 81 Myr, with relatively few (on average 38.89%) being lost in individual branches of the primate phylogeny (Fig. 5a and Supplementary Figs. 23 and 24). By contrast, S4 and S5 experienced a much more rapid gene loss in Simiiformes species with on average 84.79% and 94.93% gene loss over the past 81 and 42 Myr, respectively (Fig. 5a, Supplementary Fig. 23 and Supplementary Data 10 and 11), probably due to a stronger Hill–Robertson interference effect when more genes are present in younger strata[46]. Of the at least 16 ancestral genes in S4 and S5, a maximum of 4 have been maintained as gametologue pairs within the same species. With more primate species, we were able to refine the degeneration model of these two young strata on the basis of the number of the surviving Y chromosomal genes in each evolutionary node. The observed patterns within S4 and S5 are generally consistent with the hypothesized degeneration model[11,16] (Fig. 5b and Supplementary Data 11). Notably, more Y-linked genes of the Platyrrhini have undergone degeneration in both S4 and S5, resulting in fewer Y-linked genes being present in this region (on average 0.80 and 0.20 in Platyrrhini versus 1.31 and 0.41 in Catarrhini for S4 and S5, respectively) (Fig. 5a). We were also able to trace at which ancestral nodes the degeneration of each Y-linked gene occurred (Supplementary Fig. 25). Thus, for example, in human the extant S4 Y-linked gene set had already formed at the Simiiformes MRCA; by contrast, the S5 Y-linked genes experienced a two-step inactivation that occurred at the MRCA of Catarrhini and apes.

Although our observation indicates that most genes in S4 and S5 were lost ancestrally at Simiiformes MRCA, one exception is provided by *NLGN4Y* of S4 which was still present in over a half of extant Simiiformes, suggesting that the gene may have played an important role in this lineage. A recent study postulated that one of the amino acid differences, Pro93Ser, between the human NLGN4X and NLGN4Y proteins might be key to understanding the mechanism underlying the male bias observed in NLGN4X-associated autism spectrum disorder[47]. We found that this amino acid substitution in NLGN4Y occurred in the ancestral node of Simiiformes and has been maintained in all descendant lineages except *Callithrix jacchus* and *Sapajus apella* (Supplementary

Fig. 4 | Structural and chromatin configuration changes in S4 and S5 during primate sex chromosome evolution. a, Reconstruction of the ancestral Y chromosome sequence order with conserved syntenic blocks under 50 kb resolution. A total of 21 conserved blocks were used in reconstruction. Each arrow block represents a conserved segment, colour coded in PAR, S5 and S4. The postulated genomic re-arrangement events marked at each branch were inferred by GRIMM and manual curation. INV, inversion; INVTR, inverted translocation. **b**, Homologous blocks between X and Y (blue) harbour lower chromatin configuration similarity than the regions between homologous chromosomes (red), indicating that the contact map changed after X/Y divergence. A, diploid autosomal background; XY, X/Y homologous block. n = 7 and 2,335 for XY and A dataset, respectively. **c**, Comparison of Hi-C contact maps as an example of an X/Y homologous block in *Macaca mulatta* (chrX:5,249,201-8,310,709(+), chrY:2,669,614-4,768,941(−)). Upper triangle depicts the Y Hi-C matrix, whereas the lower triangle depicts the X Hi-C matrix. **d**, TAD boundaries are conserved on the X chromosome after the evolutionary emergence of S4 and S5. Top: Hi-C map and TADs in *Nycticebus pygmaeus*. Middle: alignment between *Nycticebus pygmaeus* and *Homo sapiens*. Bottom: Hi-C map and TADs in *Homo sapiens*. Conserved TAD boundaries are highlighted with red lines, while species-specific TAD boundaries are delineated with black lines. In this example, six of the ten TAD boundaries in the recently degenerated S4 and S5 of *Homo sapiens* are conserved with those of the PAR of *Nycticebus pygmaeus*. This level is similar to that in the ancestral stratum S3 where six of the nine TAD boundaries are conserved between the two species.

https://doi.org/10.1038/s41559-022-01974-x

Fig. 26). We also identified three other X/Y variants in this gene that were conserved across the Catarrhini (Supplementary Fig. 26). The high level of conservation of two of the X/Y variants in NLGN4Y implies that these amino acid residues have been important for male development in simians, with a hint of a potentially deleterious side effect by virtue of a contribution to the male bias in autism spectrum disorder.



https://doi.org/10.1038/s41559-022-01974-x

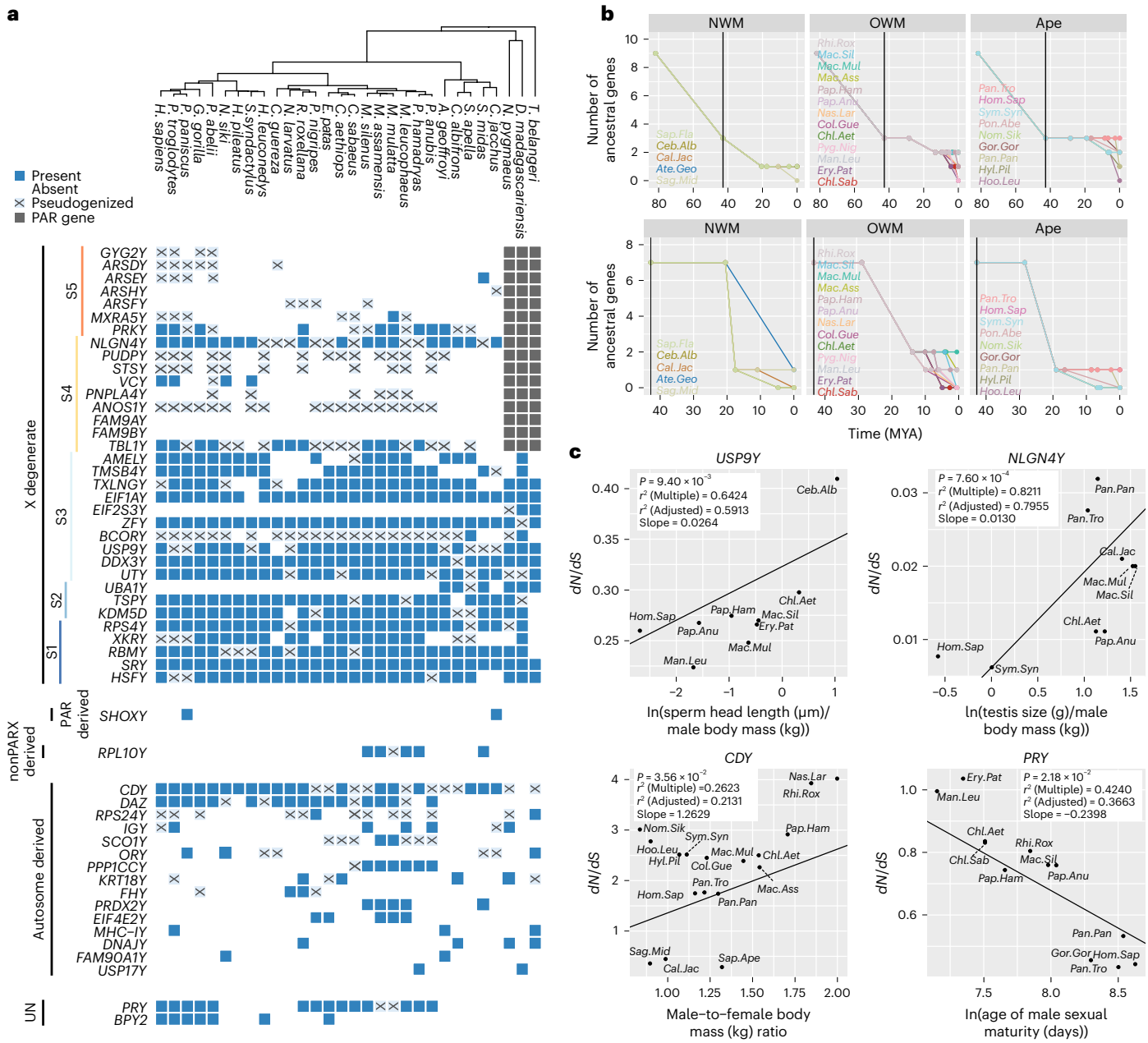


Fig. 5 | Dynamic evolution of primate Y chromosomes. a, Heat map depicts the state of Y-linked genes (present, pseudogenized or absent) in each species in this study. For genes outside S4 and S5, only Y-linked genes available in more than one species are shown. X-degenerate, PAR derived and nonPARX derived genes are ordered according to the genomic coordinates of their X-linked gametologues. Autosome-derived genes are ordered on the basis of the number of species carrying these genes. UN, unknown. **b**, Degeneration model for Y-linked genes in S4 and S5 during primate evolution. In both strata, the Y gene degeneration rate is fast initially, but then decelerates eventually reaching a stable level in recent time. The degeneration rate of NWM is more rapid than that of OWM and apes in both S4 and S5. The Platyrrhini–Catarrhini split is marked by a black line. **c**, *dN*/*dS* values of the Y-linked genes *USP9Y*, *NLGN4Y*, *CDY* and *PRY* exhibit a correlation with the sperm head length (μm) to male body mass (kg) ratio, testis size (g) to male body mass (kg) ratio, male-to-female body mass ratio and age of male sexual maturity (days), respectively, with PGLS analysis. Log transformation was applied to all traits except body mass dimorphism ratio. Two-sided *F*-test is used in PGLS analysis, and we do not apply multiple testing correction to adjust the *P* value.

We also detected a significant correlation between the *dN*/*dS* values of several Y-linked genes and male developmental traits across the primates, for example, age of male sexual maturity, sperm morphologies, testis size and level of sexual dimorphism in relation to body mass (Fig. 5c, Supplementary Fig. 27 and Supplementary Data 12 and 13), after accounting for the phylogenetic covariance. These traits are widely believed to evolve in response to sexual selection which has been reported to favour genes on the Y chromosome that enhance male fertilization success[48–50]. We have found that *USP9Y* and *NLGN4Y* exhibit a positive correlation between *dN*/*dS* values and sperm head length and testis size (both scaled by body mass and log-transformed), respectively. Moreover, the evolution of *CDY* may correlate with sexual dimorphism in relation to body mass and the age of male sexual maturity. Interestingly, *CDY* is one of the few Y-linked genes that have not degenerated from the ancestral sex chromosome but were instead derived from an autosome prior to the adaptive radiation of the primates. In contrast to the autosomal gene *CDYL* from which *CDY* originated, the Y-linked gene has experienced a change from a ubiquitous

https://doi.org/10.1038/s41559-022-01974-x

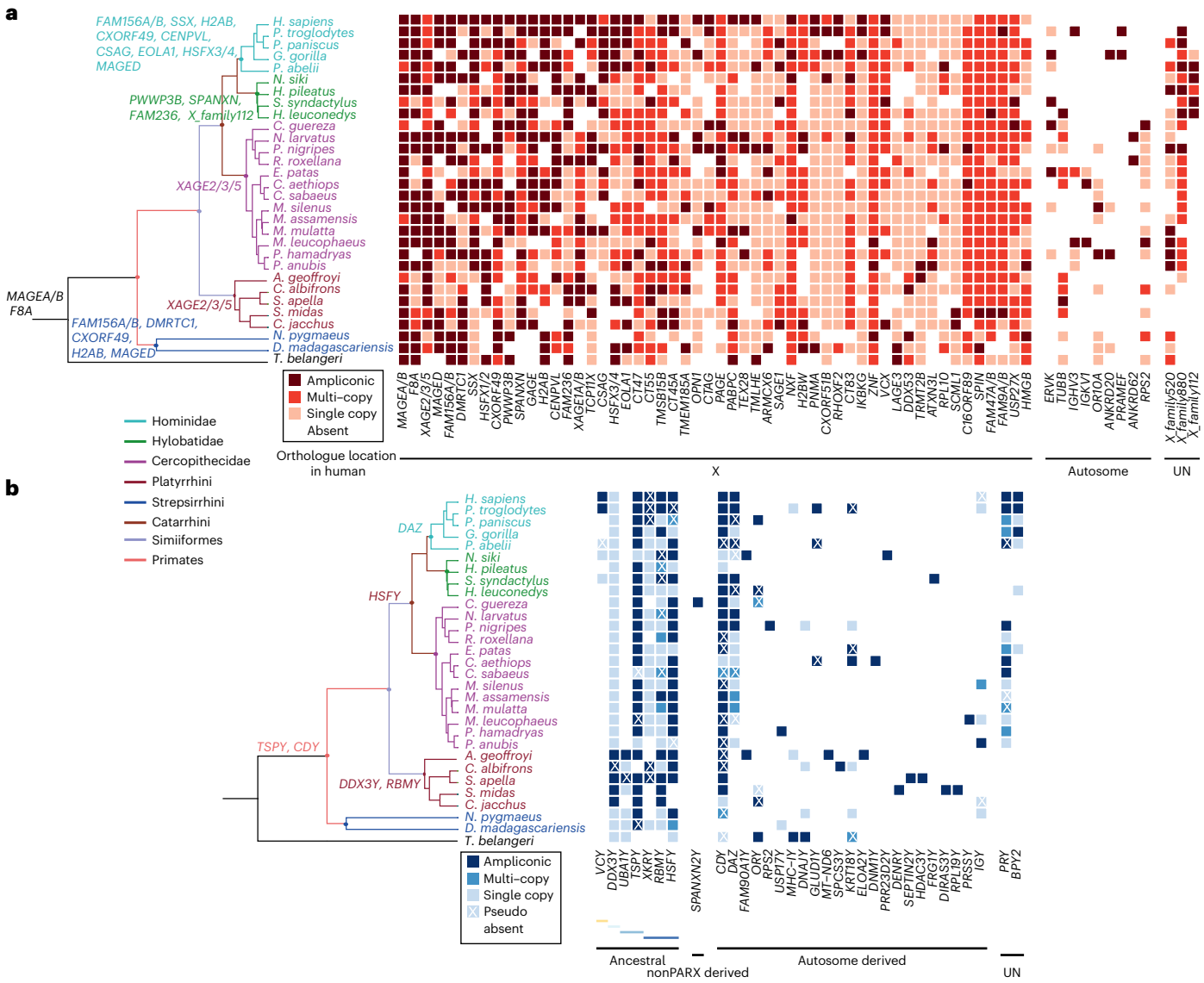


Fig. 6 | Comparison of X-linked (a) and Y-linked (b) families involving AGs across the primates and treeshrew. a, Comparison of the X-linked gene families involving AGs across the primates. Only families with AGs in more than one species are shown. Families were ordered by the number of species containing AGs. **b**, Comparison of the Y-linked gene families involved AGs. Pseudogenes are marked by crosses. Families were organized in the order of ancestral, nonPARX derived, autosome derived and unknown (UN). Ancestral Y-linked genes were ordered by stratum age. The other three types of gene were ordered by the number of species containing AGs. Differently coloured branches in the phylogenetic tree represent different taxa. Genes in different colours represent different ancestral origins.

pattern of expression to testis-specific expression, an example perhaps of neofunctionalization[51–53]. We were also able to detect dozens of new Y chromosomal genes that appeared during the process of primate diversification (Fig. 5a). These were mainly derived from the autosomes with only a few cases being duplicated from the X chromosome. Most of these genes are present in only one or a few species (Fig. 5a), suggesting that their acquisition was lineage specific (Supplementary Fig. 28).

AG evolution across primates

AGs, defined as genes with highly similar copies (sequence identity >99%), are common in most sex chromosome systems, including human, chimpanzee and mouse[22,51,54] as well as *Drosophila*[55]. These genes are predominantly expressed in the testis, are often important for male fertility[54,56], and have been proposed to be associated with sexual selection through traits such as mating system[17,57] and sex-linked meiotic drive[22,58]. We searched for AGs on both sex chromosomes in all primate species by identifying genes located in genomic regions that occur as duplicated copies with high identity, or which have more than twofold sequencing depth compared with single-copy regions and hence may have been collapsed in the assemblies. Employing this method, we were able to detect AGs on both the X and Y chromosomes of all primate species (Supplementary Data 14 and 15), with the exception of the Y chromosome of the aye-aye (*D. madagascariensis*) which possesses a very small ampliconic region (Fig. 1b). We also identified genes characterized by multiple duplicated copies but with sequence identity lower than 99%, and found that these multiple copy genes generally belong to gene families that also harbour AGs (Fisher's exact test, $P = 1.09 \times 10^{-11}$; Supplementary Data 16), suggesting that at least a portion of these multiple copy genes may once have been AGs but accumulated substitutions after the cessation of gene conversion.

In total, we identified 178 X-linked gene families across all primates that harboured AGs in at least one species. Becoming ampliconic

appears to be a highly species-specific phenomenon, with only 66 X-linked gene families containing AGs in more than one species (Fig. 6a). The MAGEA/B family is one of the most highly conserved gene families with AGs in 24 species spanning all major primate groups (Fig. 6a). This family also has AGs in mice[54] and members of the family are predominantly expressed in testis[59], suggesting that this AG family has been evolving at least since the Euarchontoglires, having been maintained ampliconic for more than 96 Myr[60].

As in the situation encountered with the X-linked AG families, from a total of 32 Y-linked families containing AGs in primates, 21 are AG in only one species. Nevertheless, we found that several Y-linked AGs, such as *TSPY*, *HSFY* and *CDY*, are AGs in over half of the primates studied (Fig. 6b). For example, *TSPY* is an important factor for spermatogenic efficiency[61], and the amplification of *TSPY* might therefore have been favoured by sexual selection during the evolution of the primates. Of note, some AGs are only present in specific taxa. For instance, *DDX3Y* and *UBA1Y* were found to be AG specifically in NWM. In particular, *DDX3Y* was present as AGs in most investigated species of NWM. This gene is thought to be a major azoospermia factor in humans, and its knockout results in the reduction or even absence of germ cells[62]. The amplification of *DDX3Y* could be attributed to the multimale–multifemale and polyandry mating systems in NWM[63], which increase the degree of sperm competition. A previous study reported three waves of duplication and transposition of *DAZ*, another azoospermia factor, into the human Y chromosome[64], starting with the ancestors of the Catarrhini. Consistent with this, we found that *DAZ* constituted an AG family in most Catarrhini, with the exception of the Cercopithecinae, in which multiple copies have accumulated mutations, probably due to a lack of gene conversion events between copies.

Previous studies have also found massive co-amplification of AGs on the X and Y chromosomes in mouse and bovine, suggesting a possible role in genomic conflicts over sex chromosome transmission[22,58]. Interestingly, we detected two examples of co-amplified genes on the X and Y chromosomes, *HSFX/Y* and *VCX/Y*, in several primate lineages. The co-amplification of *VCX/Y* was found to be confined to human and chimpanzee, consistent with a previous study[17,21]. By contrast, co-amplified *HSFX/Y* were found in 11 species from NWM, OWM and apes (Supplementary Data 17). This gametologue pair has also been reported to be co-amplified in bovine, with high copy number variation on the Y chromosome[58]. It is possible that the co-amplification of *HSFX/Y* might have been the consequence of sexually antagonistic selection or meiotic drive.

Discussion

Our study fills a number of important phylogenetic gaps pertaining to the structure and gene properties of the sex chromosomes in primates and broadens our knowledge of sex chromosome evolution across the primate phylogeny. This unprecedented data resource has allowed us to identify the key changes that has occurred on the sex chromosomes in each evolutionary node and to disentangle the evolutionary patterns that occurred in the common ancestor of all primates as well as particular events in specific lineages. Compared with the X, our analyses revealed rapid dynamic change on the Y chromosome that accompanied the diversification of primates, for example, the two young strata (S4 and S5) that evolved in the Simiiformes ancestor and Catarrhini (apes and OWMs) ancestor respectively, while the NWMs also evolved an S5 stratum quite independently. Interestingly, the two new strata would appear to have experienced different rates of degeneration across primate groups, probably due to different background selection or rate of genetic drift. In line with previous findings in humans and a few other primates, we found that primates have not only expanded their male-specific regions by extending the X-degenerate region, but have also continually accumulated new gene copies either by translocation from the rest of the genome or by the amplification of Y-linked genes, mediated potentially by palindromic sequences. These AG families are dominated by testis-specific genes that have remained highly identical as a consequence of the action of interallelic Y–Y gene conversion, suggesting a selective advantage conferred by their increased gene copy number. Interestingly, we observed that AG families were highly dynamic across primate lineages, possibly due to variation in sexual selection and background selection pressures between species. An accurate and complete indexing of all ampliconic and palindromic sequences would require a complete telomere to telomere assembly of the Y chromosome, which still remains a technical challenge even for the human genome[15]. Nevertheless, our study has illuminated many key evolutionary features of the Y chromosomes of primates. With improvements in sequencing technology and the sequencing of additional primate species, comparative analyses on a phylogenetic scale will enable us to acquire a much better understanding of the molecular basis of the diversity of sex-linked biological traits across the primates, with potential medical significance for studies on human reproduction and health that pertain to Y-linked genes[14,57,65–67].

Methods

Data type and species used in each analysis are listed in Supplementary Data 18.

Ethics, sample preparation and sequencing

Blood samples of female *Macaca silenus*, *Papio hamadryas*, *Erythrocebus patas* and *Colobus guereza* and a muscle sample of female *Hylobates pileatus* were collected from the Japan Monkey Centre, Japan. The use of genetic materials of Japan Monkey Centre was approved by the Research Ethics Committee from the Japan Monkey Centre (#2017-018) and performed in accordance with the Ethical Guidelines for Research at the Japan Monkey Centre (1 April 2016). DNA was extracted from the tissue samples with QIAGEN DNeasy Blood & Tissue Kit (50), following the manufacturer's instructions. Library preparation was performed with Illumina TruSeq DNA PCR-Free Kit (Illumina) and sequencing was run on the Illumina NovaSeq 6000 system (151 bp × 2 PE).

Identification of X-linked and Y-linked sequences

To identify the X-linked and Y-linked sequences in each primate species, we used a similar method to that employed in ref. [25]. Male and female resequencing short reads were mapped to each genome using BWA MEM (v0.7.17) (ref. [68]). For species lacking female short reads, we used short reads from female individuals from the same genus (Supplementary Data 1 and Supplementary Notes). Coverage was extracted with samtools (v.1.9) (ref. [69]), normalized by the peak coverage, and was then calculated in 5 kb windows with bedtools (v2.29.2) (ref. [70]). Scaffolds (>10 kb) of over 60% of windows with normalized F/M coverage ratio between 1.5 and 2.5 were identified as X-linked, and between 0.0 and 0.3 as Y-linked. Coverage of candidate X- and Y-linked scaffolds was also visualized with ggplot2 (v3.3.5) (ref. [71]) and manually examined to delineate the PAR and nonPARX/Y within each scaffold. For species with PacBio or Nanopore long reads, we also mapped long reads to the genome, and further de-collapsed the nonPARY region with SDA (git commit 4ca0c0709dc86181afaeeee862543dc19a411eb3) (ref. [72]), using the half value of the peak long reads coverage of the genome (Supplementary Notes). The sequences assembled by SDA were merged with other sequences using the method described in ref. [25] and polished with male short reads data by means of the freebayes-bcftools polishing pipeline (freebayes v1.3.1-17-gaa2ace8, bcftools v1.11) described in ref. [73]. We further BLASTPed the protein sequence translated from the nonPARY genes to the NR database (access date: 20190727) with diamond (v2.0.0) (ref. [74]). The BLAST hit was then further processed with BASTA (v1.3.2.3) (ref. [75]) for categorization. If over 80% of the genes on a nonPARY candidate scaffold were categorized to be from a non-primate source, the scaffold was removed.

Y gene set annotation

We combined the homologous method and the de novo method to annotate genes on nonPARY sequences. Homologous gene annotation was first performed with GeneWise (v2.4.1) (ref. [76]) using protein sets from the X and Y chromosomes of human, rhesus macaque and mouse collected from Ensembl 98. De novo gene annotation was performed with Augustus (v3.2.3) (ref. [77]), with pre-trained human parameters on N-masked sequences. We further excluded possible retrogenes or pseudogenes from the annotation using the following criteria: (1) containing more than two frameshifts or premature termination signals, (2) annotated as a single-exon gene while the query was a multi-exon gene, (3) the coding sequence (CDS) was covered with ≥60% of repeats. These two datasets were further merged to construct a non-redundant gene set, and the dataset was then used to BLASTP the Swiss-Prot database (release date May 2020). Genes BLASTed to the database but with an alignment rate ≤0.7 were excluded, and gene names were obtained by their best hits on the Swiss-Prot database. Frameshift signals within genes that passed the above criteria were confirmed with male short reads from the same individual (Supplementary Notes and Supplementary Data 19 and 20). The absence of Y-linked genes was confirmed with male long reads and short reads (Supplementary Notes and Supplementary Data 21–23).

Y completeness evaluation

Karyotype images of several studied primates were collected from previous studies (Supplementary Data 2) and the expected size ratios between the X and Y chromosomes were calculated using the method described in ref. [73]. We used the size of identified X-linked and PAR scaffolds to calculate the size of the assembled X chromosome. The assembled X chromosome size and the expected X-to-Y ratio were then used to calculate the estimated Y chromosome size in each species. The completeness of the assembled Y-linked scaffolds was calculated as the proportion of the assembled ungapped Y-linked scaffold length relative to the expected Y chromosome length.

Y structural analysis

We split the Y-linked sequences into 5 kb windows and performed all-versus-all BLASTN sequence comparisons[78]. If a BLASTN hit had an identity >99% and an alignment rate >50%, the windows were considered as candidate ampliconic regions. We also used the male sequencing depth to identify windows that might still have collapsed in the assembly and hence could not be detected by means of the BLASTN method. We compared the mean sequencing depth of each window to the mean sequencing depth of the windows covering the single-copy orthologues in most species (*SRY, KDM5D, TXLNGY, XKRY, UTY, USP9Y, TMSB4Y, AMELY and NLGN4Y*). Where the sequencing depth fold change to the control depth was ≥2, the window was considered to be a candidate ampliconic region. The two resulting ampliconic datasets were merged with bedtools to obtain a final ampliconic region. The X-transposed windows were defined by 5 kb windows that overlapped with X-transposed genes characterized in the strata analysis. We also BLASTNed the nonPARY 5 kb windows against the nonPARX sequences. Windows with a BLASTN hit of identity >90% and an alignment rate >80% were retained. These 5 kb windows were then merged with 'bedtools merge -d 15000'. Only merged regions of >1 Mb were retained as being X-transposed region. We further curated the human X-transposed region by taking the smallest continuous region that could cover the merged segments. Heterochromatic regions were defined as windows for which >80% of the sequences were satellite repeats, based on the definition in ref. [8]. When combining the three datasets, we assigned the windows with the following priority: X-transposed, ampliconic, heterochromatic. The remaining regions were defined as X-degenerate. Estimated heterochromatic regions were defined as regions that were not assembled but were instead estimated from karyotype. To confirm that the large X-transposed region is unique to human, we obtained each X-transposed gene and its flanking 50 kb region sequence. The X counterpart and the flanking 50 kb region were also obtained. The Y region was split into 100 bp windows and BLASTNed to the X region. Solar (v0.9.6) was used to chain the alignment with parameter '-a est2genome2' and the longest chained block was retained. Average identity in every neighbouring three windows (that is, 300 bp) was calculated and visualized by pyGenomeTracks (v3.7) (ref. [79]).

PAR identification and PAB inference

Some PARs could be identified by the sequencing depth method if the sequence was assembled with nonPARX or Y, but using only this method could have generated some false negatives. On the basis of the assumption that the PAR and its flanking region are relatively well conserved among primates, we used human as the reference and produced whole genome alignments to identify PARs in other species. We treated a scaffold as a candidate PAR scaffold if over 70% of the sequence was aligned to the human X and most of the scaffold was aligned to the human X 0–12 Mb region. We further confirmed the candidate as a true PAR if it had similar normalized male and female resequencing depth. To annotate genes on PAR1, PAR2 and their flanking region, genes from the first 12 Mb (from *PLCXD1* to *MID1*) and the last nine genes (from *TMLHE* to *IL9R*) on human chrX were used as queries. Homologous annotation was performed with GeneWise (v2.4.1) and Exonerate (v2.4.0) (ref. [80]), and de novo gene annotation was performed with Augustus (v3.2.3). Human PAR was obtained from https://www.ncbi.nlm.nih.gov/grc/human. To check if PAB is conserved among Simiiformes[16], we first performed whole-genome pairwise lastZ (v1.04.00) (ref. [81]) alignment between each species with human, with parameter set '–step=19 –hspthresh=2200 –inner=2000 –ydrop=3400 –gappedthresh=10000'. The human PAB location was then liftOvered into each species. We further examined the normalized male and female sequencing coverage under 200 bp window resolution. If the normalized F/M coverage ratio was changed abruptly at the liftOvered location, we considered the PAB conserved between human and the primate species. Marmoset PAR was obtained from ref. [25]. The ancestral PAB was inferred in a parsimonious way, under the assumption that once X/Y divergence had occurred and transformed a gene from PAR to SDR, it could not then revert to PAR. We also considered the evolutionary history of S4 and S5 in the inference.

Strata analysis

To establish X/Y gametologues, we searched for the closest homologues between X and Y genes using encoded amino acid sequences and running a BLASTP search, with e-value cut-off 1×10^{-5}. The gametologue pairs were further examined in terms of their hits to the Swiss-Prot database. Only pairs of both X-linked and Y-linked genes that BLASTPed to the same gametologue were retained for further analysis. We then aligned the CDSs of candidate X/Y gametologue pairs and their orthologues with PRANK (v170427) (ref. [82]), built phylogenetic trees with RAxML (v8.2.4) (ref. [83]) and manually examined to remove false positive gametologue pairs caused by paralogues. Additionally, we incorporated gametologues found in other studies[25]. Gametologues recovered from raw reads in the Y gene-set annotation procedure were included but were not used for evolutionary analysis due to their high sequencing error rate. To calculate the pairwise *dS* value, we first aligned the protein sequence of each gametologue pair with PRANK and converted the protein alignment into the CDS alignment. Pairwise *dS* values were calculated with KaKs_Calculator (v2.0) (ref. [84]). Gametologue pairs were grouped into different strata according to the pairwise *dS* values and their assignment in human[10,11]. To further confirm the evolutionary history of S4 and S5, we aligned the entire gametologue gene region, including both CDS and intron, with MAFFT (v7.471) (ref. [85]), and constructed the phylogenetic tree with RAxML.

dS analysis

To increase the accuracy of estimation of species-specific population genetic parameters, we concatenated CDSs of gametologues within each stratum. To quantify the change of mutation rate between different strata and species, we calculated the *dS* of the concatenated sequences using the KaKs_Calculator. The expected values of *dS* (dS_{exp}) were based on empirically estimated extant mutation rates as shown in Supplementary Data 5.

The expected dS_{exp} values were calculated as

$$dS_{exp} = \frac{4(1+2\alpha)}{3(1+\alpha)}\mu t,$$

where α is the male mutation bias (in this case we used $\alpha = 3$ for all strata and species, as determined for mammals[86]), μ is the annual autosomal mutation rate and t is the time of stratum formation[5,9,11]; S1 = 181 MYA, S2 = 117 MYA, S3 = 116 MYA, S4 = 42.8 MYA and S5 = 33 MYA. If the empirical μ was not available for a given species, we assigned it the μ of its closest available relative. Given that dS_{exp} is based on μ values estimated for current populations, the ratio dS/dS_{exp} is informative in relation to the extent of change in the mutation rate experienced throughout the 181 Myr period of stratum evolution, relative to the current empirical mutation rate.

Ancestral Y sequence order reconstruction in S4 and S5

X chromosome conservation was confirmed with lastZ alignment among multiple primates (Supplementary Notes). For eight species with high-level assemblies on both X and Y (*Nycticebus pygmaeus*, *Callithrix jacchus*, *Macaca mulatta*, *Rhinopithecus roxellana*, *Colobus guereza*, *Hylobates pileatus*, *Pan troglodytes* and *Homo sapiens*), pairwise alignments between X and Y sequences were performed using lastZ with the same parameter set as above. Dot plots were produced using custom Python scripts for reciprocal best MAF results. To reconstruct the ancestral sequence order, the Y assemblies in these eight species were aligned to human X with lastZ using the same parameters mentioned above. PAR sequences were added to the Y if PAR was linked with nonPARX in the assembly. Conserved segments under 50 kb or 30 kb were then extracted with DESCRAMBLER (git commit c52d775), and the reconstruction was performed using the MLGO web server (http://www.geneorder.org/server.php)[87], which allows for segment gains/losses during evolution, with tree structure: '((NYCPYG,(CALJAC,((MACMUL,(RHIROX,COLGUE)),(HYLPIL,(PANTRO,HOMSAP))))),HOMSAPX);'. The initial reconstruction at *Simiiformes* MRCA contained SVs involving both S4 and S5, which contradicted the argument that S5 evolved after the Platyrrhini–Catarrhini split. Thus, we manually curated the reconstruction at *Simiiformes* MRCA. PAR segments of MACMUL were also curated in order to be included. Conserved blocks in both parental and child nodes were retained when we inferred the evolutionary events between each of the two closest nodes with GRIMM (v2.01) (ref. [88]) by manual curation.

Chromatin configuration evolution analysis

Hi-C reads were sampled into ~30X for *Nycticebus pygmaeus*, *Callithrix jacchus*, *Macaca mulatta*, *Rhinopithecus roxellana*, *Colobus guereza*, *Hylobates pileatus*, *Homo sapiens* and *Saguinus midas*, and mapped to the corresponding genome with juicer (v1.6). The resulting 'merged_nodups.txt' were used to generate Hi-C contact maps under different resolutions (10 kb, 20 kb, 50 kb, 100 kb, 200 kb, 500 kb, 1 Mb, 2 Mb and 5 Mb) with cooler (v0.8.11) (ref. [89]), converted to h5 format and normalized using the KR method with hicexplorer (v3.7.2) (ref. [90]). Regional Hi-C maps were visualized by hicPlotMatrix in the hicexplorer package.

To obtain homologous blocks between X and Y within a species, Y-linked sequences were aligned to X-linked sequences with LastZ with the same parameter set as above. Reciprocal best results were obtained according to the guidance provided at http://genomewiki.ucsc.edu/index.php/HowTo:_Syntenic_Net_or_Reciprocal_Best. X/Y homologous blocks within S4 and S5 were obtained on the basis of the reciprocal NET result by 'netFilter' from kent package (v351). We required homologous blocks of ≥500 kb in both X and Y, and a length ratio of X and Y syntenic blocks between 0.2 and 5. Filtered homologous block pairs were manually examined with MAF alignment to exclude false positives. To evaluate the similarity of the Hi-C contact maps between homologous blocks, chess (v0.3.7) (ref. [43]) was used to calculate the structural similarity scores (SSIM) and signal-to-noise ratio (SN). When examining the changes between X and Y, we used the 25 kb diploid human autosomal Hi-C map generated by ref. [91]. Chess was run in 1 Mb windows as a control, given that the median length of the X/Y homologue blocks is about 1.5 Mb. After excluding blocks with low SN (estimated as 10% percentile of the background), *P* values were calculated on the basis of the background SSIM distribution and adjusted with FDR. When examining the changes from PAR to nonPARX, we used the regions on the X chromosome that had diverged ancestrally as the control. Hi-C maps were at 20 kb resolution and the synteny block was required to be ≥500 kb. Chess was run between each primate and *Nycticebus pygmaeus* or treeshrew. After excluding blocks with low SN, a Wilcoxon rank-sum test was used to test for significant SSIM differences between ancestral PAR (Simiiformes S4, S5 and PAR) and ancestral SDR that evolved before the primate MRCA. TAD boundaries were identified with hicexplorer in 50 kb resolution Hi-C map under 0.01 FDR-corrected *P* value cut-off, and visualized with hicPlotTADs. We determined the conservation of a boundary following a similar method to that in ref. [92]. Specifically, TAD boundaries were lifted over to the coordinates in *Nycticebus pygmaeus* (or treeshrew) with CrossMap (v0.5.4) (ref. [93]). If the lifted boundary was within a 100 kb (two bins) region to a *Nycticebus pygmaeus* (or treeshrew) TAD boundary, the TAD boundary was defined as being conserved between the two species.

Y gene evolutionary analysis

All-versus-all BLASTP was performed with Y gene protein sequences from primates and mouse using BLAST+ (v2.2.31) under an 1×10^{-5} e-value cut-off and the gene family clustering was performed with mcl (v14-137) (ref. [94]) (Supplementary Notes). We manually examined the mcl output and curated the clusters based on the gene name and the pairwise BLASTP score. Y-linked genes recovered from raw reads and Y-linked gene states (presence/pseudo/absence) from refs. [7,11,21,95] were used to curate a presence/absence matrix. When drawing inferences as to whether a Y-linked gene was present in each primate ancestor, we first converted the gene number into a presence/absence matrix and inferred with Count (v10.04) (ref. [96]). Pseudogenes were treated as 'present' before inference, and converted back into 'pseudo' after Count inference. The X-degenerate Y-linked genes were inferred with Dollo parsimony, whereas the other Y-linked genes were inferred with Wagner parsimony. We also considered the evolutionary history of S5 when inferring the ancestral states of S5 genes. Gene loss was visualized with the 'ete3' Python package (v3.1.2) (ref. [97]).

Primate species' life history traits, including the age of male sexual maturity, male-versus-female body mass dimorphism, testis mass and morphology of sperm (sperm total length and head length) were collected from previous studies and four databases: AnAge[98], the Animal Diversity Web (https://animaldiversity.org/), All the World's Primates[63] and Pan-THEIRA[99] (Supplementary Data 12). Orthologous Y-linked genes were obtained from the MCL clustering result. If multiple copies were available for one species, we randomly selected one. CDS of orthologous genes were aligned with PRANK. The alignment was further cleaned by Gblock (v0.91b)[100] with default parameters and estimate *dN* and *dS* values for each branch with codeml (paml package[101]) under the free-ratios model. Root-to-tip *dN*, *dS* and *dN*/*dS* values were calculated. Only species with root-to-tip *dN*/*dS* ratios between 0.001 and 5 were retained for downstream analysis. To test the potential relationships between gene evolution (root-to-tip

dN/*dS*) and life history traits, the R package 'caper' (v1.0.1) was used to perform phylogenetic generalized least squares (PGLS) analysis between root-to-tip *dN*/*dS* and each trait. At least eight data points were required. Sperm total length, sperm head length and testis size were first scaled with male body mass, then log-transformed. The age of male sexual maturity was log-transformed. Outlier values that fell outside of the interquartile range were excluded and PGLS was re-run to confirm the regression significance. We required P values <0.05, and plotted the transformed traits and *dN*/*dS* with the regression lines excluding the outliers.

AG analysis

If >50% of a gene overlapped with the ampliconic region, we considered it to be a candidate AG. Sequences of candidate AGs, including CDS and introns, were BLASTNed to the ampliconic region again. Only genes with >99% identity and >80% alignment rate were retained. The BLASTed regions were merged with bedtools to remove redundancy. Candidate AGs that hit against only themselves were removed. To obtain the gene orthologous relationships between species, we clustered the X-linked or Y-linked genes using the same mcl workflow described in 'Y gene evolutionary analysis'. Families that contained AGs in at least one species were retained for evolutionary analysis. For each family, if multiple genes were found in the same species but were not defined as ampliconic, these genes were classified as multiple copy. We calculated the number of gene families that contained both AGs and multi-copy genes, contained only AGs, contained only multi-copy genes, and contained neither AGs nor multi-copy genes. The resulting contingency table (Supplementary Data 16) was used for a Fisher's exact test.

Reporting summary

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

Data availability

Primate long- and short-read sequencing data were obtained from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) Database (https://www.ncbi.nlm.nih.gov/sra/) under accession code PRJNA785018, PRJNA658635 and PRJEB49549, and the GSA database with project no. PRJCA003786. Sequencing data and curated assemblies used in this study have been deposited in the NCBI Assembly Database (https://www.ncbi.nlm.nih.gov/assembly/) under accession code PRJNA790674 and the CNGB Sequence Archive (CNSA) of China National GeneBank DataBase (CNGBdb) with accession number CNP0002500. Human diploid Hi-C mapping data were obtained from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525. Other data are available in the main text or the supplementary data. Analytical data have been deposited into figshare with the link https://doi.org/10.6084/m9.figshare.20115467.v1.

Code availability

Custom scripts are available at: https://github.com/zy041225/primate_sex_chromosome.

References

1. Bachtrog, D. The Y chromosome as a battleground for intragenomic conflict. *Trends Genet.* **36**, 510–522 (2020).
2. Kitano, J. et al. A role for a neo-sex chromosome in stickleback speciation. *Nature* **461**, 1079–1083 (2009).
3. Bachtrog, D. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nat. Rev. Genet.* **14**, 113–124 (2013).
4. Wright, A. E., Dean, R., Zimmer, F. & Mank, J. E. How to make a sex chromosome. *Nat. Commun.* **7**, 1–8 (2016).
5. Vicoso, B. & Charlesworth, B. Evolution on the X chromosome: unusual patterns and processes. *Nat. Rev. Genet.* **7**, 645–653 (2006).
6. Ohno, S. *Sex Chromosomes and Sex-Linked Genes* Vol. 1 (Springer Science & Business Media, 2013).
7. Cortez, D. et al. Origins and functional evolution of Y chromosomes across mammals. *Nature* **508**, 488–493 (2014).
8. Skaletsky, H. et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825–837 (2003).
9. Ross, M. T. et al. The DNA sequence of the human X chromosome. *Nature* **434**, 325–337 (2005).
10. Lahn, B. T. & Page, D. C. Four evolutionary strata on the human X chromosome. *Science* **286**, 964–967 (1999).
11. Bellott, D. W. et al. Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* **508**, 494–499 (2014).
12. Rozen, S. et al. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* **423**, 873–876 (2003).
13. Hughes, J. F. & Page, D. C. The biology and evolution of mammalian Y chromosomes. *Annu Rev. Genet* **49**, 507–527 (2015).
14. Trombetta, B., D'Atanasio, E. & Cruciani, F. Patterns of inter-chromosomal gene conversion on the male-specific region of the human Y chromosome. *Front. Genet.* **8**, 54 (2017).
15. Tomaszkiewicz, M., Medvedev, P. & Makova, K. D. Y and W chromosome assemblies: approaches and discoveries. *Trends Genet.* **33**, 266–282 (2017).
16. Hughes, J. F. et al. Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* **483**, 82–86 (2012).
17. Hughes, J. F. et al. Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* **463**, 536–539 (2010).
18. Hallast, P. & Jobling, M. A. The Y chromosomes of the great apes. *Hum. Genet.* **136**, 511–528 (2017).
19. Tomaszkiewicz, M. et al. A time-and cost-effective strategy to sequence mammalian Y chromosomes: an application to the de novo assembly of gorilla Y. *Genome Res.* **26**, 530–540 (2016).
20. Shao., Y. et al. Phylogenomic analyses provide insights into primate genomic and phenotypic evolution. *Submitted* (2021).
21. Cechova, M. et al. Dynamic evolution of great ape Y chromosomes. *Proc. Natl Acad. Sci. USA* **117**, 26273–26280 (2020).
22. Soh, Y. Q. et al. Sequencing the mouse Y chromosome reveals convergent gene acquisition and amplification on both sex chromosomes. *Cell* **159**, 800–813 (2014).
23. Bostock, C., Gosden, J. & Mitchell, A. Localisation of a male-specific DNA fragment to a sub-region of the human Y chromosome. *Nature* **272**, 324–328 (1978).
24. Zhou, Q. et al. Complex evolutionary trajectories of sex chromosomes across bird taxa. *Science* **346**, 1246338 (2014).
25. Yang, C. et al. Evolutionary and biomedical insights from a marmoset diploid genome assembly. *Nature* **594**, 227–233 (2021).
26. Fan, Y. et al. Chromosomal level assembly and population sequencing of the Chinese tree shrew genome. *Zool. Res.* **40**, 506 (2019).
27. Ye, M. S. et al. Comprehensive annotation of the Chinese tree shrew genome by large-scale RNA sequencing and long-read isoform sequencing. *Zool. Res* **42**, 692–709 (2021).
28. Charlesworth, D., Charlesworth, B. & Marais, G. Steps in the evolution of heteromorphic sex chromosomes. *Heredity* **95**, 118–128 (2005).

29. Otto, S. P. et al. About PAR: the distinct evolutionary dynamics of the pseudoautosomal region. *Trends Genet.* **27**, 358–367 (2011).
30. Ellis, N., Yen, P., Neiswanger, K., Shapiro, L. J. & Goodfellow, P. N. Evolution of the pseudoautosomal boundary in Old World monkeys and great apes. *Cell* **63**, 977–986 (1990).
31. Charchar, F. J. et al. Complex events in the evolution of the human pseudoautosomal region 2 (PAR2). *Genome Res.* **13**, 281–286 (2003).
32. Weiss, J. et al. Sox3 is required for gonadal function, but not sex determination, in males and females. *Mol. Cell. Biol.* **23**, 8084–8091 (2003).
33. Carmignac, D. et al. SOX3 is required during the formation of the hypothalamo-pituitary axis. *Nat. Genet.* **36**, 247–255 (2004).
34. Berta, P. et al. Genetic evidence equating SRY and the testis-determining factor. *Nature* **348**, 448–450 (1990).
35. Koopman, P., Gubbay, J., Vivian, N., Goodfellow, P. & Lovell-Badge, R. Male development of chromosomally female mice transgenic for Sry. *Nature* **351**, 117–121 (1991).
36. Lahn, B. T. & Page, D. C. A human sex-chromosomal gene family expressed in male germ cells and encoding variably charged proteins. *Hum. Mol. Genet.* **9**, 311–319 (2000).
37. Trombetta, B., Cruciani, F., Underhill, P. A., Sellitto, D. & Scozzari, R. Footprints of X-to-Y gene conversion in recent human evolution. *Mol. Biol. Evol.* **27**, 714–725 (2010).
38. Huang, S., Li, Q., Alberts, I. & Li, X. PRKX, a novel cAMP-dependent protein kinase member, plays an important role in development. *J. Cell. Biochem.* **117**, 566–573 (2016).
39. Rosser, Z. H., Balaresque, P. & Jobling, M. A. Gene conversion between the X chromosome and the male-specific region of the Y chromosome at a translocation hotspot. *Am. J. Hum. Genet.* **85**, 130–134 (2009).
40. Cruciani, F., Trombetta, B., Macaulay, V. & Scozzari, R. About the X-to-Y gene conversion rate. *Am. J. Hum. Genet.* **86**, 495–497 (2010).
41. Moorjani, P., Amorim, C. E. G., Arndt, P. F. & Przeworski, M. Variation in the molecular clock of primates. *Proc. Natl Acad. Sci. USA* **113**, 10607–10612 (2016).
42. Chintalapati, M. & Moorjani, P. Evolution of the mutation rate across primates. *Curr. Opin. Genet Dev.* **62**, 58–64 (2020).
43. Galan, S. et al. CHESS enables quantitative comparison of chromatin contact data and automatic feature extraction. *Nat. Genet.* **52**, 1247–1255 (2020).
44. Liu, J. et al. A new emu genome illuminates the evolution of genome configuration and nuclear architecture of avian chromosomes. *Genome Res.* **31**, 497–511 (2021).
45. Xue, L. et al. Telomere-to-telomere assembly of a fish Y chromosome reveals the origin of a young sex chromosome pair. *Genome Biol.* **22**, 203 (2021).
46. Bachtrog, D. The temporal dynamics of processes underlying Y chromosome degeneration. *Genetics* **179**, 1513–1525 (2008).
47. Nguyen, T. A. et al. A cluster of autism-associated variants on X-Linked NLGN4X functionally resemble NLGN4Y. *Neuron* **106**, 759–768 e757 (2020).
48. Kappeler, P. M. & Van Schaik, C. P. *Sexual Selection in Primates: New and Comparative Perspectives* (Cambridge Univ. Press, 2004).
49. Roldan, E. & Gomendio, M. The Y chromosome as a battle ground for sexual selection. *Trends Ecol. Evol.* **14**, 58–62 (1999).
50. Williams, T. M. & Carroll, S. B. Genetic and molecular insights into the development and evolution of sexual dimorphism. *Nat. Rev. Genet.* **10**, 797–804 (2009).
51. Bhowmick, B. K., Satta, Y. & Takahata, N. The origin and evolution of human ampliconic gene families and ampliconic structure. *Genome Res.* **17**, 441–450 (2007).
52. Dorus, S., Gilbert, S. L., Forster, M. L., Barndt, R. J. & Lahn, B. T. The CDY-related gene family: coordinated evolution in copy number, expression profile and protein sequence. *Hum. Mol. Genet.* **12**, 1643–1650 (2003).
53. Seboun, E. et al. Gene sequence, localization, and evolutionary conservation of DAZLA, a candidate male sterility gene. *Genomics* **41**, 227–235 (1997).
54. Mueller, J. L. et al. Independent specialization of the human and mouse X chromosomes for the male germ line. *Nat. Genet.* **45**, 1083–1087 (2013).
55. Bachtrog, D., Mahajan, S. & Bracewell, R. Massive gene amplification on a recently formed *Drosophila* Y chromosome. *Nat. Ecol. Evol.* **3**, 1587–1597 (2019).
56. Lahn, B. T., Pearson, N. M. & Jegalian, K. The human Y chromosome, in the light of evolution. *Nat. Rev. Genet.* **2**, 207–216 (2001).
57. Schaller, F. et al. Y chromosomal variation tracks the evolution of mating systems in chimpanzee and bonobo. *PLoS ONE* https://doi.org/10.1371/journal.pone.0012482 (2010).
58. Hughes, J. F. et al. Sequence analysis in Bos taurus reveals pervasiveness of X-Y arms races in mammalian lineages. *Genome Res.* **30**, 1716–1726 (2020).
59. Jl, W. potts pr. The maGe protein family and cancer. *Curr. Opin. Cell Biol.* **37**, 1–8 (2015).
60. Nei, M., Xu, P. & Glazko, G. Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. *Proc. Natl Acad. Sci. USA* **98**, 2497–2502 (2001).
61. Krausz, C., Giachini, C. & Forti, G. TSPY and male fertility. *Genes* **1**, 308–316 (2010).
62. Foresta, C., Ferlin, A. & Moro, E. Deletion and expression analysis of AZFa genes on the human Y chromosome revealed a major role for DBY in male infertility. *Hum. Mol. Genet.* **9**, 1161–1169 (2000).
63. Rowe, N. & Myers, M. *All the World's Primates* (Pogonias Press, 2016).
64. Yu, Y.-H., Lin, Y.-W., Yu, J.-F., Schempp, W. & Yen, P. H. Evolution of the DAZ gene and the AZFc region on primate Y chromosomes. *BMC Evol. Biol.* **8**, 1–10 (2008).
65. Plavcan, J. M. Sexual dimorphism in primate evolution. *Am. J. Phys. Anthropol.* **116**, 25–53 (2001).
66. Liu, W.-S. Mammalian sex chromosome structure, gene content, and function in male fertility. *Annu. Rev. Anim. Biosci.* **7**, 103–124 (2019).
67. Wilson, M. A. The Y chromosome and its impact on health and disease. *Hum. Mol. Genet.* **30**, R296–R300 (2021).
68. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1303.3997 (2013).
69. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
70. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
71. Wickham, H. Elegant graphics for data analysis. *Media* **35**, 10.1007 (2009).
72. Vollger, M. R. et al. Long-read sequence and assembly of segmental duplications. *Nat. Methods* **16**, 88–94 (2019).
73. Rhie, A. et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).
74. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
75. Kahlke, T. & Ralph, P. J. BASTA—taxonomic classification of sequences and sequence bins using last common ancestor estimations. *Methods Ecol. Evol.* **10**, 100–103 (2019).
76. Birney, E., Clamp, M. & Durbin, R. GeneWise and genomewise. *Genome Res.* **14**, 988–995 (2004).

77. Hoff, K. J. & Stanke, M. Predicting genes in single genomes with AUGUSTUS. *Curr. Protoc. Bioinformatics* **65**, e57 (2019).
78. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
79. Lopez-Delisle, L. et al. pyGenomeTracks: reproducible plots for multivariate genomic data sets. *Bioinformatics* (2021).
80. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
81. Harris, R. S. *Improved Pairwise Alignment of Genomic DNA* (The Pennsylvania State University, 2007).
82. Loytynoja, A. Phylogeny-aware alignment with PRANK. *Methods Mol. Biol.* **1079**, 155–170 (2014).
83. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
84. Zhang, Z. et al. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics* **4**, 259–263 (2006).
85. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
86. Wu, F. & Przeworski, M. A paternal bias in germline mutation is widespread in amniotes and can arise independently of cell division numbers. *eLife* **11**, e80008–e80008 (2022).
87. Hu, F., Lin, Y. & Tang, J. MLGO: phylogeny reconstruction and ancestral inference from gene-order data. *BMC Bioinformatics* **15**, 354 (2014).
88. Tesler, G. GRIMM: genome rearrangements web server. *Bioinformatics* **18**, 492–493 (2002).
89. Abdennur, N. & Mirny, L. A. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* **36**, 311–316 (2020).
90. Ramirez, F. et al. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat. Commun.* **9**, 189 (2018).
91. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
92. Luo, X. et al. 3D genome of macaque fetal brain reveals evolutionary innovations during primate corticogenesis. *Cell* **184**, 723–740. e721 (2021).
93. Zhao, H. et al. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006–1007 (2014).
94. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
95. Martínez-Pacheco, M. et al. Expression evolution of ancestral XY gametologs across all major groups of placental mammals. *Genome Biol. Evol.* **12**, 2015–2028 (2020).
96. Csuros, M. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* **26**, 1910–1912 (2010).
97. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
98. de Magalhães, J. P. & Costa, J. A database of vertebrate longevity records and their relation to other life-history traits. *J. Evol. Biol.* **22**, 1770–1774 (2009).
99. Jones, K. E. et al. PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology* **90**, 2648–2648 (2009).
100. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552 (2000).
101. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

Acknowledgements

We thank all supervisors, collaborators and anyone else involved with the collection and processing of the primary datasets. We thank China National GeneBank for providing the computational resources. This study was supported by grants from Strategic Priority Research Program of the Chinese Academy of Sciences (XDB31020000 to G.Z.), International Partnership Program of Chinese Academy of Sciences (no. 152453KYSB20170002 to G.Z.), Villum Investigator Grant (no. 25900 to G.Z.), National Natural Science Foundation of China (31822048 to D.-D.W.), Yunnan Fundamental Research Project (2019FI010 to D.-D.W.) and The Animal Branch of the Germplasm Bank of Wild Species of Chinese Academy of Science (the Large Research Infrastructure Funding to D.-D.W.).

Author contributions

G.Z. conceived the project. D.-D.W., H.K., T.H., Y.-G.Y., La.Z., X.Q., L.K. and T.M.-B. coordinated and were involved in sample collection, extraction and sequencing. Y.Z., X.Z., J.J., Lo.Z., J.B., X.L., M.M.C.R., M.R.B., M.F., J.C. and Q.F. performed the analyses. G.Z., M.H.S. and H.Y. supervised the project. G.Z., D.N.C., M.H.S., Y.Z., M.R.B., J.B., M.M.C.R., Y.-G.Y. and T.M.-B. wrote the manuscript with input from all the authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41559-022-01974-x.

Correspondence and requests for materials should be addressed to Guojie Zhang.

Peer review information *Nature Ecology & Evolution* thanks Qi Zhou and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

[1]BGI-Shenzhen, Shenzhen, China. [2]Centre for Evolutionary & Organismal Biology, and Women's Hospital, Zhejiang University School of Medicine, Hangzhou, China. [3]Liangzhu Laboratory, Zhejiang University Medical Center, Hangzhou, China. [4]Section for Ecoinformatics & Biodiversity, Department of Biology, Aarhus University, Aarhus C., Denmark. [5]Bioinformatics Research Centre, Aarhus University, Aarhus C., Denmark. [6]College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China. [7]College of Life Sciences, Northwest University, Xi'an, China. [8]Institute of Evolutionary Biology (UPF-CSIC), PRBB, Barcelona, Spain. [9]Catalan Institution of Research and Advanced Studies (ICREA), Barcelona, Spain. [10]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. [11]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Barcelona, Spain. [12]Graduate School of Environmental Science, Hokkaido University, Sapporo, Japan. [13]Faculty of Environmental Earth Science, Hokkaido University, Sapporo, Japan. [14]Japan Monkey Centre, Inuyama, Japan. [15]Kunming College of Life Science, University of the Chinese Academy of Sciences, Kunming, China. [16]Key Laboratory of Animal Models and Human Disease Mechanisms of Chinese Academy of Sciences & Yunnan Province, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China. [17]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China. [18]National Resource Center for Non-Human Primates, Kunming Primate Research Center, and National Research Facility for Phenotypic & Genetic Analysis of Model Animals (Primate Facility), Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China. [19]KIZ-CUHK Joint Laboratory of Bioresources and Molecular Research in Common Diseases, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China. [20]James D. Watson Institute of Genome Sciences, Hangzhou, China. [21]Guangdong Provincial Academician Workstation of BGI Synthetic Genomics, BGI-Shenzhen, Shenzhen, China. [22]The Cancer Hospital of the University of Chinese Academy of Sciences (Zhejiang Cancer Hospital), Institute of Basic Medicine and Cancer (IBMC), Chinese Academy of Sciences, Hangzhou, China. [23]Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, UK. [24]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China. [25]Section for Ecology and Evolution, Department of Biology, University of Copenhagen, Copenhagen, Denmark. ✉e-mail: guojiezhang@zju.edu.cn

nature portfolio

Corresponding author(s): Guojie Zhang

Last updated by author(s): Nov 16, 2022

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	Data collection did not involve any software or code.
Data analysis	Common bioinformatic and statistical analysis software packages were used, including: caper (v1.0.1), Gblock (v0.91), PRANK (v150803), GeneWise (v2.4.1), Augustus (v3.2.3), bedtools (v2.29.2), blast+ (v2.2.31), solar (v0.9.6), KaKs_Calculator (v2.0), RAxML (v8.2.4), MAFFT (v7.471), iqtree (v2.1.3), bwa (v0.7.17), SDA (git commit 4ca0c0709dc86181afaeeee862543dc19a411eb3), ggplot2 (v3.3.5), samtools (v1.9), freebayes (v1.3.1-17-gaa2ace8), bcftools (v1.11), diamond (v2.0.0), BASTA (v1.3.2.3), exonerate (v2.4.0), pbmm2 (v1.2.0-1-g31b4be0), minimap2 (v2.17), juicer (v1.6), 3d-dna (v180922), straw (v0.0.8), matplotlib (v3.4.2), bowtie2 (v2.4.1), wtdbg2 (v2.3), lastz (v1.04.00), DESCRAMBLER (git commit c52d775), MLGO (http://www.geneorder.org/), GRIMM (v2.01), cooler (v0.8.11), hicexplorer (v3.7.2), liftOver (v351), netFilter (v351), chess (v0.3.7), CrossMap (v0.5.4), mcl (v14-137), Count (v10.04), ete3 (v3.1.2), hmmer (v3.1b2) , paml (v4.8)

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

Data

Policy information about availability of data

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

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

Primate long- and short-read sequencing data were obtained from National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) Database (https://www.ncbi.nlm.nih.gov/sra/) under accession code PRJNA785018, PRJNA658635 and PRJEB49549, and GSA database with project no. PRJCA003786. Sequencing data and curated assemblies in used this study have been deposited into National Center for Biotechnology Information (NCBI) Assembly Database (https://www.ncbi.nlm.nih.gov/assembly/) under accession code PRJNA790674 and CNGB Sequence Archive (CNSA) of China National GeneBank DataBase (CNGBdb) with accession number CNP0002500. Human diploid Hi-C map was obtained from: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525. Other data are available in the main text or the supplementary materials. Analysis data have been deposited into figshare with the link: https://doi.org/10.6084/m9.figshare.20115467.v1

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	Study only uses human genome (hg38) which contains both X and Y chromosomes.
Population characteristics	Study does not involve human population data
Recruitment	Study does not recruit human participants
Ethics oversight	Study only uses human genome (hg38) which contains both X and Y chromosomes.

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

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	The determination of sample size for genome sequencing is not applied in this study. Bioinformatic analyses were performed with all available data.
Data exclusions	Bioinformatic analyses were performed with all available data.
Replication	Replication is not applied for genomes sequencing.
Randomization	Randomization for genome sequencing is not applied in this study.
Blinding	Blinding was not necessary for genome sequencing

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	Blood samples of female Macaca silenus, Papio hamadryas, Erythrocebus patas, Colobus guereza and muscle sample of female Hylobates pileatus were collected from Japan Monkey Centre, Japan.
Wild animals	Study did not involve wild animals.
Reporting on sex	Both sex genomic data are used in this study.
Field-collected samples	Study did not involve field-collected samples.
Ethics oversight	The use of genetic materials of Japan Monkey Centre was approved by the Research Ethics Committee of the Japan Monkey Centre (#2017-018) and performed in accordance with the Ethical Guidelines for Research at the Japan Monkey Centre (1 April 2016).

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

nature ecology & evolution

Article

https://doi.org/10.1038/s41559-022-01974-x

Eighty million years of rapid evolution of the primate Y chromosome

In the format provided by the authors and unedited

Supplementary Information

Supplementary Notes

Evaluation of the use of female short reads from closely related species for X- and Y-linked sequence identification

We evaluated the utility of using female short reads from closely related species to identify sex-linked sequences using the *Cebus albifrons* genome and resequencing short reads. To identify the X-linked and Y-linked sequences, we used the same male short reads dataset of *Cebus albifrons*. For the female short reads, here we utilized three sets of reads:

1. female short reads of *Cebus albifrons* (the same species)
2. female short reads of *Sapajus apella*, which diverged from *Cebus albifrons* ~4.70 million years (Myr)[1]
3. female short reads of *Pithecia pithecia*, which diverged from *Cebus albifrons* ~21.67 Myr[1]

For each male-female short reads combination, we first aligned the male and female short reads to the *Cebus albifrons* genome with BWA MEM (v0.7.17), then identified each set of X-linked and Y-linked sequences using the procedure described in "**Identification of X-linked and Y-linked sequences**" of the **Methods** section. The X-linked and Y-linked sequence datasets identified via male and female short reads both from *Cebus albifrons* were used as the gold standard. In total, we identified 144,195,170 bp nonPARX sequences and 9,560,224 bp nonPARY sequences with the male and female reads of *Cebus albifrons*.

Next, we evaluated the performance of sex-linked sequence identification using female short reads from another species. Taking the nonPARX dataset as an example, we defined true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) as:

$$TP = X_a \cap X_b$$

$$FP = (G - X_a) \cap X_b$$

$$FN = X_a \cap (G - X_b)$$

$$TN = (G - X_a) \cap (G - X_b)$$

where X_a, X_b represent the nonPARX dataset identified with female reads of *Cebus albifrons* (golden standard) and female reads of another species (*Sapajus apella* or *Pithecia pithecia*), respectively, and G represent the whole genome of *Cebus albifrons* (including autosomal, X- and Y-linked sequences). Then we calculated the true positive rate (TPR), false positive rate (FPR) and false discovery rate (FDR) in the other two datasets as

$$TPR = TP/(TP + FN)$$

$$FPR = FP/(FP + TN)$$

$$FDR = FP/(FP + TP)$$

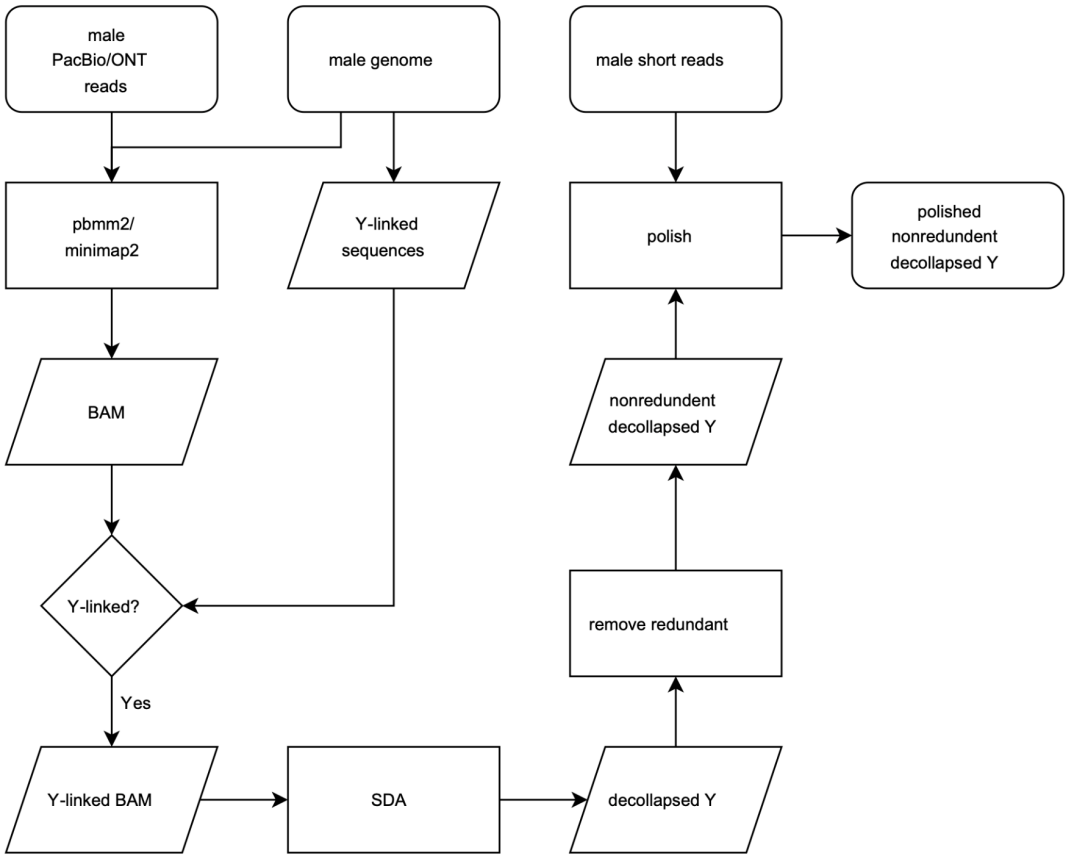
Our results (**Supplementary Table 1**) showed that, when using female short reads from species that had diverged from the target species for ~ 5 Myr, the TPR of the identified X-linked sequences was > 97%, suggesting that this would not have affected the reliability of the X-linked sequence dataset. As for the Y-linked dataset, when using female short reads from species that had diverged from the target species for about 5 Myr, the TPR was 71%. However, when more distant species that had diverged for ~20 Myr were used, the true positive rate decreased considerably for both datasets whereas the false positive rate increased. In our study, we used female reads of *C. sabaeus* and *H. hoolock* to identify the sex-linked sequences of *Chlorocebus aethiops* and *Hoolock leuconedys*, respectively. Both divergence times were smaller than 5 Myr (*C. sabaeus – C. aethiops* 2.1 Myr[1], *H. hoolock – H. leuconedys* 3.8 Myr[2]). Thus, we were able to argue that when using species from the same genus, both the identified X-linked and Y-linked sequence datasets should be sufficiently reliable for further analysis.

Supplementary Table 1. Evaluation of using closely related species in X/Y identification.

		C. albifrons – S. apella (4.7MYA)	*C. albifrons – P. pithecia* (21.67MYA)
X	TP (bp)	140,613,752	84,067,076
	TN (bp)	2,729,079,926	2,730,152,381
	FP (bp)	1,167,141	94,686
	FN (bp)	3,581,418	60,128,094
	TPR (%)	97.52	58.30
	FPR (%)	0.04	0.16
	FDR (%)	0.82	0.11
Y	TP (bp)	6,788,503	8,896,758
	TN (bp)	2,862,924,883	2,850,829,073
	FP (bp)	1,957,130	14,052,940
	FN (bp)	2,771,721	663,466
	TPR (%)	71.01	38.77
	FPR (%)	0.07	95.49
	FDR (%)	22.38	61.23

De-collapsing Y-linked sequence

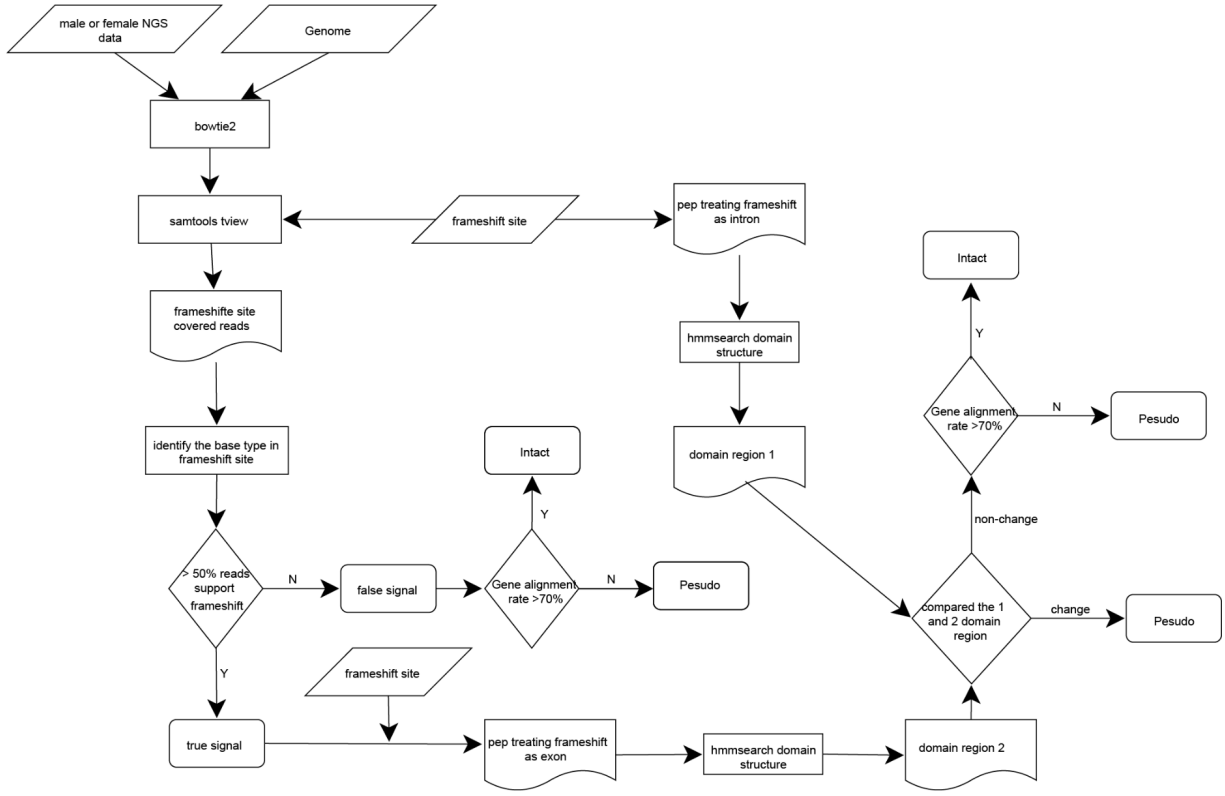
We used SDA to de-collapse Y-linked sequences in species with available male long reads (**Supplementary Note Fig. 1**). PacBio long reads were mapped using pbmm2 (v1.2.0-1-g31b4be0) with parameter "--preset SUBREAD -N 50 -l 1000". Nanopore long reads were mapped with minimap2 (v2.17) with parameter "-ax map-ont -m 1000". The BAM files of reads alignment to the Y-linked sequences were extracted and input into SDA along with the half value of the whole-genome peak depth, since there is only one copy of the Y chromosome in a male genome. The de-collapsing result was merged with the input Y-linked sequences with in-house scripts (available at: https://github.com/gf777/misc/tree/master/marmoset%20Y) to a non-redundant Y dataset. To further polish the de-collapsed sequences, we first combined the non-redundant Y-linked sequences with other (autosomal and X-linked) sequences into a new genome. Short reads of the same individual were mapped to the new genome with BWA MEM and the genome was polished with the freebayes-bcftools polishing pipeline (freebayes v1.3.1-17-gaa2ace8, bcftools v1.11) described in ref[3]. We did not perform de-collapsing with *H. sapiens*, *P. troglodytes* and *M. mulatta* since the Y sequences are rather complete, have been constructed using the SHIMS method[4-6].



Supplementary Note Figure 1. Y de-collapsing workflow.

Frameshift confirmation with short reads

We confirmed the frameshift signals by short reads, in a two-round reads mapping procedure. In the first round, male short reads were mapped to the genome with bowtie2 (v2.4.1)[7] using default parameters. To reduce the impact on mapping due to duplicated genes, we performed a second round of reads mapping. Reads mapped to the raw gene set were extracted and mapped to the genome with an additional parameter '-a' to increase efficiency. The resulting BAM generated in the second round were used to examine the frameshift signals reported by annotation. We also checked if the frameshift signals would affect the functional domains by hmmsearch (v3.1b2) and database Pfam-A.hmm (Pfam35.0)[8] obtained from http://pfam.xfam.org/. If the annotation contained no or false positive frameshifts, or if the frameshift signal does not affect the number and type of the functional domains, we would treat it as an intact gene; otherwise, the gene would be considered as a pseudogene. Check results are available at **Supplementary Data 19** & **20**.

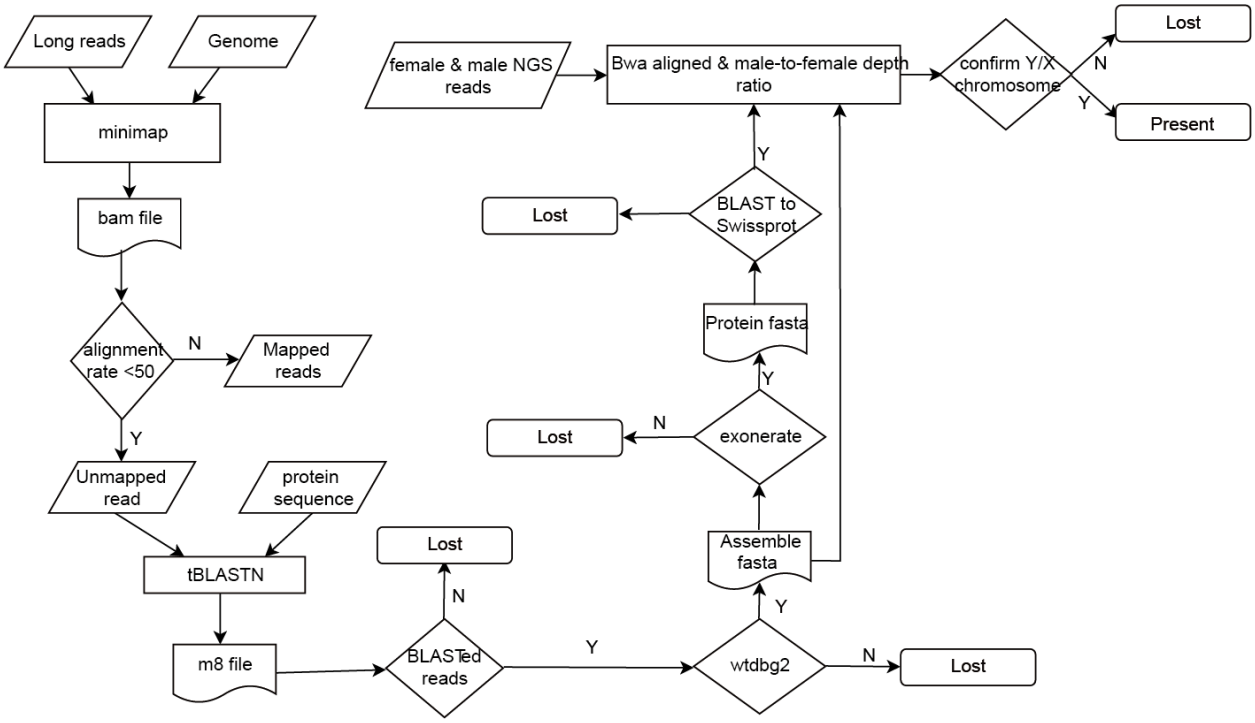


Supplementary Note Figure 2. Frameshift signal confirmation in potential X or Y pseudogenes with male short reads.

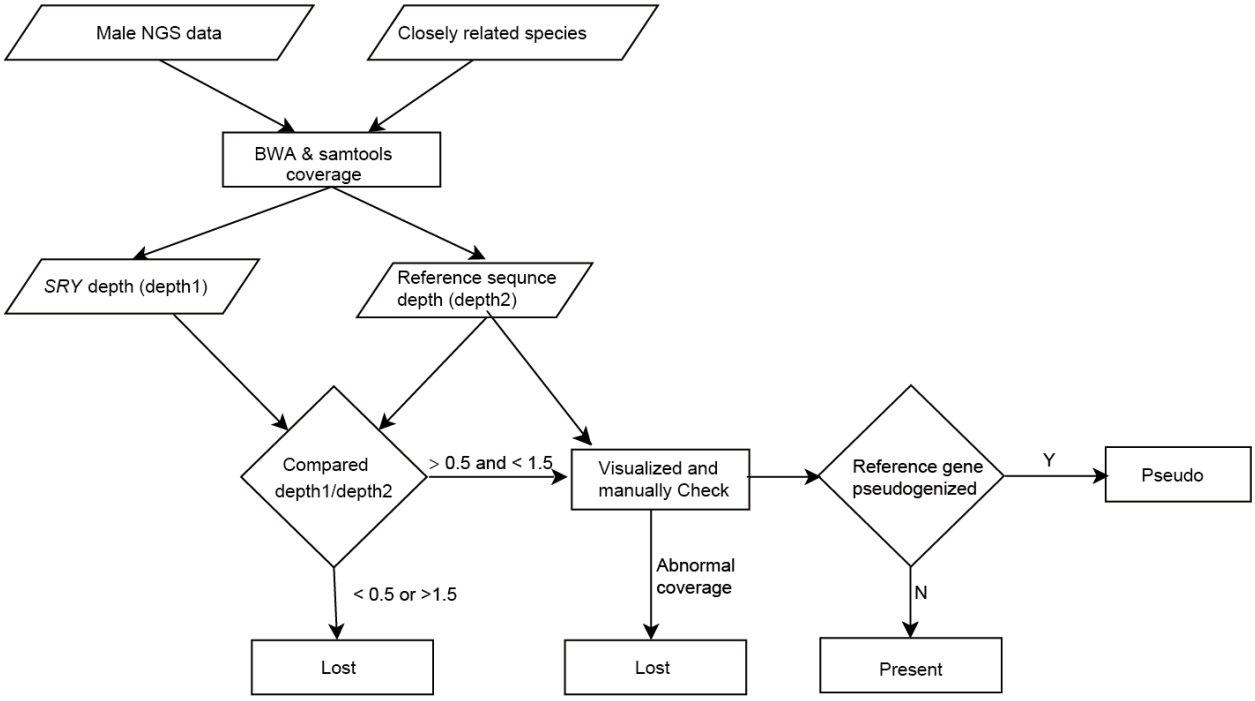
X or Y-linked gene absence confirmation

To confirm the absence of X-linked or Y-linked genes from a given species, for species that had available male long reads (**Supplementary Note Fig. 3**), we first mapped long reads to the genome with minimap2 (v2.17) and extracted the poorly aligned reads (alignment rate < 50%).

These raw long reads were then BLASTed against Y-linked genes from other primate species. The resulting BLASTed long reads were used for assembling and polishing with wtdbg2 (v2.3)[9] and genes were annotated with Exonerate (v2.4.0). We also mapped the male and female reads to the assembled sequence with bowtie2 (v2.4.1) to exclude false positives from autosomal and X-linked sequences following the same strategy as that applied above. We further BLASTPed the annotated protein sequence against the Swiss-Prot database (release date 2020/05). Only annotated entries that BLASTed against the correct record with an identity > 60% and an alignment rate > 70% were retained. Genes that did not satisfy the criteria were classified as pseudogenes. We did not consider frameshift signals here because the assembly may have been error-prone due to the high sequencing error rate in the long reads, and the low sequencing depth extracted from the poorly aligned reads may not have been enough for sequence polishing. Thus, these Y-linked genes were not used in follow-up sequence-based analyses such as pairwise *dS* and evolutionary analysis. For species where only male short reads were available, i.e., *C. sabaeus, M. leucophaeus* and *P. abelii*, we mapped the male short reads to the closely related species where a high-quality Y assembly was available (**Supplementary Note Fig. 4**). Specifically, we mapped the male short reads of *C. sabaeus* and *M. leucophaeus* to the *M. mulatta* genome and the male short reads of *P. abelii* to the *H. sapiens* genome. To confirm the loss of a certain gene, we checked the mapping depth of the exons of the gene. The sequencing depth of exons of *SRY*, which is present and intact in both species, was used as a control. The mapping depth of the exons was visualized and manually examined to confirm the gene loss. The gene was considered to be intact if the following requirements were met: 1. the average exon depth was between 0.5x and 1.5x of the depth of *SRY*, 2. the gene in the genome of the reference species was intact (i.e., contained no frameshifts & had an alignment rate > 70%). We also visualized the depth distribution along each locus with pygenometracks and manually examined the result. Check results are available at **Supplementary Data 21-23** and **Supplementary Note Fig 4-7**.

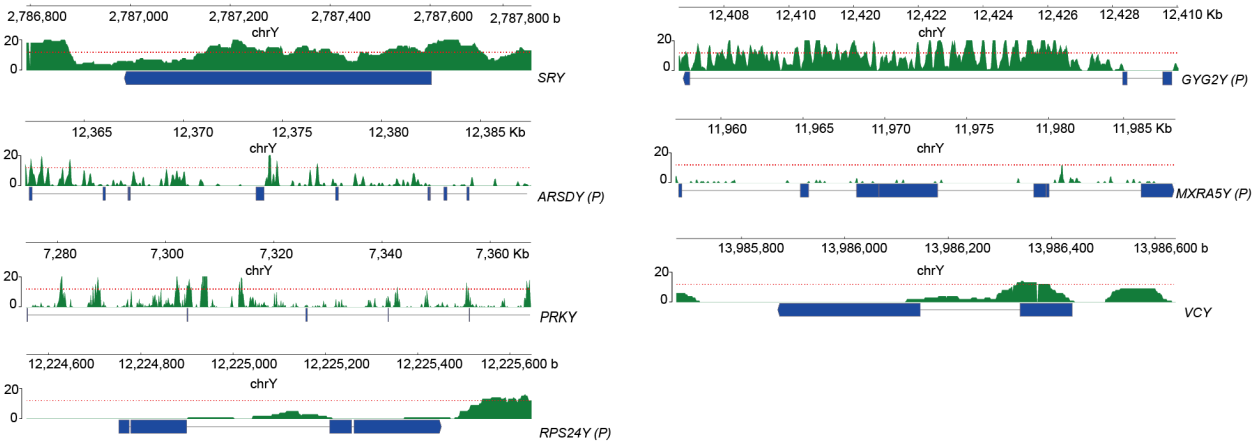


Supplementary Note Figure 3. Absent X or Y chromosome gene confirmation by long reads.

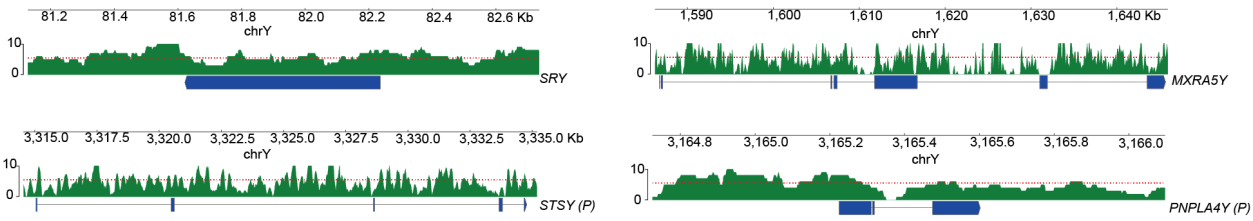


Supplementary Note Figure 4. Potential Y gene loss confirmation by short reads

Supplementary Note Figure 5. The absent Y chromosome gene of *C. sabaeus* confirmed by short reads. Exons are drawn as dark blue boxes below each gene. The red dashed line indicates the average depth of *SRY*.



Supplementary Note Figure 6. The absent Y chromosome gene of *P. abelii* confirmed by short reads. Exons are drawn as dark blue boxes below each gene. The red dashed line indicates the average depth of *SRY*.



Supplementary Note Figure 7. The absent Y chromosome gene of *M. leucophaeus* confirmed by short reads. Exons are drawn as dark blue boxes below each gene. The red dashed line indicates the average depth of *SRY*.

Assembly confirmation

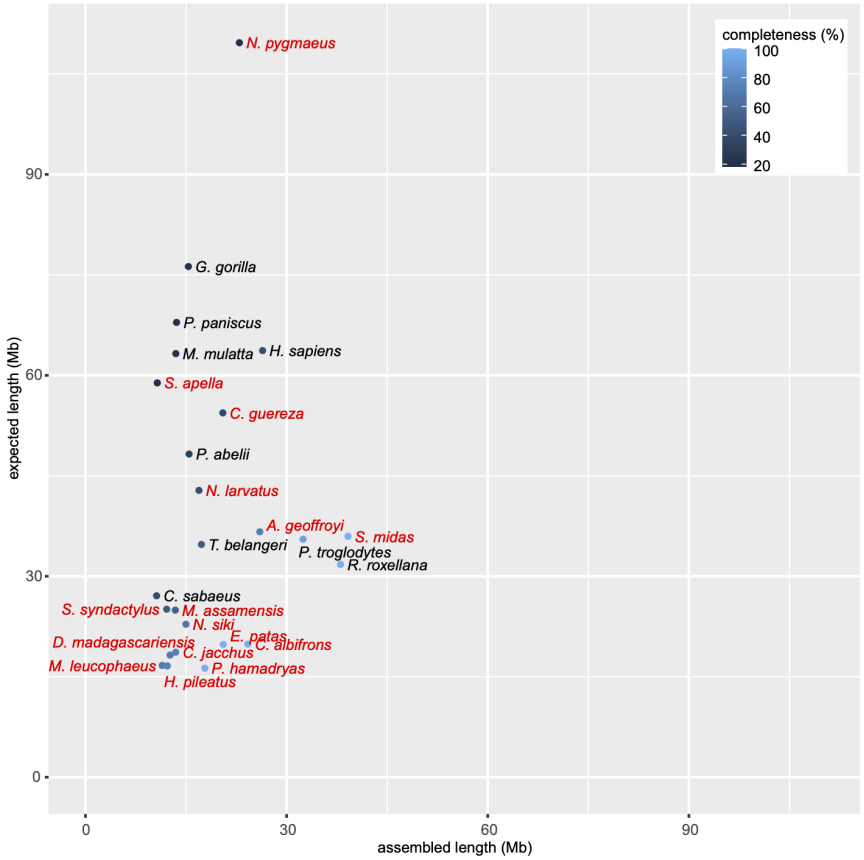
We mapped long reads, 10X linked reads and Hi-C data to the genome to confirm our assembly result. PacBio long reads were mapped using pbmm2 (v1.2.0-1-g31b4be0) with parameter "--preset SUBREAD". Nanopore long reads were mapped with minimap2 (v2.17)[10] with parameter "-ax map-ont". For 10X linked reads, barcodes were trimmed from read1 and the remaining reads were mapped to the genome with BWA MEM (v0.7.17). Hi-C matrix data were obtained via the juicer-3d-dna pipeline (v180922)[11]. Interaction strength values were obtained with straw (v0.0.8). The interaction heatmap was visualized with matplotlib (v3.4.2)[12] and the violin plot was made with ggplot2 (v3.3.5).

MCL clustering

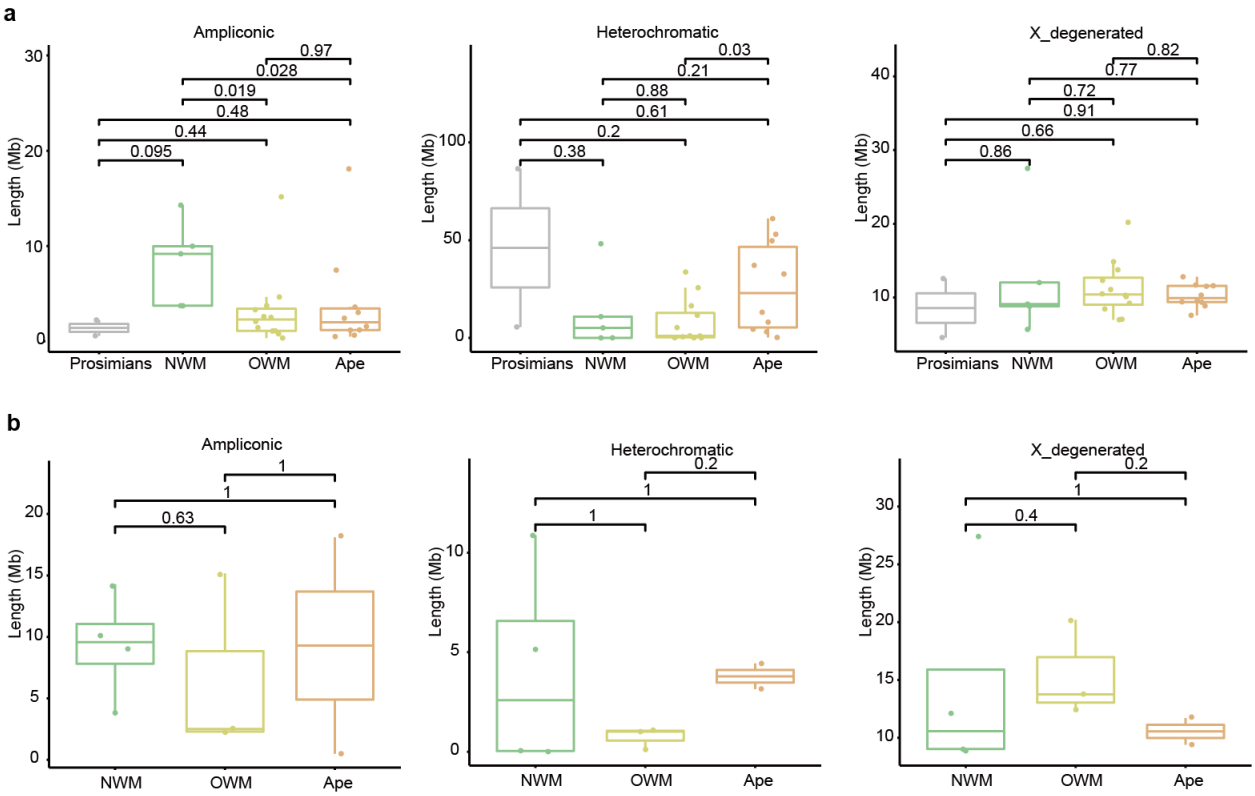
To obtain the optimal inflation value, we performed a two-step iterative process. In the first iteration, we tested the inflation values from 1.1 to 5.0 with an increment of 0.4. Parameters were selected according to previous research[13,14]. Clm info was used to determine the inflation value that captured more edges and was granular among clustering results in the input graph[15]. The optimal inflation value (2.8 in this study) was then used in the second round of clustering. During the second round of computation, inflation was iterated between 1.1 and 2.8, the optimal inflation value, with an increment of 1. The unique-function subgroups obtained from each step of the second-round computation were added to the final gene family dataset. All clustered subgroups constituted the final gene family dataset.

X chromosome conservation analysis

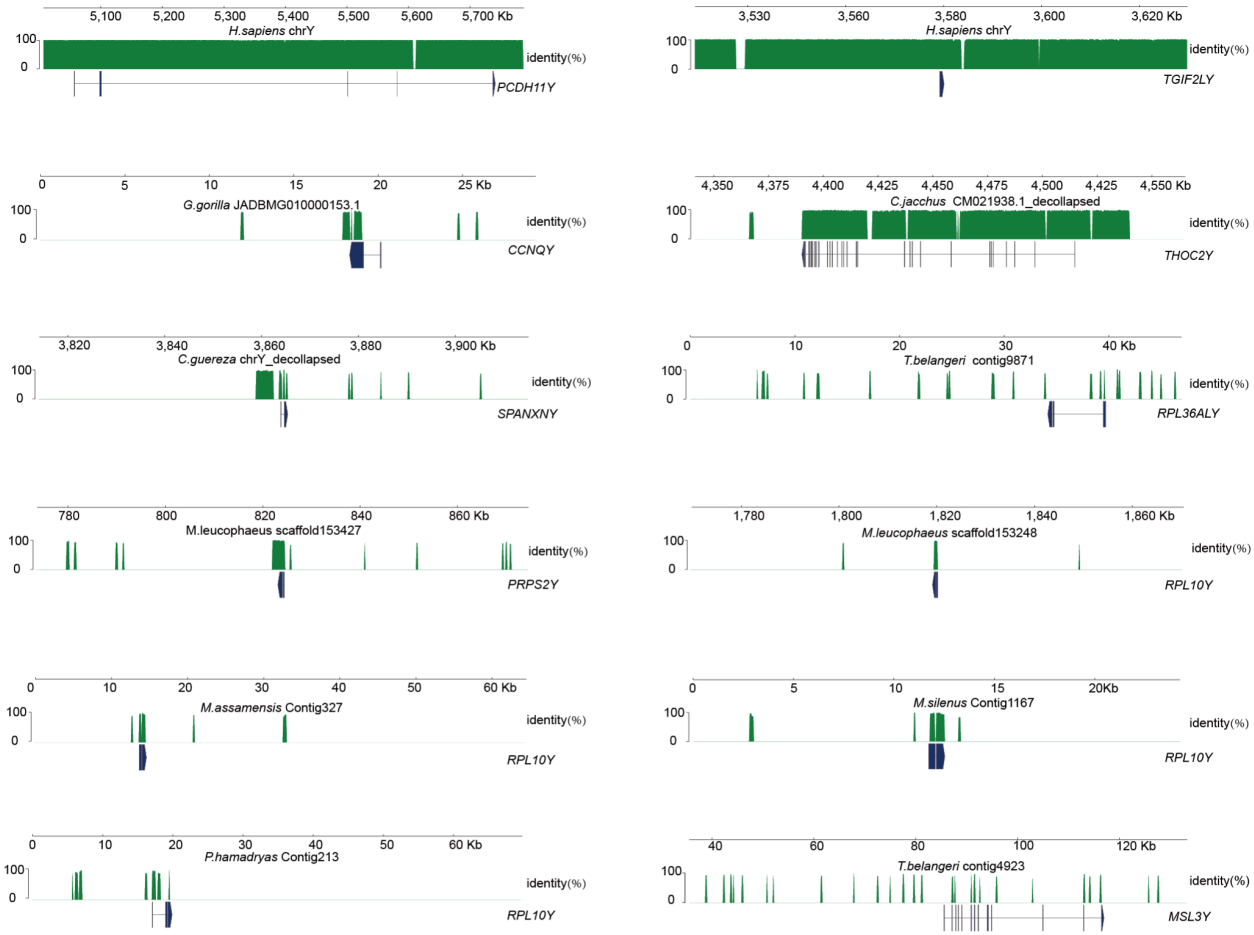
We performed X-linked sequence pairwise lastZ (v1.04.00) alignment between human and species with chromosome-level assembly of the X, with the same parameter set as above. NET and CHAIN results were then used as input for DESCRAMBLER (git commit c52d775)[16] to extract conserved segments under 100 Kb resolution. Large structural variants (SVs) of >2 Mb were confirmed with Hi-C.



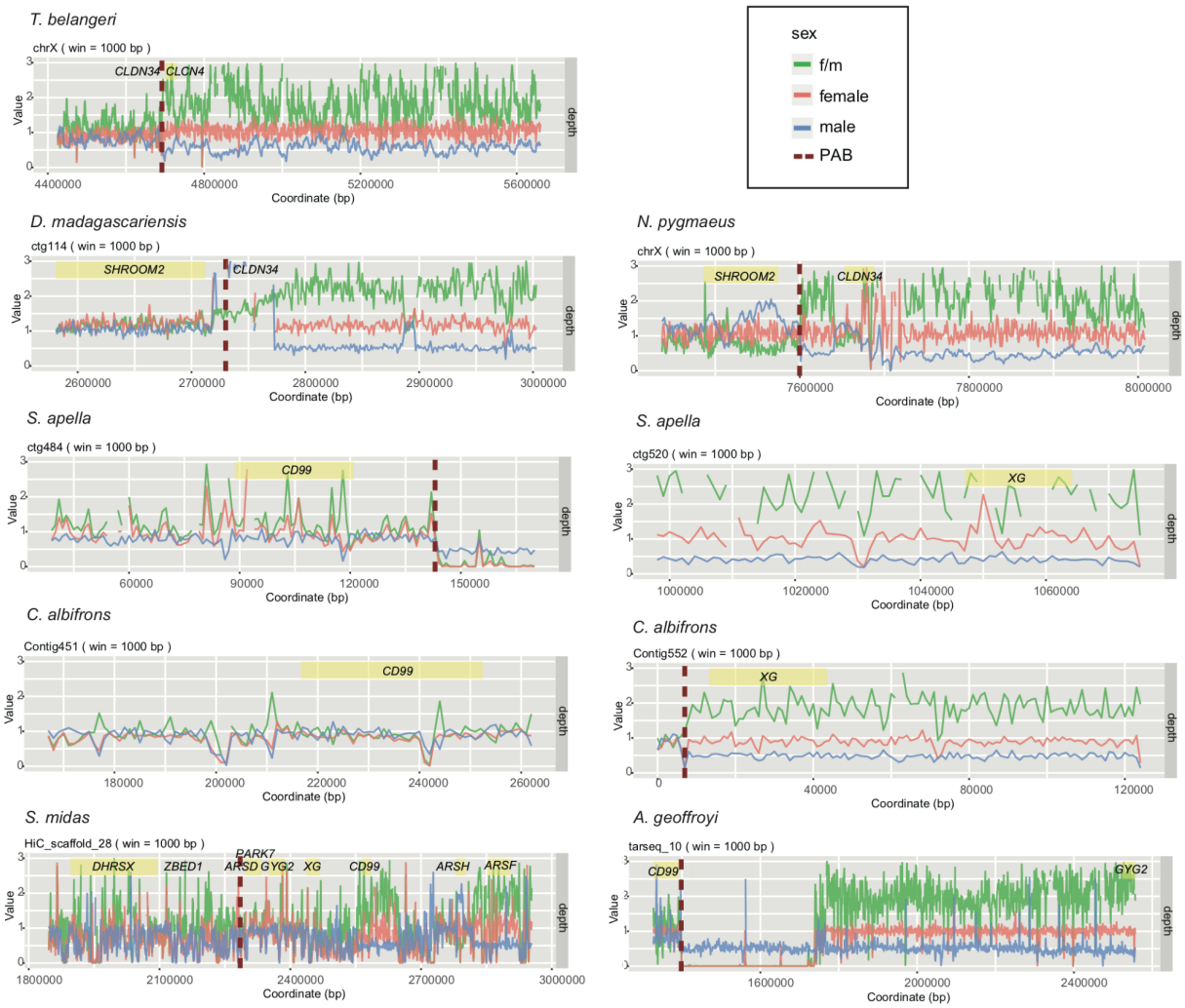
Supplementary Figure 1. Assembled Y length and completeness evaluation. The expected length was estimated based on male karyotype images. Text color: red, new assembly generated in this project; black, published assembly. Dots are color-coded to denote the estimated Y completeness. We also marked *C. jacchus* which was published recently[17] by our group, in red.



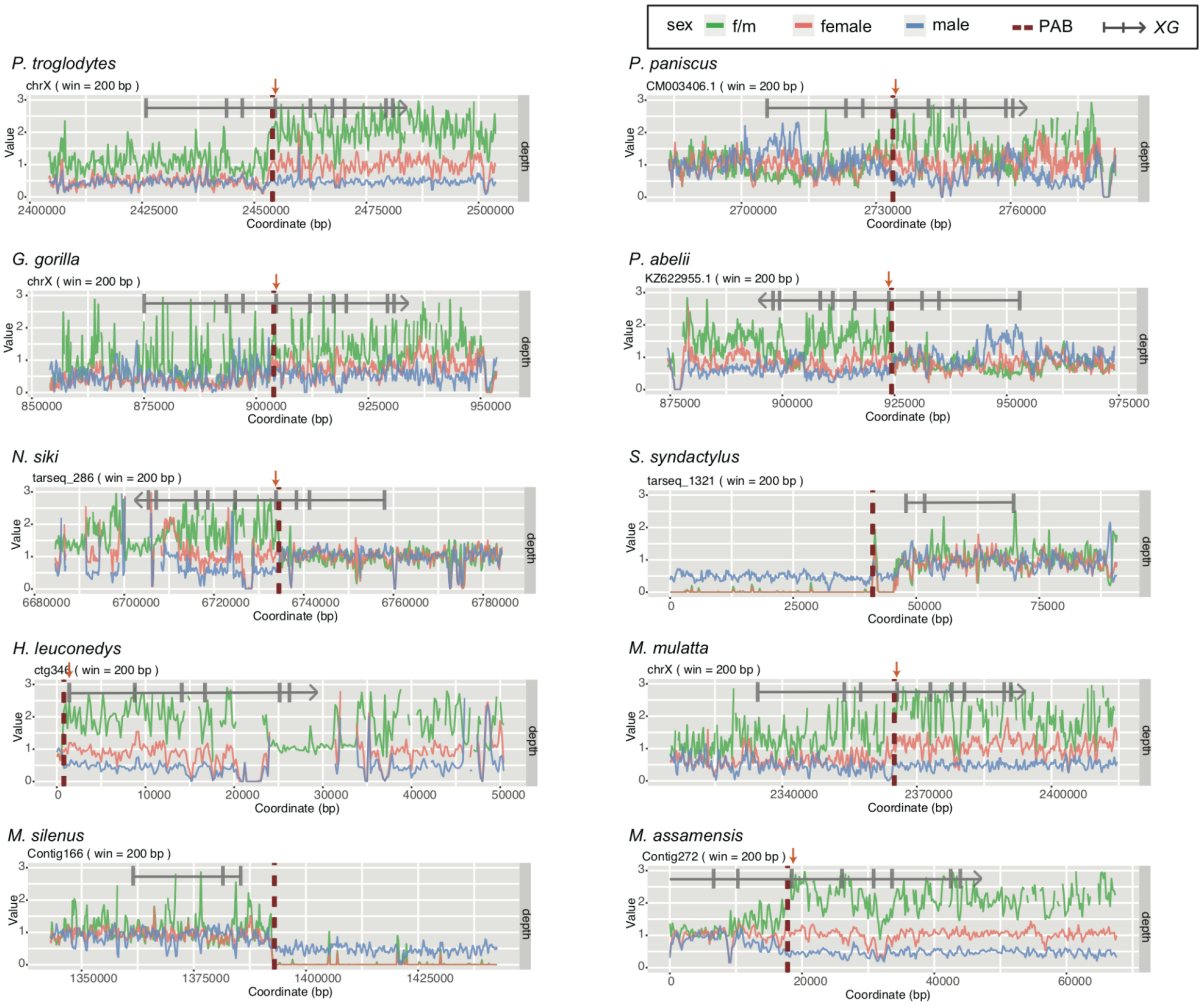
Supplementary Figure 2. Length comparison of ampliconic, heterochromatic and X degenerate regions on the Y chromosome in major clades of all primates (a). Similar patterns were observed when only primates with ≥70% Y completeness (**b**). Panel a contains all primates species in this study, including prosimians (n = 2), NWM (n = 5), apes (n = 10) and OWM (n = 12). Panel b only contains species with ≥70% Y completeness, including NWM (n = 4), apes (n = 2) and OWM (n = 3). Box plots show median, quartiles (boxes), and range (whiskers). Two-sided Wilcoxon rank sum test is performed.



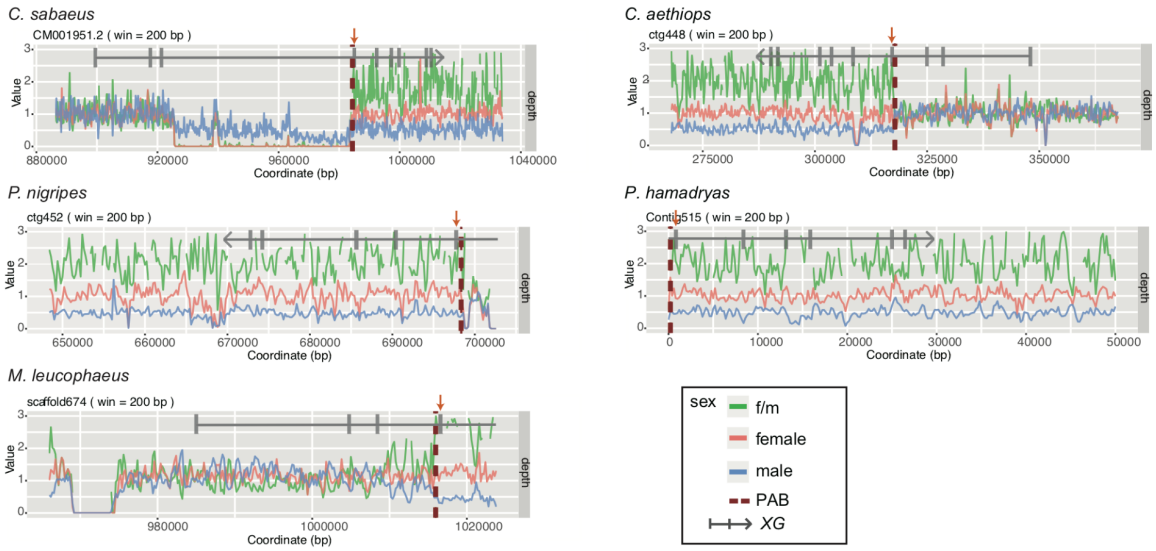
Supplementary Figure 3. The identity between X and Y at the X-transposed gene and its flanking 50 Kb region. Exons are drawn as dark blue boxes below each gene.



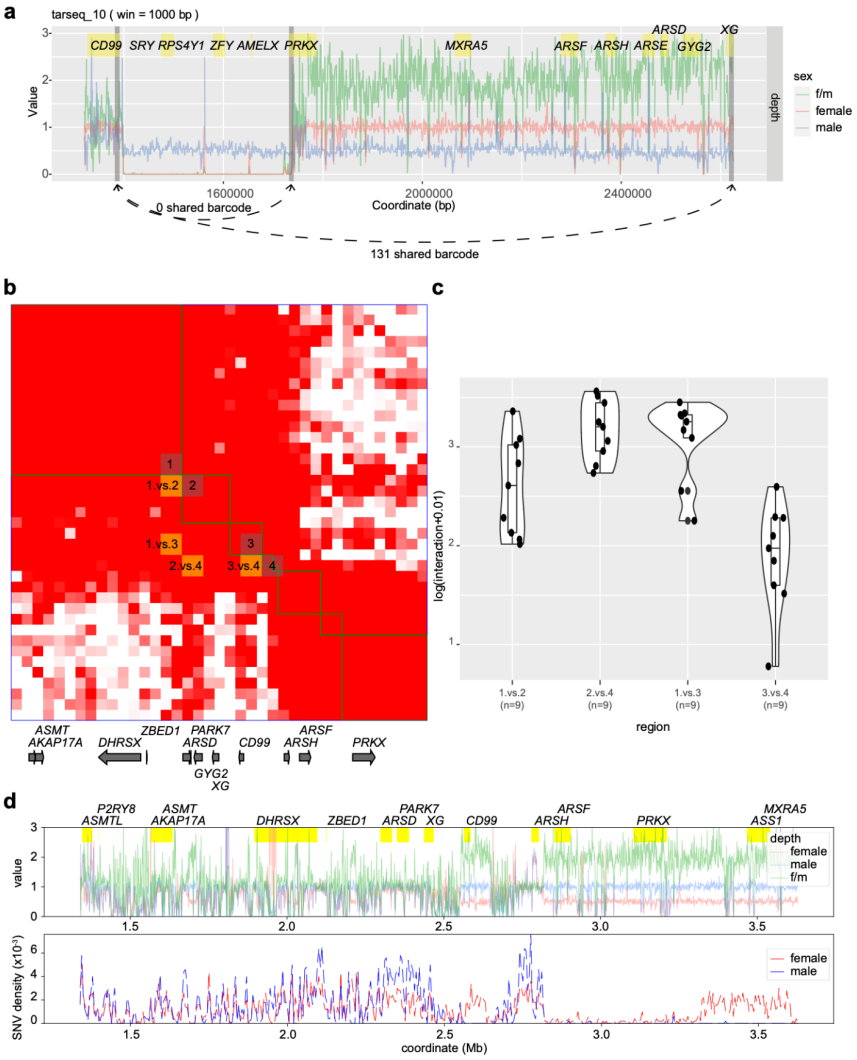
Supplementary Figure 4. Normalized male and female depth distribution around the PAR boundary in *T. belangeri, D. madagascariensis, N. pygmaeus, S. apella, C. albifrons, S. midas* and *A. geoffroyi*. Gene regions are highlighted in yellow and the PAB in each species are delineated with a red dotted line.



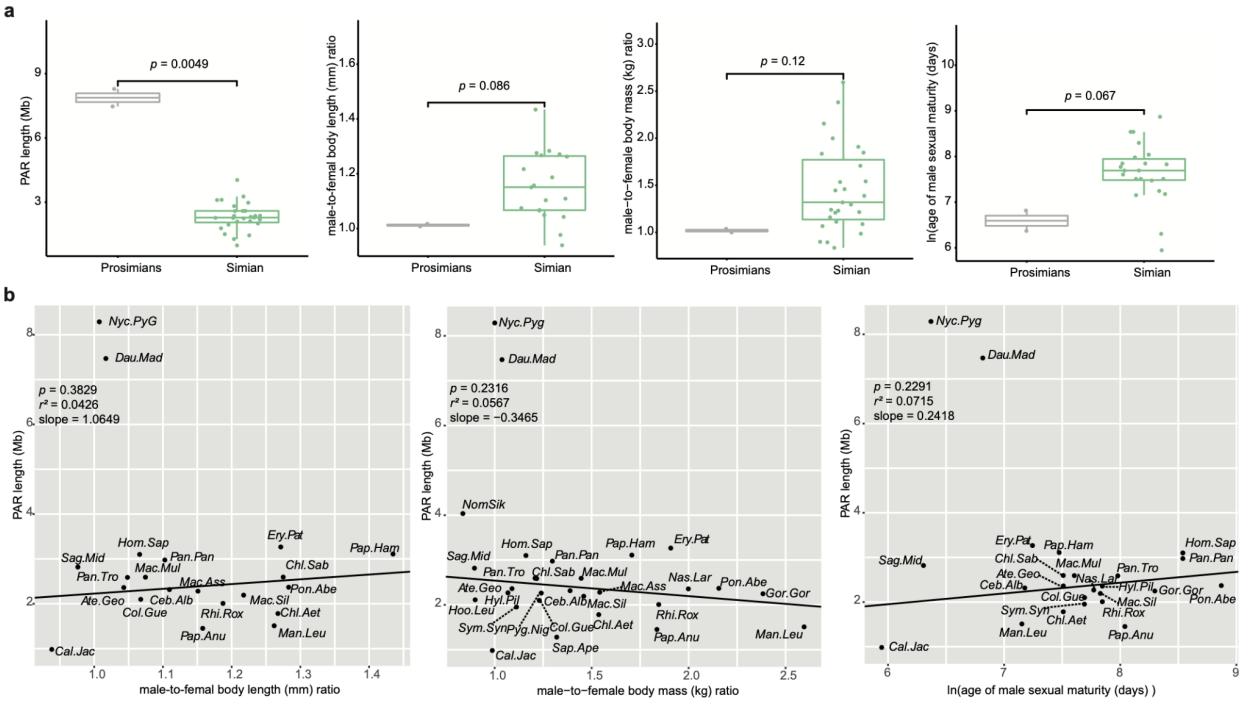
Supplementary Figure 5. Normalized male and female depth distribution around the PAR boundary in apes (*P. troglodytes, P. paniscus, G. gorilla, P. abelli, N. siki, S. syndactylus, H. leucoendys*) and Old World monkeys (*M. mulatta, M. silenus, M. assamensis,*). The orthologous position of the human PAR is marked with red dotted lines. *XG* exon positions are highlighted in grey. *XG*'s orthologous exon 4 is marked by an arrow. The 1st-3rd exons of *XG* are not assembled in *H. leuconedys* and *P. hamadryas*. The first exon of *XG* was not assembled in *M. assamensis*. In *M. silenus* and *S. syndactylus*, PAR was assembled with nonPARY; thus, the 4th-10th exons of *XG* were not annotated. *H. pileatus* is not shown as we were unable to find the orthologous coordinate of human PAB in this species and its *XG* is fragmented.



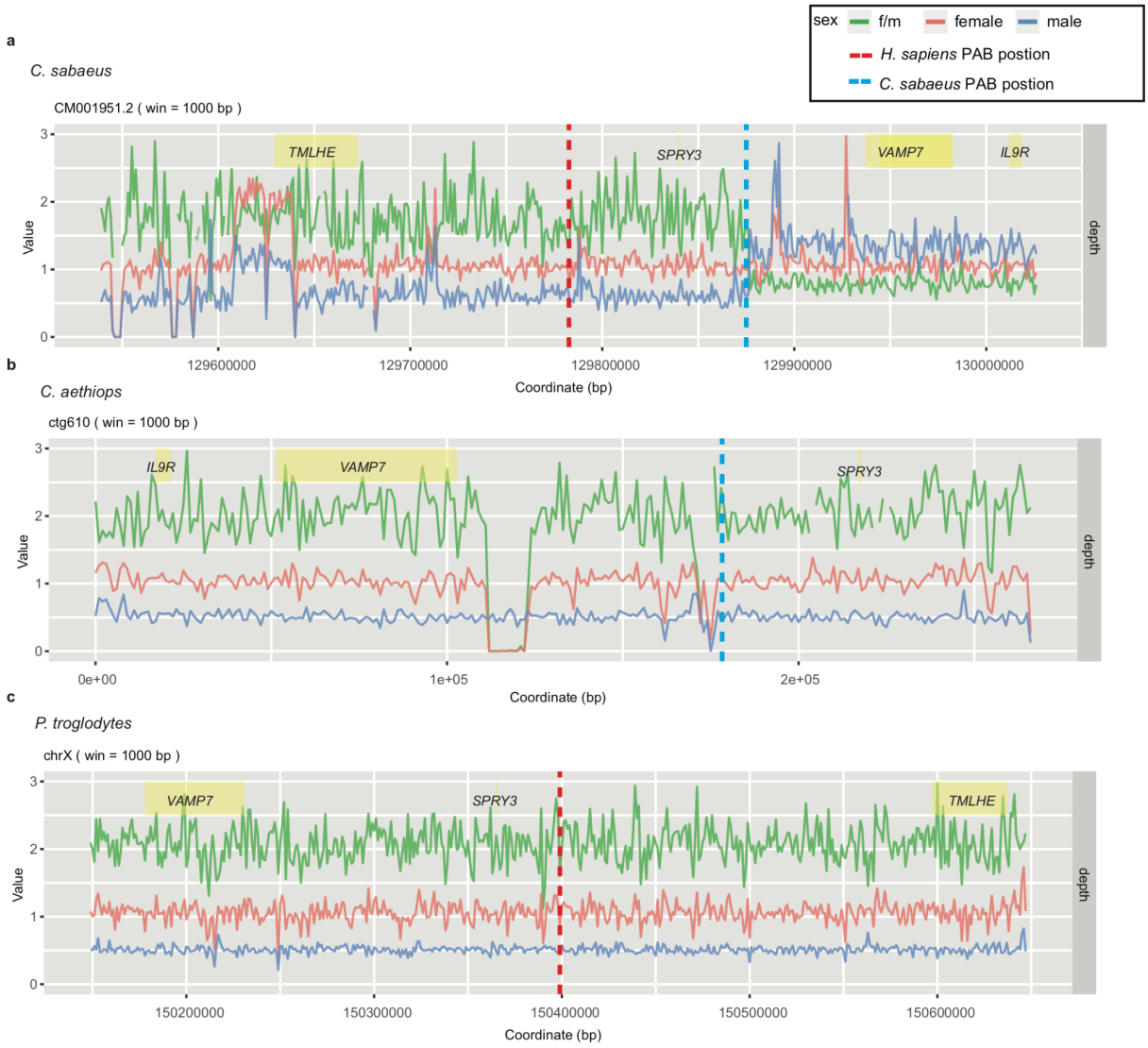
Supplementary Figure 6. Normalized male and female depth distribution around the PAR boundary in Old World monkeys (*C. sabaeus, C. aethiops, P. nigripes, P. hamadryas, M. leucophaeus*). The orthologous position of the human PAR is marked with red dotted lines. *XG* exon positions are highlighted in grey. *XG*'s orthologous exon 4 is marked by an arrow. *C. sabaeus* contains an X-Y chimeric assembly error. The 1st-3rd exons of *XG* are not assembled in *P. nigripes* and *P. hamadryas*. The 5th-last exons of XG are fragmentally assembled in different scaffolds of *M. leucophaeus. C. guereza*, *N. larvatus, R. roxellana, E. patas* and *P. anubis* are not shown as we were unable to find the orthologous coordinate of human PAB in this species and its *XG* is fragmented.



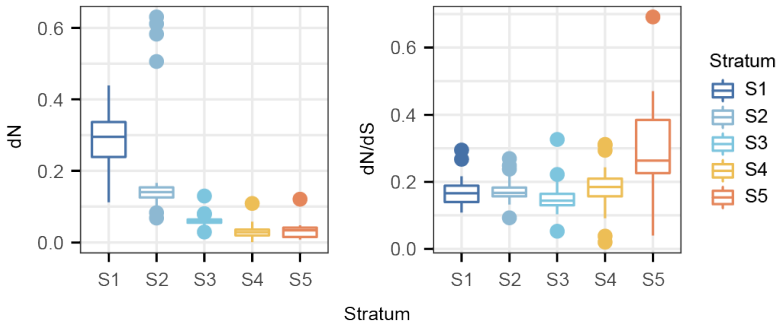
Supplementary Figure 7. Examination of gene order in *A. geoffroyi* (a) and *S. midas* (b-d). **a** 10X-linked reads suggest that *XG* should be closer to *CD99* than *PRKX* in *A. geoffroyi*. There are 133 10X-linked reads supporting the linkage between *XG* end and *CD99* whereas there were no 10X-linked reads to support the linkage between *PRKX* end and *CD99*, suggesting that the correct gene order should be: *CD99, XG, GYG2, ..., PRKX* in this species. Examined 1 Kb regions are shown in grey. **b** Hi-C map under 25 Kb resolution shows the inversion (covering genes from *ARSD* to *CD99*) and the flanking region. Pairwise interactions between the four 100 Kb regions (1, 2, 3, 4) were extracted and compared. Green boxes delineate contig boundaries. **c** Pairwise interaction strength comparison shows higher interaction between region 2 and region 4 as well as between region 1 and region 3, suggesting that the inversion is an assembly artifact in *S. midas*, and the correct gene order should be: *DHRSX, ZBED1, CD99, XG, PARK7, ARSD, ARSH, ARSF*. $n = 9$ interaction values for every region pair. Box plots show median, quartiles (boxes), and range (whiskers). **d** Although sequencing depth does not differ in *ARSD*, *PARK7*, *XG* and *ARSH*, SNV density is higher at *ARSD*, *PARK7*, *XG* and *ARSH* in male than in female, suggesting that the region is under X/Y divergence in *S. midas*.



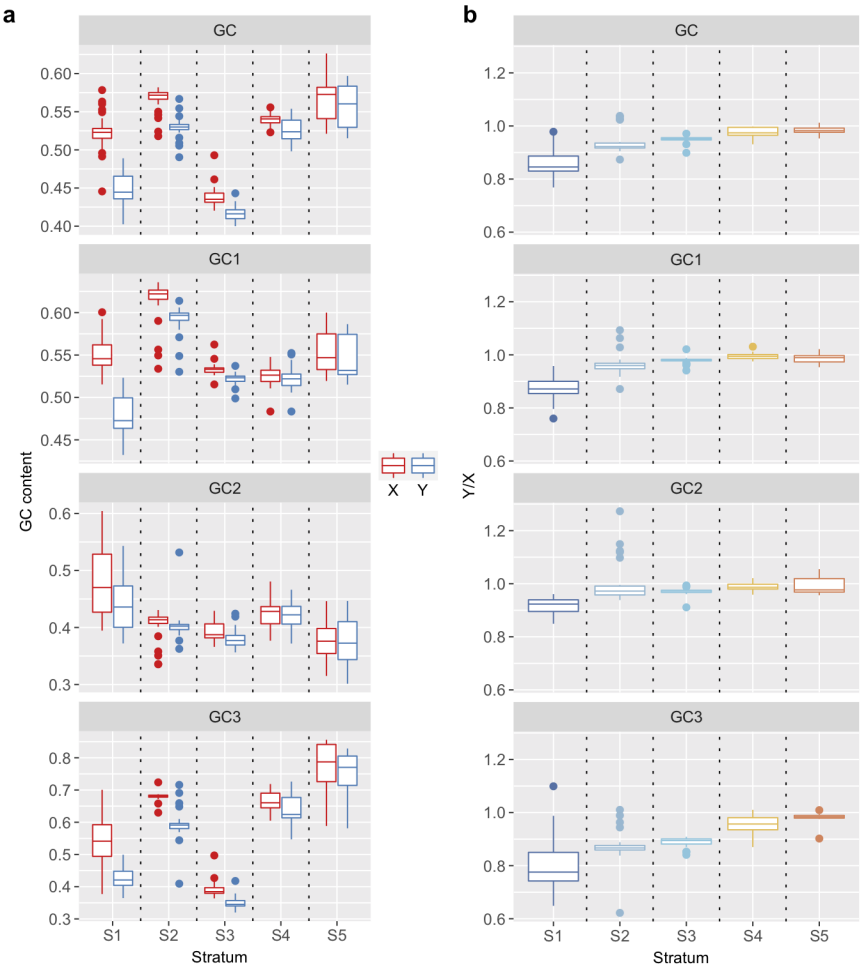
Supplementary Figure 8. PAR length and life history trait comparison between prosimians and *Simiiformes*. a Only PAR length is significantly longer in prosimians than other *Simiiformes*. $n = 2$ for prosimians in every panel and $n = 27, 19, 27, 22$, from left to right for simians. Box plots show median, quartiles (boxes), and range (whiskers). Two-sided Wilcoxon rank sum test is performed. **b** PGLS analysis suggests no correlation between PAR length and body length dimorphism, or body mass dimorphism, or age of male sexual maturity (lambda = 0 for all three analyses). Two-sided F-test is used in PGLS analysis and we do not apply multiple testing correction to adjust the p-value.



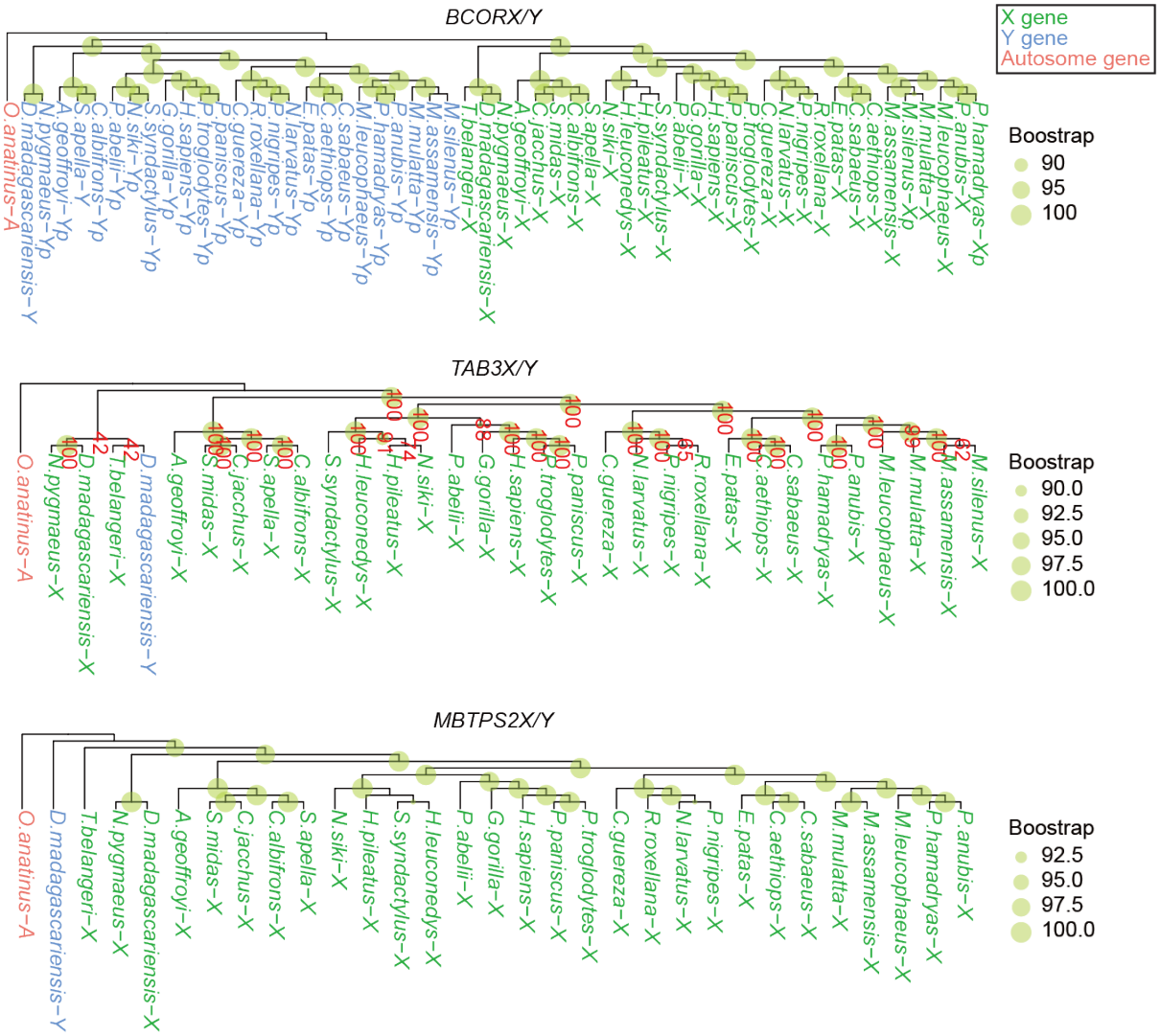
Supplementary Figure 9. Normalized male and female depth distribution around the human PAR2 orthologous region in *C. sabaeus* (a), *C. aethiops* (b, closely related to *C. sabaeus*) and *P. troglodytes* (c, closely related to human). The *VAMP7* and *IL9R* located region harbors similar male and female normalized depths in *C. sabaeus* but not in *C. aethiops*, suggesting that PAR2 containing *VAMP7* and *IL9R* evolved independently in *C. sabaeus*. Human orthologous PAB is delineated with a red dotted ine and *C. sabaeus* (orthologous) PAB is delineated with blue dotted line.



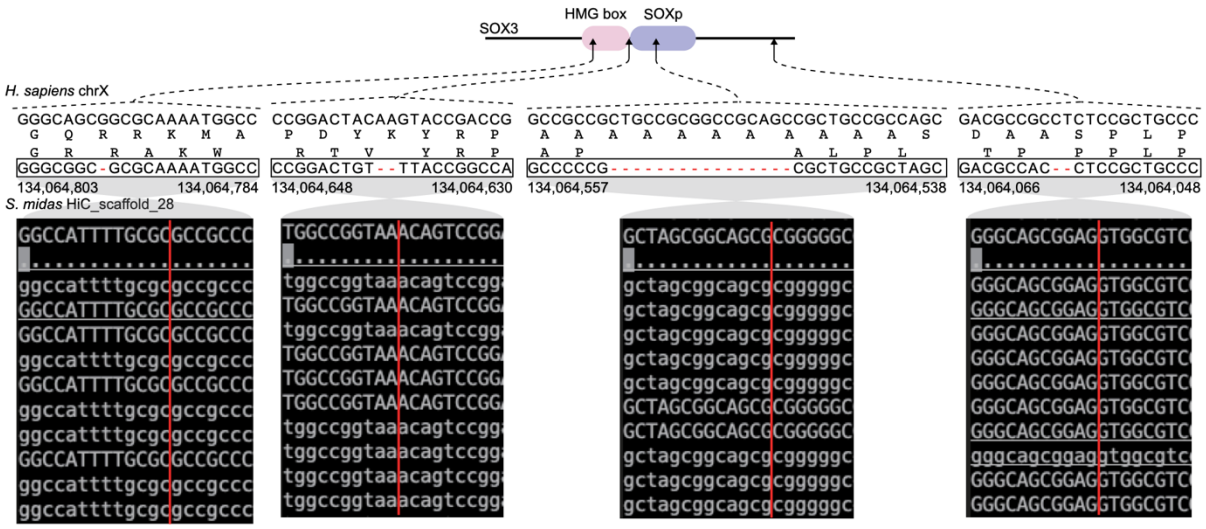
Supplementary Figure 10. Distributions of *dN* and *dN/dS* values calculated using the KaKs_Calculator for concatenated stratum sequences of each species. *n* = 30, 30, 30, 19 and 8 in the order from S1 to S5 (treeshrew and prosimians are included in S1-S3). Box plots show median, quartiles (boxes), and range (whiskers).



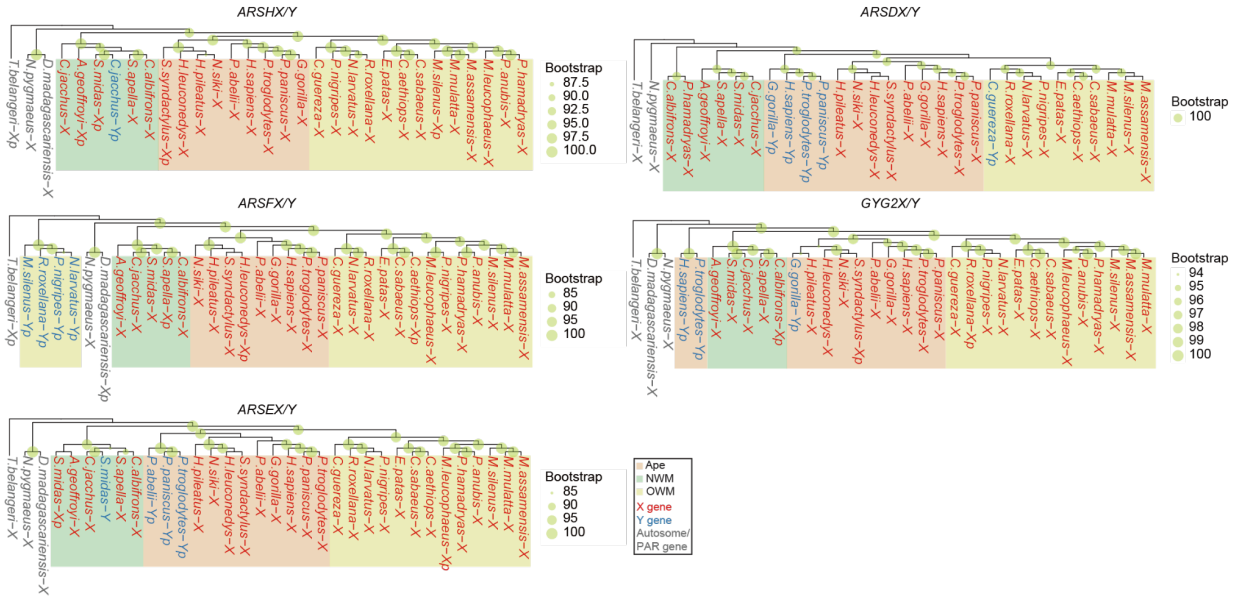
Supplementary Figure 11. GC proportion of X and Y gametologs for concatenated stratum sequences of each species. a GC proportion of X and Y gametologs for concatenated stratum sequences of each species. Box plots show median, quartiles (boxes), and range (whiskers). **b** Ratio of GC proportions between Y and X for concatenated stratum sequences of each species (GC = all sites; GC1 = sites at first codon position only; GC2 = sites at second codon position only; GC3 = sites and third codon position only). *n* = 30, 30, 30, 19 and 8 in the order from S1 to S5 in both panels (treeshrew and prosimians are included in S1-S3). Box plots show median, quartiles (boxes), and range (whiskers).



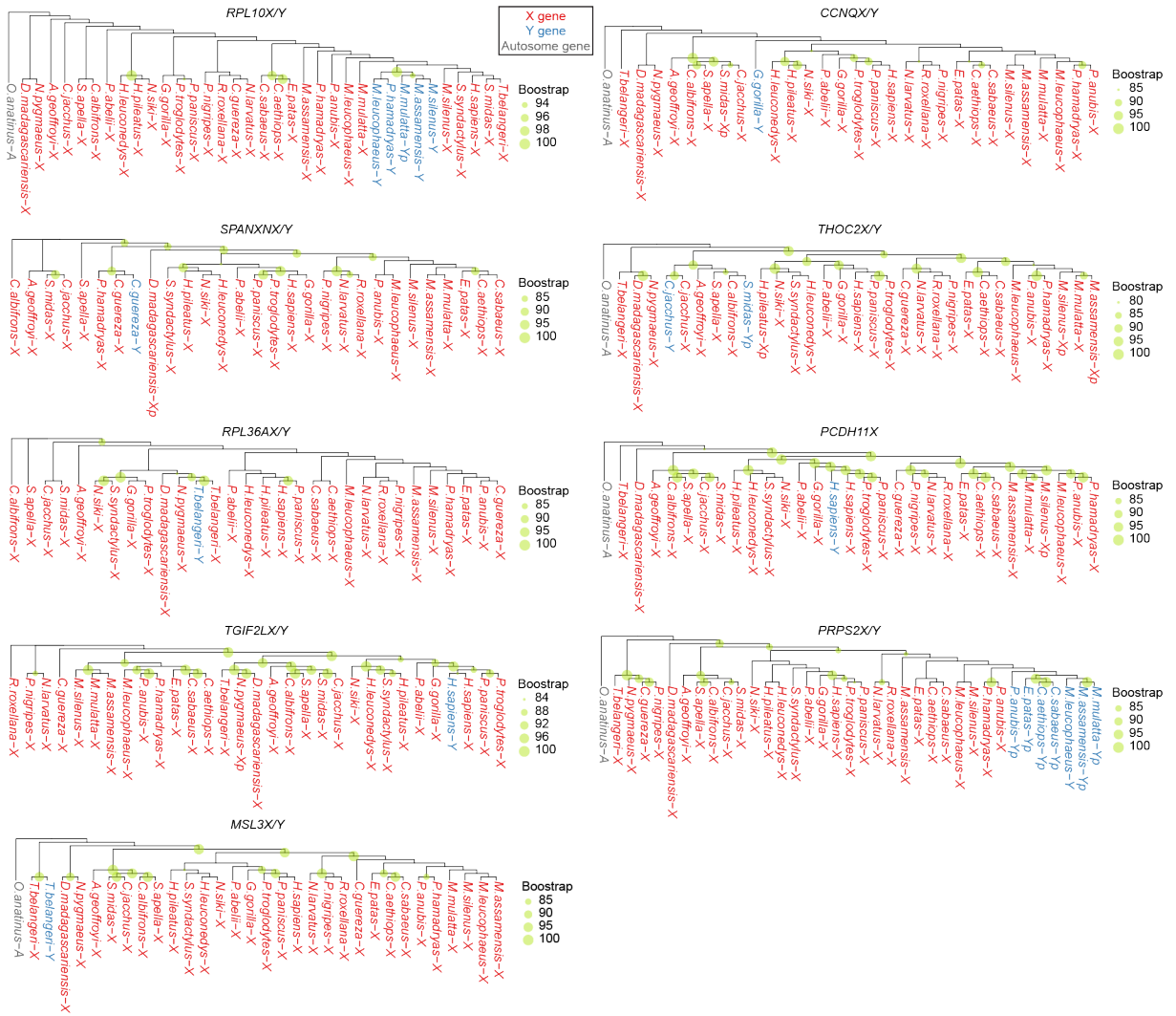
Supplementary Figure 12. Gamtolog phylogenetic trees of *BCORX/Y*, *TAB3X/Y* and *MBTPS2X/Y*. Gene names are color-coded according to chromosome (green: X, blue: Y, red: autosome). Pseudogenes are marked by a "p" suffix. Nodes with bootstrap ≥ 80 are displayed with green circles with size representing the bootstrap values. All bootstraps in TAB3X/Y are labeled in red.



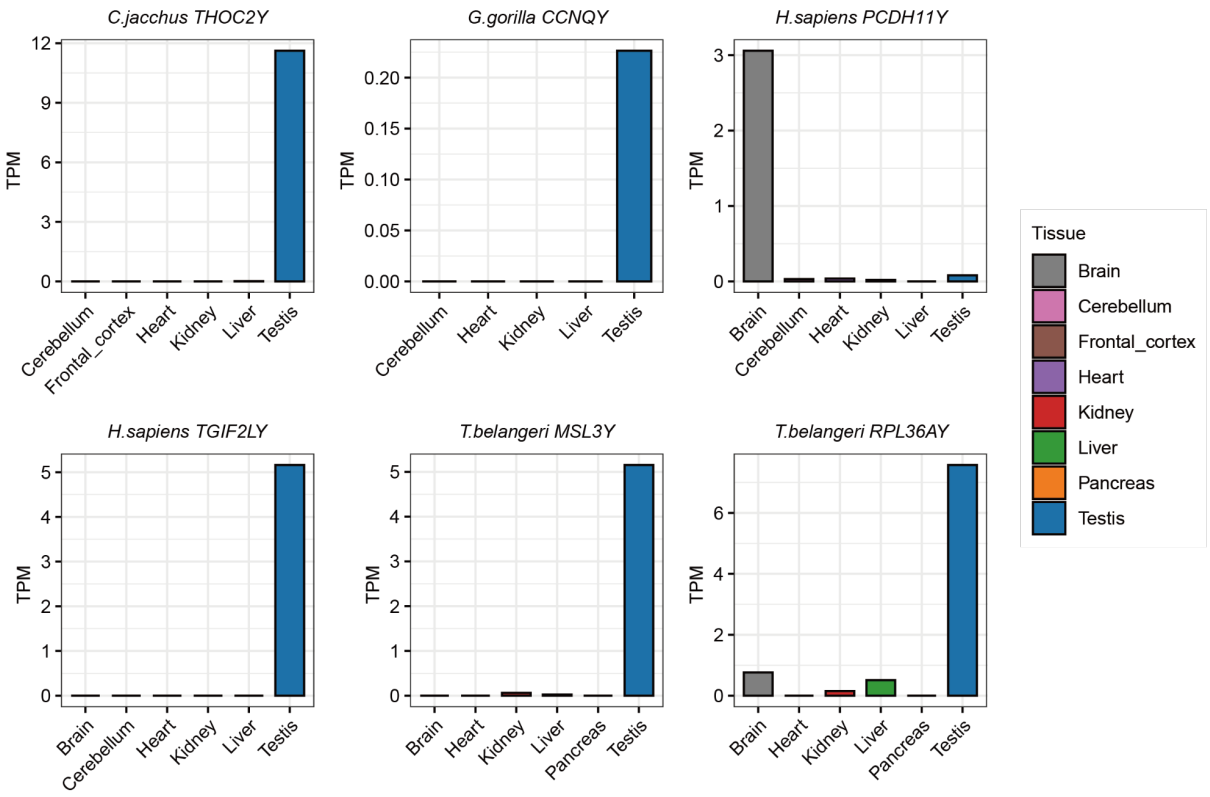
Supplementary Figure 13. Examination of frameshift signals in *Saguinus midas SOX3*. Illumina short reads confirmed the four frameshift signals caused by nucleotide deletion within the two functional domains, resulting in the fragmentation of the HMG box and the loss of the SOXp domain.



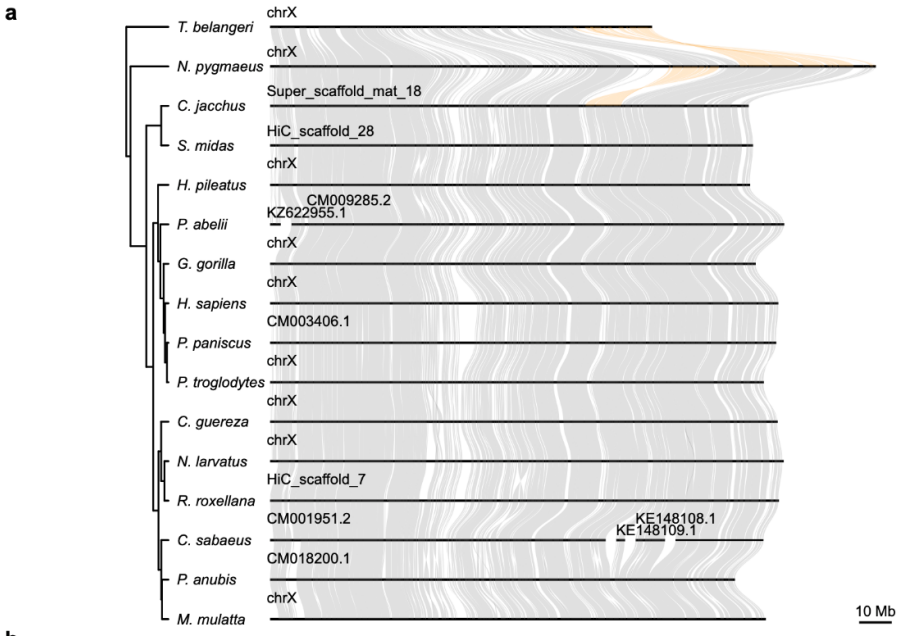
Supplementary Figure 14. Phylogenetic trees of S5 X/Y gametologs *ARSHX/Y*, *ARSFX/Y*, *ARSEX/Y, ARSDX/Y* and *GYG2X/Y*. Gene names are color-coded according to chromosomes (red: X, blue: Y, grey: PAR). Leaf node background is color-coded according to lineage (yellow: OWM, red: apes, green: NWM). Pseudogenes are marked by a “p” suffix.

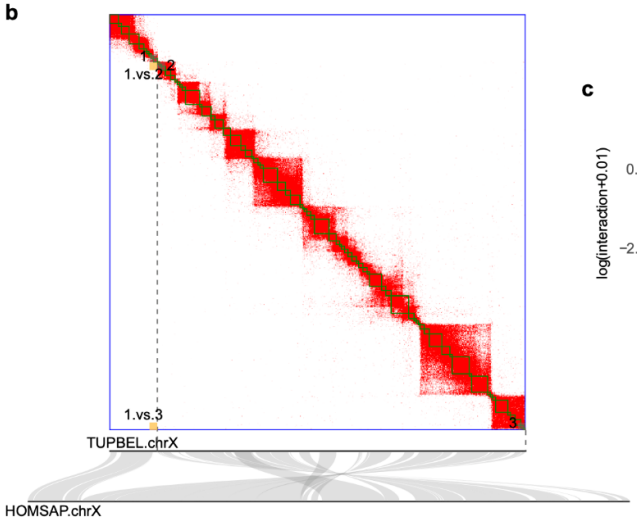


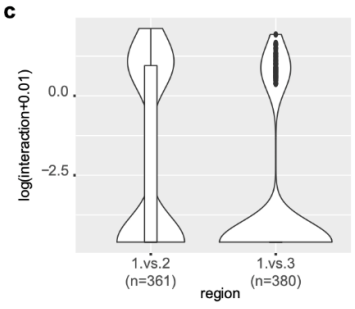
Supplementary Figure 15. Phylogenetic trees of X-transpose Y-linked gene with the X homolog. The Y gametologs are clustered with the X gametologs, suggesting that they were recently duplicated from the X. Gene names are color-coded according to chromosomes (red: X, blue: Y, grey: autosome). Pseudogenes are also included and are marked with a “p” suffix.

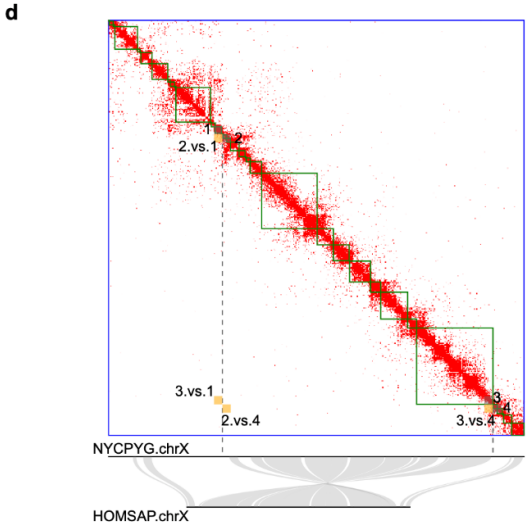


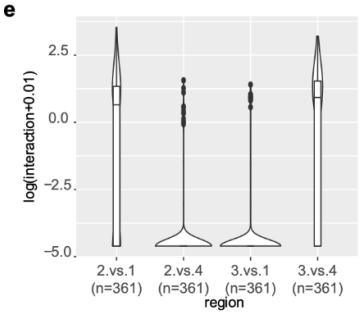
Supplementary Figure 16. Expression of six X-transposed Y genes in multiple male tissues, including *THOC2Y* in *C. jacchus*, *CCNQY* in *G. gorilla*, *PCDH11Y* and *TGIF2LY* in *H. sapiens*, and *MSL3Y* and *RPL36AY* in *T. belangeri*.


a
T. belangeri
N. pygmaeus
C. jacchus
S. midas
H. pileatus
P. abelii
G. gorilla
H. sapiens
P. paniscus
P. troglodytes
C. guereza
N. larvatus
R. roxellana
C. sabaeus
P. anubis
M. mulatta
chrX
Super_scaffold_mat_18
HiC_scaffold_28
CM009285.2
KZ622955.1
CM003406.1
HiC_scaffold_7
CM001951.2
KE148108.1
KE148109.1
CM018200.1
10 Mb

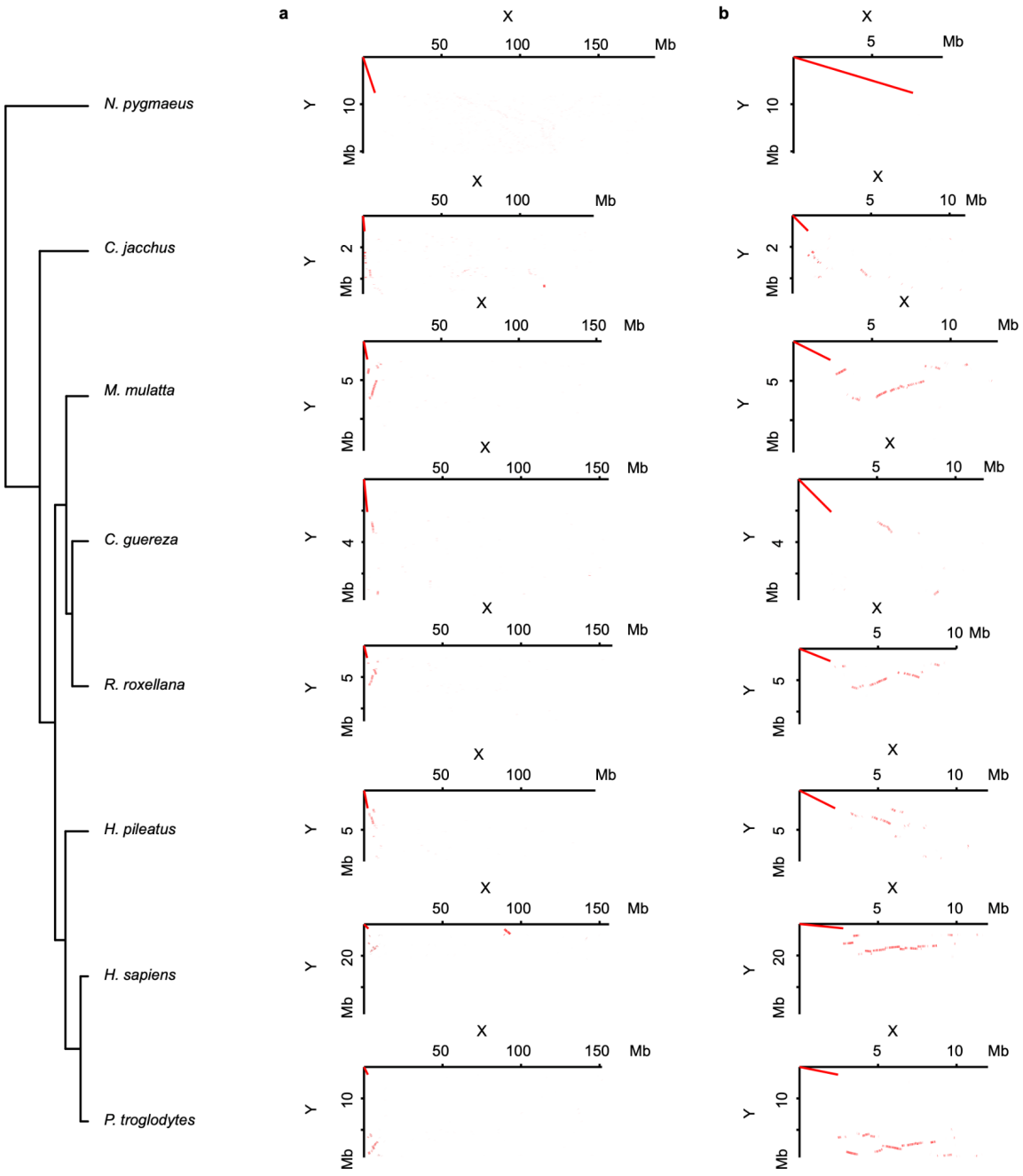

b
1.vs.2
1.vs.3
TUPBEL.chrX
HOMSAP.chrX


c
log(interaction+0.01)
0.0
−2.5
1.vs.2
(n=361)
1.vs.3
(n=380)
region


d
2.vs.1
3.vs.1
2.vs.4
3.vs.4
NYCPYG.chrX
HOMSAP.chrX

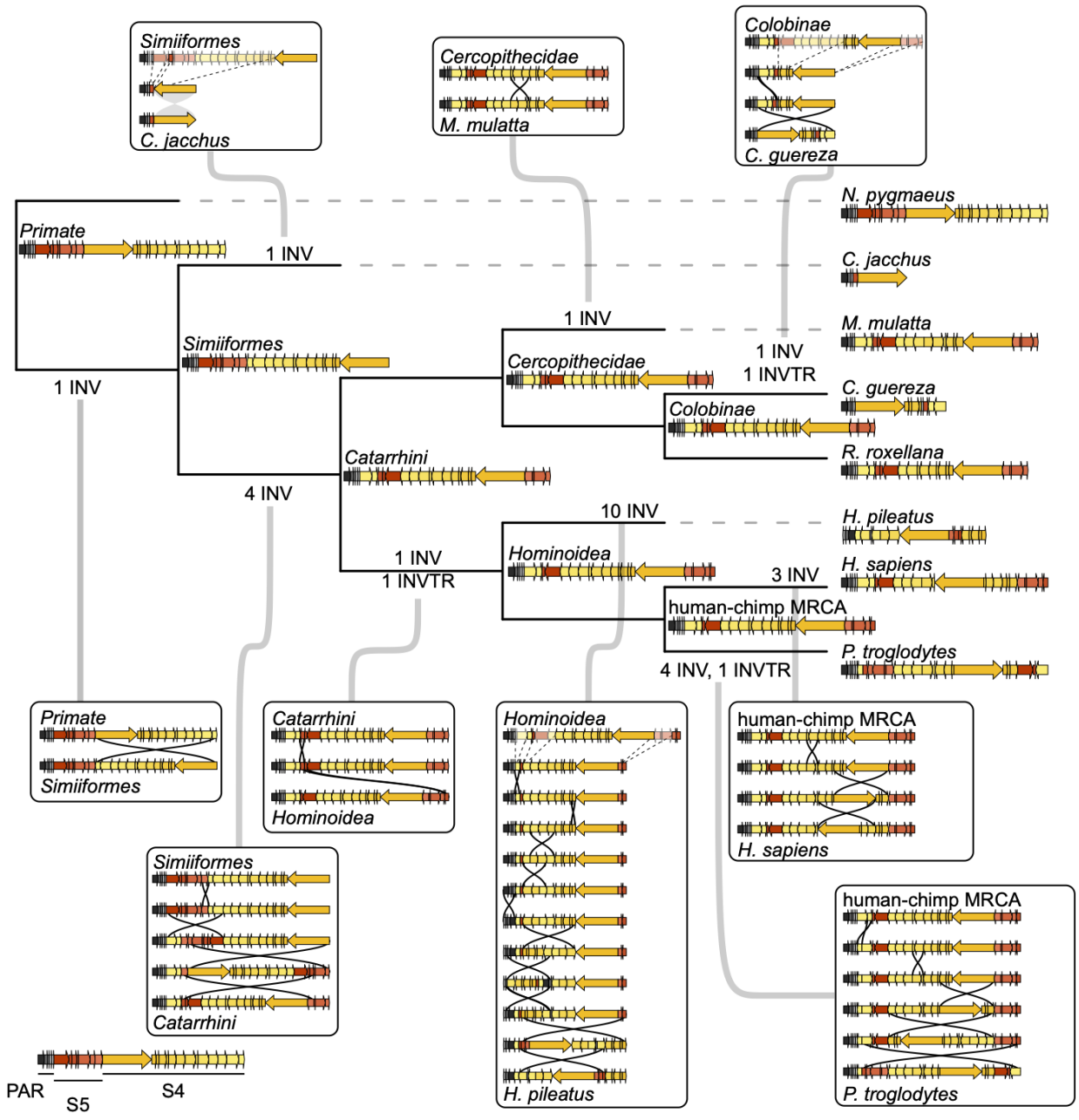

e
log(interaction+0.01)
2.5
0.0
−2.5
−5.0
2.vs.1
(n=361)
2.vs.4
(n=361)
3.vs.1
(n=361)
3.vs.4
(n=361)
region

Supplementary Figure 17. X chromosome synteny and confirmation among studied species with chromosome-level assembly. a After confirmation with Hi-C data, only two large inversions (> 2 Mb, marked in orange) were found: one in treeshrew (*T. belangeri*) and one in pygmy slow loris (*N. pygmaeus*). **b** Hi-C map under 25 Kb resolution shows the inversion (chrX:93,175,425-117,245,019) and the flanking region (chrX:90,000,000-117,263,549) in *T. belangeri*. Pairwise interaction (orange boxes) between the three 200 Kb regions (1-3, gray boxes) were extracted and compared. Green boxes delineate contig boundaries. **c** Pairwise interaction strength comparison shows greater interaction between region 2 and region 1 (1.vs.2) than that between region 3 and region 1 (1.vs.3), suggesting that the inversion in **(b)** is genuine. *n* = 361 and 380 interaction values for 1.vs.2 and 1.vs.3, respectively. Box plots show median, quartiles (boxes), and range (whiskers). **d** Hi-C map under 25 Kb resolution shows the inversion (chrX:123,742,252-137,640,438) and the flanking region (chrX:115,376,843-139,776,886) in *N. pygmaeus*. Pairwise interactions (orange boxes) between the four 500 Kb regions (1-4, gray boxes) were extracted and compared. Green boxes delineate contig boundaries. **e** Pairwise interaction strength comparison shows greater interaction between region 1 and region 2 (2.vs.1), and between region 3 and region 4 (3.vs.4), suggesting that the inversion is genuine. *n* = 361 interaction value for every region pair. Box plots show median, quartiles (boxes), and range (whiskers).

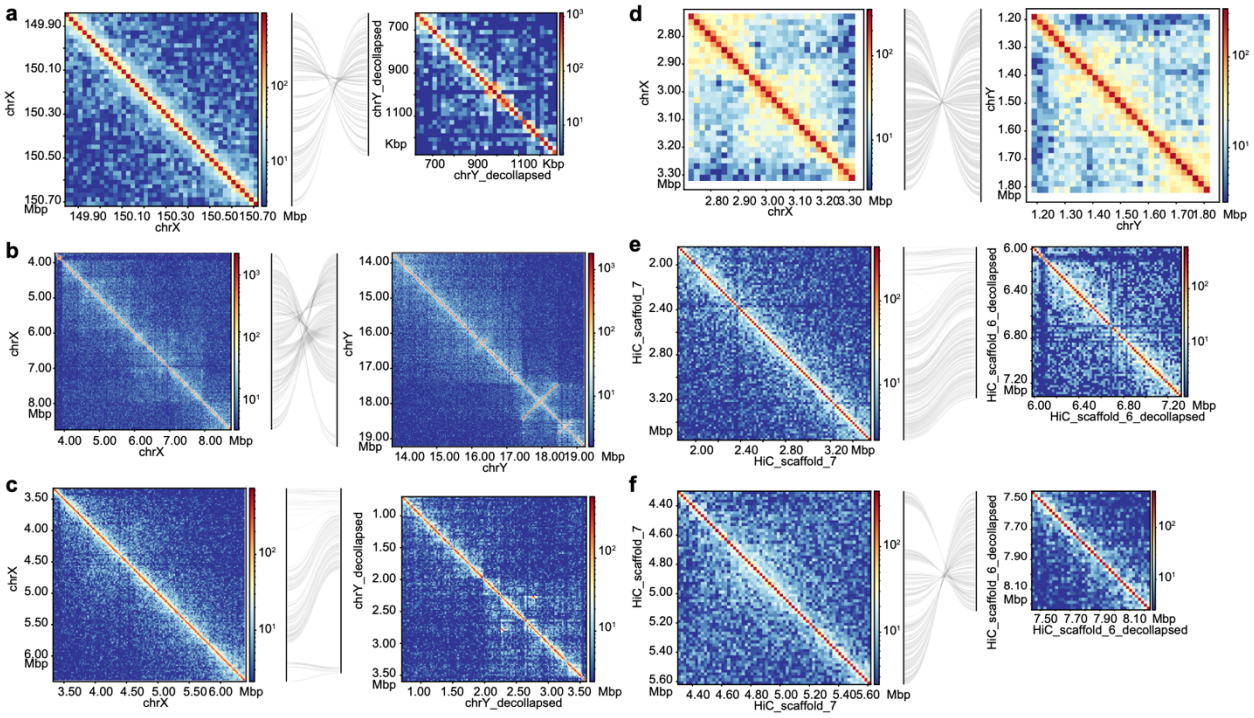


Supplementary Figure 18. Dotplots showing the LastZ alignment between X-linked and Y-linked sequences in representative primate species. a Alignment between the whole X and Y chromosomes showing that most of the alignable regions locate near PAR (upper right of the dotplot). *N. pygmaeus*: chrX vs. chrY_decollapsed, *C. jacchus*: Super_scaffold_mat_18 vs. CM021938.1_decollapsed, *M. mulatta*: chrX vs. chrY, *C. guereza*: chrX vs. chrY_decollapsed, *R. roxellana*: HiC_scaffold_7 vs. HiC_scaffold_6_decollapsed, *H. pileatus*: chrX vs. chrY_decollapsed, *H. sapiens*: chrX vs. chrY, *P. troglodytes*: chrX vs. chrY. **b** The zoomed-in dotplot of the Y versus the PAR-S5-S4 region of the X. *N. pygmaeus*: chrX:0-9500000, *C.*

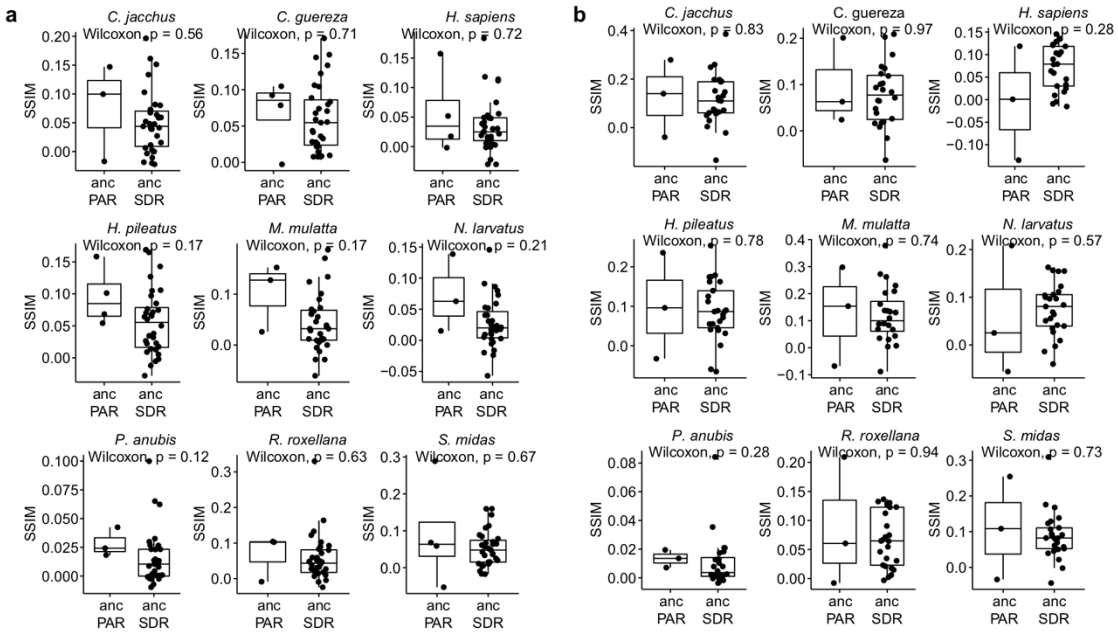
jacchus: Super_scaffold_mat_18:0-11000000, *M. mulatta*: chrX:0-13000000, C. guereza: chrX:144000000-155790761, *R. roxellana*: HiC_scaffold_7:0-10000000, *H. pileatus*: chrX:0-12000000, *H. sapiens*: chrX:0-12000000, *P. troglodytes*: chrX:0-12000000.



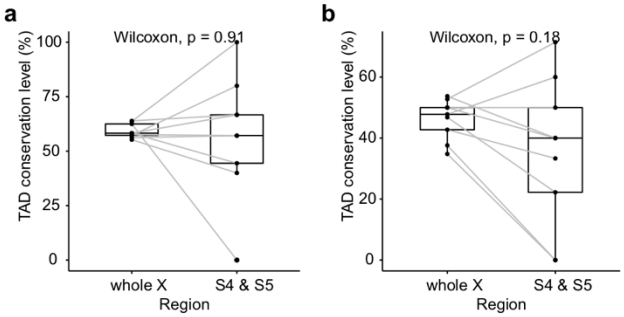
Supplementary Figure 19. Y ancestral order reconstruction and inference of structural variations with conserved blocks in PAR, S5 and S4 under 30 Kb resolution. Reconstruction of the ancestral Y sequence order with conserved blocks under 30 Kb resolution. A total of 26 conserved blocks were used in reconstruction. Each arrow block represents a conserved segment, color coded in PAR, S5 and S4. Structural variation events marked at each branch were inferred with GRIMM. Lost blocks are marked with light color and dashed lines.



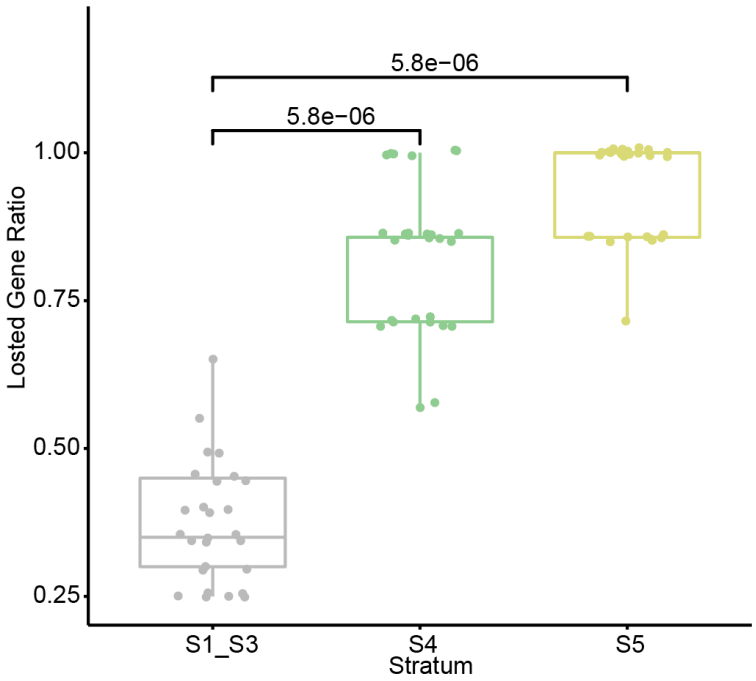
Supplementary Figure 20. Hi-C contact map between homologous blocks between X and Y in *Colobus guereza* (a), *Homo sapiens* (b), *Hylobates pileatus* (c), *Macaca_mulatta* (d) and *Rhinopithecus roxellana* (e, f).



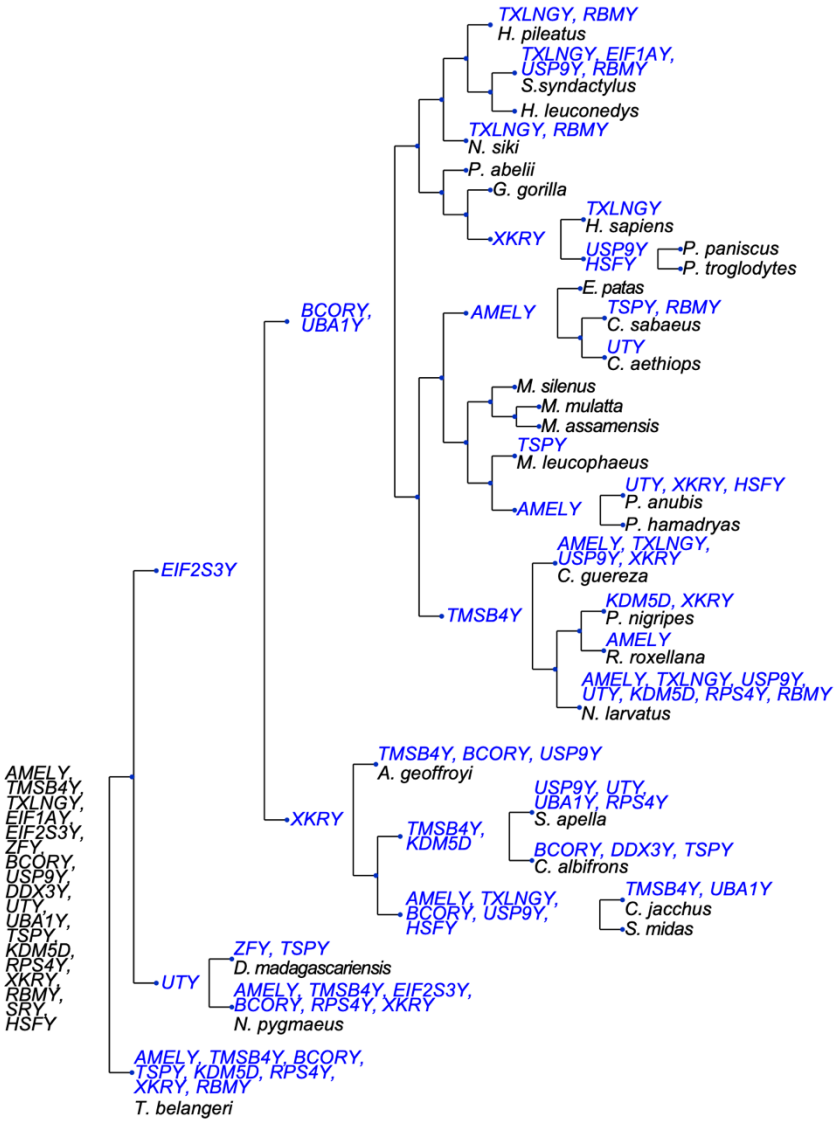
Supplementary Figure 21. SSIM comparison between the ancestral PAR (i.e. *Simiiformes* S4, S5 and PAR) and the ancestral SDR suggesting that chromatin configuration is at the same level between these two regions (two-sided Wilcoxon rank-sum test, *p-value* > 0.01), between *Simiiformes* and *N. pygmaeus* (a) or *T. belangeri* (b). *n* = 41, 40, 41, 41, 41, 39, 39, 41, 41 SSIM values calculated based on the syntenic block between *N. pygmaeus* and each of *C. jacchus*, *C. guereza*, *H. sapiens*, *H. pileatus*, *M. mulatta*, *N. larvatus*, *P. anubis*, *R. roxellana*, *S. midas*, respectively. *n* = 42, 41, 42, 42, 42, 40, 40, 42, 42 SSIM values calculated based on the syntenic block between *T. belangeri* and each of *C. jacchus*, *C. guereza*, *H. sapiens*, *H. pileatus*, *M. mulatta*, *N. larvatus*, *P. anubis*, *R. roxellana*, *S. midas*, respectively. Box plots show median, quartiles (boxes), and range (whiskers).



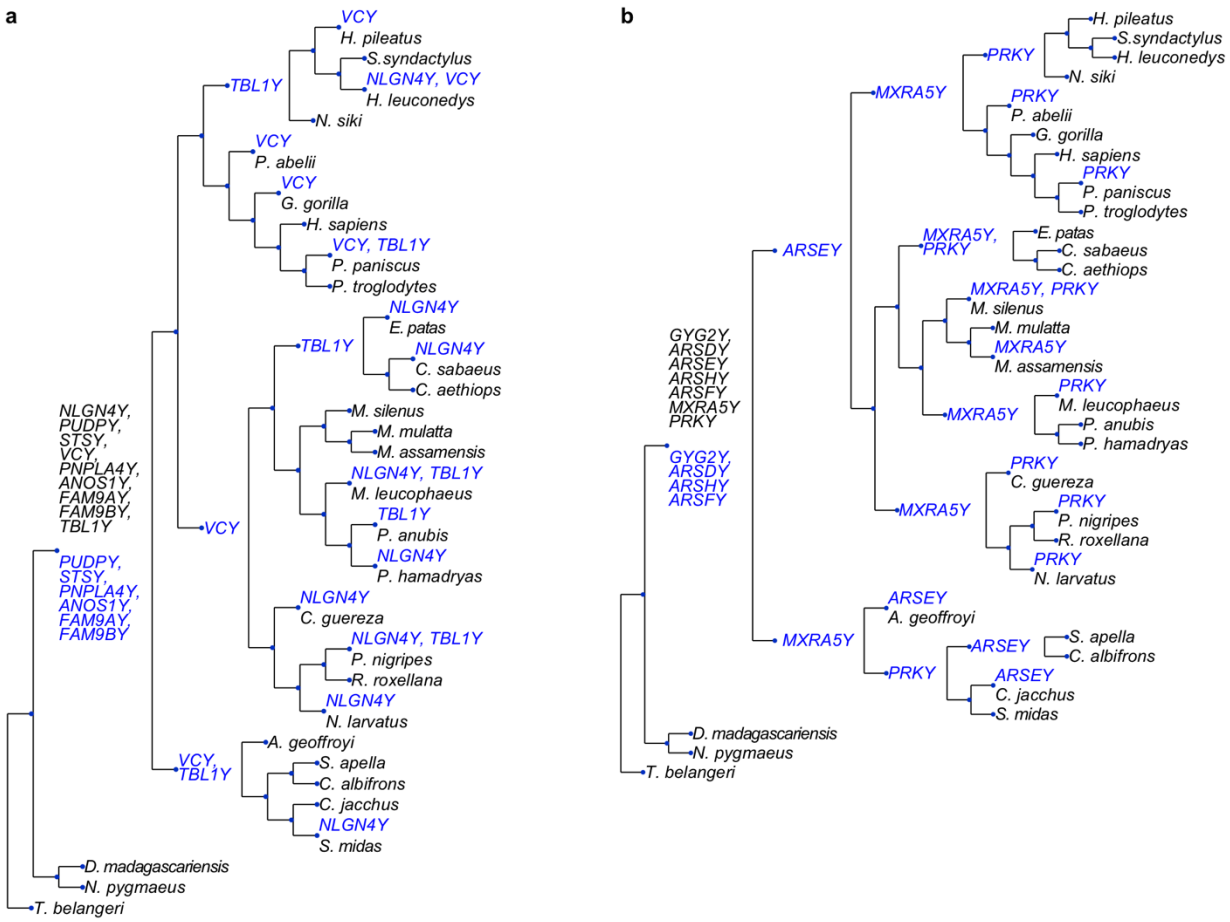
Supplementary Figure 22. TAD boundary conservation level between *Simiiformes* and *N. pygmaeus* (a) or *T. belangeri* (b) in different X regions. Comparison between the whole X and *Simiiformes* S4 and S5 suggests that the conservation level is similar between these two regions (paired two-sided Wilcoxon rank-sum test *p-value* > 0.05). $n = 9$ comparison between *Simiiformes* and *N. pygmaeus* (or *T. belangeri*). Box plots show median, quartiles (boxes), and range (whiskers).



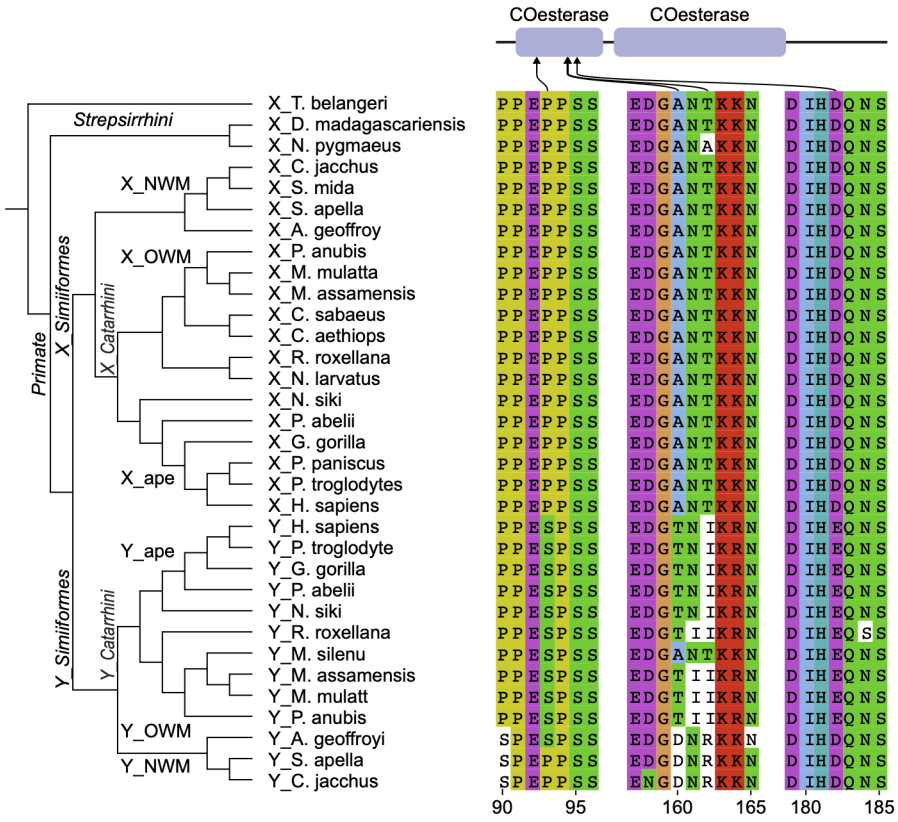
Supplementary Figure 23. The comparison of lost gene ratio between S1-3, S4 and S5. The lost gene ratio of S1-3 is significantly less than that of S4 and S5 (two-sided Wilcoxon rank-sum test, p-value < 0.01). *n* = 27 *Simiiformes* species are included for each box. Box plots show median, quartiles (boxes), and range (whiskers).



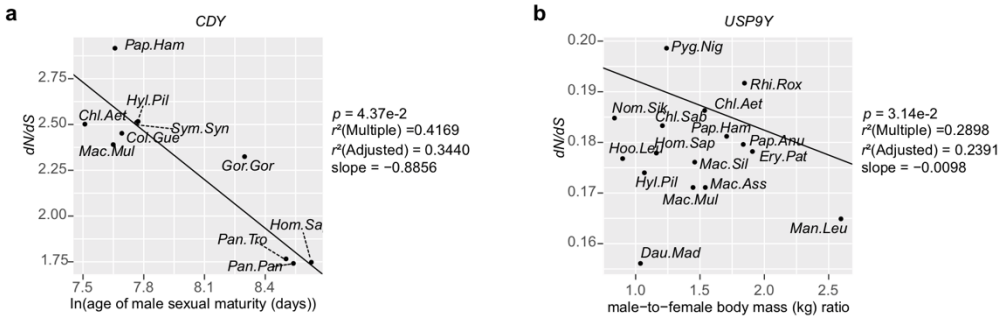
Supplementary Figure 24. The evolution of X-degenerate Y-linked genes within S1-S3 during primate evolution. Gene losses are marked above each node. Black: Y-linked genes at *Euarchontoglires* MRCA; blue: genes lost. Tree is not drawn to scale.



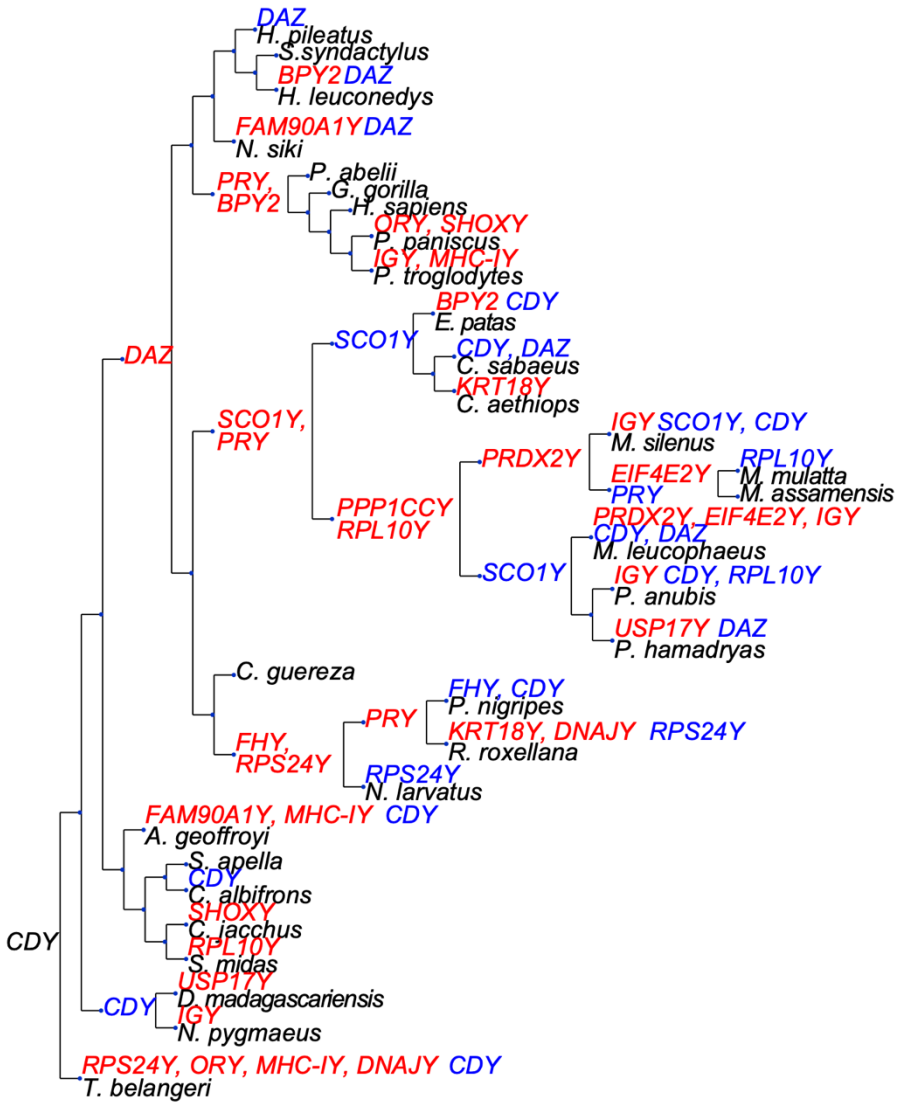
Supplementary Figure 25. The evolution of X-degenerate Y-linked genes of S4 (a) and S5 (b) during primate evolution. Gene losses are marked above each node. Black: Y-linked genes at *Simiiformes* MRCA; blue: genes lost. Tree is not drawn to scale.



Supplementary Figure 26. The key amino acid residue, Pro93Ser, and other three amino acids positions that differ between NLGN4X and NLGN4Y in primates.



Supplementary Figure 27. The dN/dS values of *CDY* (a) and *USP9Y* (b) show a correlation with various primate reproductive traits. Log transformation was applied to all traits except body mass dimorphism. Two-sided F-test is used in PGLS analysis and we do not apply multiple testing correction to adjust the p-value.



Supplementary Figure 28. The evolution of non X-degenerate Y-linked genes during primate evolution. Gene gains or losses are marked above each node. Black: Y-linked genes at *Euarchontoglires* MRCA; blue: gene loss, red: gene gain. Tree is not drawn to scale. Only genes present in more than one species are plotted.

References

1 Shao., Y. *et al.* Phylogenomic analyses provide insights into primate genomic and phenotypic evolution. *Submitted* (2021).

2 Reis, M. D. *et al.* Using phylogenomic data to explore the effects of relaxed clocks and calibration strategies on divergence time estimation: primates as a test case. *Systematic Biology* **67**, 594-615 (2018).

3 Rhie, A. *et al.* Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737-746, doi:10.1038/s41586-021-03451-0 (2021).

4 Skaletsky, H. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825-837 (2003).

5 Hughes, J. F. *et al.* Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* **463**, 536-539, doi:10.1038/nature08700 (2010).

6 Hughes, J. F. *et al.* Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* **483**, 82-86, doi:10.1038/nature10843 (2012).

7 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

8 Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic acids research* **44**, D279-D285 (2016).

9 Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods* **17**, 155-158, doi:10.1038/s41592-019-0669-3 (2020).

10 Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95-98, doi:10.1016/j.cels.2016.07.002 (2016).

11 Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing in science & engineering* **9**, 90-95 (2007).

12 Birney, E., Clamp, M. & Durbin, R. GeneWise and genomewise. *Genome research* **14**, 988-995 (2004).

13 Li, L., Stoeckert, C. J., Jr. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178-2189, doi:10.1101/gr.1224503 (2003).

14 van Dongen, S. & Abreu-Goodger, C. Using MCL to extract clusters from networks. *Methods Mol Biol* **804**, 281-295, doi:10.1007/978-1-61779-361-5_15 (2012).

15 Martínez-Pacheco, M. *et al.* Expression evolution of ancestral XY gametologs across all major groups of placental mammals. *Genome biology and evolution* **12**, 2015-2028 (2020).

16 Hu, F., Lin, Y. & Tang, J. MLGO: phylogeny reconstruction and ancestral inference from gene-order data. *BMC Bioinformatics* **15**, 354, doi:10.1186/s12859-014-0354-6 (2014).

17 Yang, C. *et al.* Evolutionary and biomedical insights from a marmoset diploid genome assembly. *Nature* **594**, 227-233 (2021).