The Complete Sequence and Comparative Analysis of Ape Sex Chromosomes

Kateryna D. Makova[1*^], Brandon D. Pickett[2*], Robert S. Harris[1*], Gabrielle A. Hartley[3*], Monika Cechova[4*], Karol Pal[1], Sergey Nurk[2], DongAhn Yoo[5], Qiuhui Li[6], Prajna Hebbar[4], Barbara C. McGrath[1], Francesca Antonacci[7], Margaux Aubel[8], Arjun Biddanda[6], Matthew Borchers[9], Erich Bomberg[8,10], Gerard G. Bouffard[2], Shelise Y. Brooks[2], Lucia Carbone[11,12], Laura Carrel[13], Andrew Carroll[14], Pi-Chuan Chang[14], Chen-Shan Chin[15], Daniel E. Cook[14], Sarah J.C. Craig[1], Luciana de Gennaro[7], Mark Diekhans[4], Amalia Dutra[2], Gage H. Garcia[5], Patrick G.S. Grady[3], Richard E. Green[4], Diana Haddad[16], Pille Hallast[17], William T. Harvey[5], Glenn Hickey[4], David A. Hillis[18], Savannah J. Hoyt[3], Hyeonsoo Jeong[5], Kaivan Kamali[1], Sergei L. Kosakovsky Pond[19], Troy M. LaPolice[1], Charles Lee[17], Alexandra P. Lewis[5], Yong-Hwee E. Loh[18], Patrick Masterson[16], Rajiv C. McCoy[6], Paul Medvedev[1], Karen H. Miga[4], Katherine M. Munson[5], Evgenia Pak[2], Benedict Paten[4], Brendan J. Pinto[20], Tamara Potapova[9], Arang Rhie[2], Joana L. Rocha[21], Fedor Ryabov[22], Oliver A. Ryder[23], Samuel Sacco[4], Kishwar Shafin[14], Valery A. Shepelev[24#], Viviane Slon[25], Steven J. Solar[2], Jessica M. Storer[3], Peter H. Sudmant[21], Sweetalana[1], Alex Sweeten[2,6], Michael G. Tassia[6], Françoise Thibaud-Nissen[16], Mario Ventura[7], Melissa A. Wilson[20], Alice C. Young[2], Huiqing Zeng[1], Xinru Zhang[1], Zachary A. Szpiech[1], Christian D. Huber[1], Jennifer L. Gerton[9], Soojin V. Yi[18], Michael C. Schatz[6], Ivan A. Alexandrov[25], Sergey Koren[2], Rachel J. O'Neill[3], Evan Eichler[5,26^], Adam M. Phillippy[2^]

*Authors contributing equally

#Currently retired

^Co-corresponding authors

1) Penn State University, University Park, PA, USA
2) National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA
3) University of Connecticut, Storrs, CT, USA
4) University of California Santa Cruz, Santa Cruz, CA, USA
5) University of Washington School of Medicine, Seattle, WA, USA
6) Johns Hopkins University, Baltimore, MD, USA
7) Università degli Studi di Bari Aldo Moro, Italy
8) University of Münster, Münster, Germany
9) Stowers Institute, Kansas City, MO, USA
10) MPI for Developmental Biology, Tübingen, Germany
11) Oregon Health & Science University, Portland, OR, USA
12) Oregon National Primate Research Center, Hillsboro, OR, USA
13) Penn State University School of Medicine, Hershey, PA, USA
14) Google, Mountain View, CA, USA
15) Foundation of Biological Data Sciences, Belmont, CA, USA
16) National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA
17) The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA
18) University of California Santa Barbara, Santa Barbara, CA, USA
19) Temple University, Philadelphia, PA, USA
20) Arizona State University, Tempe, AZ, USA

21) University of California Berkeley, Berkeley, CA, USA
22) Masters Program in National Research University Higher School of Economics, Moscow, Russia
23) San Diego Zoological Society, San Diego, CA, USA
24) Institute of Molecular Genetics, Moscow, Russia
25) Tel Aviv University, Israel
26) Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA

Abstract

Apes possess two sex chromosomes—the male-specific Y and the X shared by males and females. The Y chromosome is crucial for male reproduction, with deletions linked to infertility. The X chromosome carries genes vital for reproduction and cognition. Variation in mating patterns and brain function among great apes suggests corresponding differences in their sex chromosome structure and evolution. However, due to their highly repetitive nature and incomplete reference assemblies, ape sex chromosomes have been challenging to study. Here, using the state-of-the-art experimental and computational methods developed for the telomere-to-telomere (T2T) human genome, we produced gapless, complete assemblies of the X and Y chromosomes for five great apes (chimpanzee, bonobo, gorilla, Bornean and Sumatran orangutans) and a lesser ape, the siamang gibbon. These assemblies completely resolved ampliconic, palindromic, and satellite sequences, including the entire centromeres, allowing us to untangle the intricacies of ape sex chromosome evolution. We found that, compared to the X, ape Y chromosomes vary greatly in size and have low alignability and high levels of structural rearrangements. This divergence on the Y arises from the accumulation of lineage-specific ampliconic regions and palindromes (which are shared more broadly among species on the X) and from the abundance of transposable elements and satellites (which have a lower representation on the X). Our analysis of Y chromosome genes revealed lineage-specific expansions of multi-copy gene families and signatures of purifying selection. In summary, the Y exhibits dynamic evolution, while the X is more stable. Finally, mapping short-read sequencing data from >100 great ape individuals revealed the patterns of diversity and selection on their sex chromosomes, demonstrating the utility of these reference assemblies for studies of great ape evolution. These complete sex chromosome assemblies are expected to further inform conservation genetics of nonhuman apes, all of which are endangered species.

Introduction

Therian X and Y chromosomes are thought to have originated from a pair of autosomes approximately 170 million years ago (MYA)[1]. The X chromosome, typically present in two copies in females and one copy in males, has mostly retained the gene content and order from the original autosomal pair[2]. The Y chromosome, typically present in one copy in males only, has acquired the sex-determining gene *SRY*[3] and other male-specific genes and mutations that were fixed on the Y due to a series of inversions preventing recombination between the X and the Y over most of their lengths[4,5]. Because of this lack of recombination, the Y has contracted in size and accumulated deleterious mutations and repetitive elements. As a result, human X and Y differ substantially in size and gene content. The recent 'T2T' (gapless and complete) assemblies of these chromosomes revealed a human X of ~154 Mb harboring 796 protein-coding genes[6], and a human Y of only ~63 Mb harboring a mere 107 protein-coding genes[7]. In addition to the pseudoautosomal regions (PARs), where the Y still recombines with the X, and X-degenerate regions, which originated from the ancestral autosomal pair, the human Y has long ampliconic regions with extensive intrachromosomal homology. Within these ampliconic regions, the Y harbors numerous palindromes—long inverted repeats undergoing gene conversion between their arms, which counteracts the accumulation of deleterious mutations[8]. Similar to the human Y, the human X possesses PARs[6] and harbors several long palindromes[9]. Yet, whereas our understanding of the human sex chromosomes has greatly increased over the last few years, the complete sequence and structure of sex chromosomes in our closest relatives—non-human apes—have so far remained enigmatic despite their importance in informing human disease and evolution, and species conservation.

In general, therian sex chromosomes, particularly the Y, have been critically understudied. Both sex chromosomes have high repetitive element content (e.g., [6,7]) that, until very recently, has prohibited accurate sequencing and assembly[10]. Because male genomes have haploid sequences of both the X and the Y, most previous studies assembled female genomes, omitting the Y chromosome[10]. The Y chromosomes were sometimes sequenced and assembled via targeted methods[11–13] or long-read shotgun sequencing of male genomes, yet such assemblies were usually fragmented, collapsed, and incomplete[14,15]. The X chromosomes for several great apes have been deciphered to a greater level of contiguity (e.g.,[16,17]), but their assemblies, particularly for long satellite arrays, remained unfinished, preventing their complete characterization.

Earlier cytogenetic studies demonstrated a high level of lineage-specific amplifications, diversifications, and rearrangements leading to large size variation among great ape Y chromosomes (e.g.,[18] and reviewed in[19]). The initial assemblies of the human and chimpanzee Ys revealed remarkable differences in structure and gene content[11,12] despite divergence of only ~6 MY[20], and an acceleration of substitution rates and gene loss on the Y was observed in the common ancestor of bonobo and chimpanzee[15]. The human Y was found to be more similar in terms of alignability and gene content to the gorilla Y than to the more closely related chimpanzee Y[13]. The Y chromosome of the common ancestor of great apes had likely already possessed ampliconic sequences and multi-copy gene families[15], and all ape sex chromosomes share the same evolutionary strata[14]. Lineage-specific amplification and loss of ampliconic genes on the Y have been noted[14,15].

This progress notwithstanding, many questions about the evolution of great ape sex chromosomes have remained unanswered due to their incomplete assemblies. These include inquiries into the evolution of ampliconic regions and palindromes, satellites and heterochromatin, transposable elements (TEs), gene copy number, as well as segmental duplications and structural variants. Moreover, studies considering evolution of both X and Y chromosomes together have been lacking in apes. Utilizing the experimental and computational methodologies developed for the complete assembly of the human genome[7,21], we have deciphered the complete sequences of sex chromosomes from six ape species and studied the intricacies of their structure and evolution.

Results

Complete sex chromosome assemblies for all major great ape lineages

To perform a comparative analysis of great ape sex chromosomes, we built genome assemblies for most extant great ape species—chimpanzee, bonobo, western lowland gorilla (later called 'gorilla'), Bornean orangutan, and Sumatran orangutan. We also assembled the genome of an outgroup—the siamang, representing one of four gibbon genera of lesser apes. The assemblies included two pairs of closely related species: Bornean and Sumatran orangutans (*Pongo*), which diverged from each other ~0.5-1 MYA[22,23], and chimpanzee and bonobo (*Pan*), which diverged from each other ~1-2 MYA[24–27]. The human lineage diverged from the *Pan*, gorilla, *Pongo*, and gibbon lineages approximately 6, 7, 12, and 17 MYA[20,28], respectively (Fig. 1A). The studied species differ in their dispersal and mating patterns[29] (Table S1), potentially affecting sex chromosome structure and evolution.

We built a collection of male fibroblast and lymphoblastoid cell lines for these species (Table S2, Note S1, Note S2), each karyotyped (Fig. S1) to confirm absence of large-scale chromosomal rearrangements. From these cell lines, we isolated high-molecular-weight DNA for high-coverage Pacific Biosciences (PacBio) HiFi, Ultra-Long Oxford Nanopore Technologies (UL-ONT), and Hi-C sequencing (Methods). The sequencing depth among samples ranged from 54-109× for HiFi, 28-73× for UL-ONT, and 30-78× for Hi-C (Table S3). We also had access to parental DNA for the studied bonobo and gorilla individuals (Table S4), which was sequenced to 51-71× depth with Illumina short-read technology (Table S3).

Genome assemblies were generated with Verkko[30], which uses PacBio HiFi reads to create a diploid assembly graph that is further resolved using alignments of the ONT reads to the graph. Haplotypes were then resolved using either parental-specific *k*-mers when trios were available or Hi-C binned assemblies in the absence of trios (Methods). Components of the assembly graph representing the X and Y chromosomes were identified based on graph topology and confirmed by alignment to known reference sequences (Methods; Fig. S2). In all cases, the sex chromosomes were fully separated from the autosomes by Verkko, and several X and Y chromosomes were determined to be complete. The remaining sex chromosomes were finished via manual curation of the assembly graphs and validated using several approaches (Table S5; Methods), while the autosomes were left as draft assemblies for future analyses. Below we focus on the analysis of sex chromosomes from these assemblies (version 1.1).

New assemblies gain sizeable proportion of sequences, including challenging regions

Altogether, we generated new, T2T assemblies for siamang and Bornean orangutan X and Y chromosomes, for which prior assemblies were unavailable, and T2T assemblies for bonobo, chimpanzee, gorilla, and Sumatran orangutan X and Y chromosomes, for which lower-quality assemblies were available[15–17,31] (Fig. 2), and resolved their palindromes, centromeres, and subtelomeric satellite repeats for the first time (see below). On the Y, new sequences accounted for 24–45% of the total chromosome length (8.6–30 Mb of sequence, Table S6). On the X, new sequences accounted for 2.6–16% of the total chromosome length (3.9–28 Mb of sequence, Table S6). For example, we added 30 Mb of sequence for the Sumatran orangutan Y and 28 Mb of sequence for the gorilla X. The sequences gained in the new assemblies had a high frequency of motifs able to form non-canonical DNA structures, or non-B DNA (Fig. 2; $p<2.2\times10^{-16}$ for logistic regressions in each species with previous assemblies, Table S7), which are known to be problematic sequencing targets[32]. Combining sequencing technologies, as was done here, is expected to remedy sequencing limitations in such regions[6,7,32].

Immense variation in size and low alignability of great ape Y chromosomes

The overall variation in size was larger among the Y than among the X chromosomes across the studied ape species (including the previously assembled reference human X and Y[6,7], Fig. 2). Ape Y chromosomes ranged from 30 Mb in siamang to 68 Mb in Sumatran orangutan. The X chromosomes ranged from 154 Mb in chimpanzee and human to 178 Mb in gorilla. For pairs of closely related species, the Y chromosomes differed in length (19 Mb between the two orangutans and 11 Mb between bonobo and chimpanzee) more so than the X chromosomes (1.5 Mb between the two orangutans and 6.3 Mb between bonobo and chimpanzee).

Across all pairwise species comparisons, the percentage of sequence aligned was lower for ape Y than ape X chromosomes (Fig. 1B). For example, only 14–27% of the human Y was covered by alignments to the other ape Y chromosomes, whereas as much as 93–98% of the human X was covered by alignments to the other ape X chromosomes. The same pattern was observed for closely related species, with only 60–87% of the Y chromosome, but >95% of the X chromosome, aligned between them. Taken together, these observations suggest that, during evolution, the Y chromosome experienced a greater degree of sequence turnover compared to the X chromosome.

A new PAR2 in bonobo and high structural variation on the Y chromosomes

By analyzing sequence similarity between the X and the Y of the same species, we identified PARs (Fig. 1C; Table S8; Methods), which undergo recombination and thus differ only at the haplotype level between the two sex chromosomes. All species possessed a homologous 2.1–2.5-Mb PAR1, but independently acquired PAR2 sequences were identified in human and bonobo. The PAR2 is ~330 kb long in human[7] and ~95 kb in bonobo (this study), and they are not homologous (Note S3).

We found that, in the sequences experiencing interspecies variation, there were more base pairs affected by large-scale structural variants (SVs) than by single-nucleotide variants (SNVs) or small insertions/deletions on the X (83–86%) and particularly on the Y chromosomes (99%; not considering PARs; Fig. 1C, Fig. S3, Fig. S4; Table S9; Methods). Inversions were abundant on the Y (10–30% SVs by length), consistent with its palindromic architecture. Inversions and insertions were ~8-fold and 3-fold longer on the Y than the X, respectively (12.1 Mb vs. 1.5 Mb, and 38.2 kb vs. 11.9 kb, respectively; $p<2.2\times10^{-16}$, Wilcoxon ranked-sum tests). The number of SVs positively correlated with the lengths of phylogenetic branches (Fig. S5, Table S10), with a higher slope for the Y (15.8 SVs/Mb/MY) than for the X (6.1 SVs/Mb/MY), indicating a more rapid accumulation of SVs on the former than on the latter. Thus, SVs represent one of the dominant types of genetic variation on the Y chromosome.

We identified SVs with potential functional significance in the human lineage. We cataloged 334 human-specific SVs on the Y (309 insertions, 13 deletions, and 12 inversions) and 1,711 of such SVs on the X (1,339 insertions and 372 deletions; Additional File 1) and studied their overlaps with genes. On the Y, we detected an 80-bp deletion disrupting the first exon of the *DAZ4* gene, and an insertion of the previously reported 3.6-Mb X-transposed region (XTR, a human-specific duplication from the X to the Y[11]) including 13 genes (Table S11). Outside of gene copy-number changes, human-specific inversions on the Y were associated with 11 genes (Table S11). The human-specific insertions and deletions affected 23 genes on the X (Table S11).

The Y and the X differ in nucleotide substitution rates and spectra

The phylogenetic analysis of multi-species alignments for the X chromosome, and separately for the Y chromosome (Methods), revealed the expected species topology (Fig. 1A). After removing PARs, we detected

higher substitution rates on the Y than on the X for all branches of the phylogeny (Fig. 1D), consistent with male mutation bias[33–35]. For instance, the human-chimpanzee divergence was 2.68% on the Y and 0.97% on the X. For the Y, we detected a 10.7–11.0% acceleration of substitution rates in the *Pan* lineage and a 9.2% slowdown in the *Pongo* lineage, as compared to substitution rates in the human lineage ($p<10^{-5}$, relative rate tests; Fig. 1D, Table S12). The substitution rates were more similar in magnitude among the branches for the X than for the Y (Fig. 1D, Table S12). These results indicate a stronger male mutation bias for the *Pan* lineage and a weaker bias for the *Pongo* lineage, as compared to that for the human and gorilla lineages. Strong male mutation bias in bonobo and chimpanzee is consistent with greater sperm production due to a stronger sperm competition in these species than in other great apes[29].

Comparing nucleotide substitution spectra between the two sex chromosomes (excluding PARs, Fig. 1E), we found C>A, C>G, T>A, and T>G substitutions to be significantly more abundant on the Y than on the X and C>T and T>C substitutions to be more abundant on the X than on the Y. These findings are broadly consistent with sex-specific signatures of *de novo* mutations from other studies. C>A, C>G and T>G were shown to be enriched in paternal *de novo* mutations, whereas C>T mutations—in maternal *de novo* mutations[36]. C>G might be related to meiotic double-strand breaks in the male germline[37].

Rapid evolution of ampliconic regions and palindromes on the Y but not on the X

To study evolution of sex chromosomes outside of the PARs in more detail, we separated X chromosome assemblies into X-ancestral, ampliconic, and satellite regions; and Y chromosome assemblies into X-degenerate, ampliconic, X-transposed, and satellite regions (Methods; Fig. 2; Table S13; Additional File 2). The X-ancestral regions on the X, which are the remnants of the autosomal past of this chromosome, ranged from 138–147 Mb among species. The corresponding regions on the Y—the X-degenerate regions—were much shorter, from 3.6–7.5 Mb, consistent with sequence loss on the Y due to the lack of recombination. We did not find XTRs[11] on the Y chromosomes of nonhuman apes (Note S4).

Ampliconic regions, defined as long sequences present multiple times on the same chromosome (Fig. 2, Table S14; Methods), ranged from 3.8–6.9 Mb on the X, but were longer on the Y, where they ranged from 9.7–28 Mb, greatly contributing to its variation in length among species (Table S13). Notably, the length estimates of these regions were lower in previous Y chromosome assemblies[12,15] than in our T2T assemblies (a difference of 9.1 Mb, 2.5 Mb, 8.0 Mb, and 25 Mb for bonobo, chimpanzee, gorilla, and Sumatran orangutan, respectively), suggesting that they were partially collapsed in the former. Ampliconic regions on the X were shared among species to a large degree (Fig. 3A); for instance, we could detect their homology not only among closely related species, but also among the African great apes. In contrast, we could detect appreciable similarity between Y ampliconic regions only in pairs of closely related species—bonobo and chimpanzee, and Bornean and Sumatran orangutans (Fig. 3B)—yet these Y ampliconic regions still differed in organization (Fig. 3B; Fig. S6), suggesting extremely rapid evolution.

Within ampliconic regions, we located abundant palindromes (i.e. long inverted repeats separated by a spacer; Fig. 2; Additional File 3; Methods). Palindromes on the Y were two to three times longer and had higher density than on the X (Fig. S7A), consistent with their role in rescuing deleterious mutations on the Y through intrachromosomal recombination and gene conversion[4,8]. We found shorter spacers to be associated with higher sequence identity between the arms (moderate Spearman correlation coefficients, Fig. S7C, Fig. S7E; Table S15), consistent with more efficient gene conversion for sequences located closer to each other on the chromosome[38]. We detected an increase in arm length to be associated with increased sequence identity between the arms (Fig. S7D, Fig. S7F; Table S15), suggesting that longer palindromes were more likely to undergo gene conversion, and consistent with the high length of gene conversion tracts in palindromes[39]. We found a negative correlation between arms' GC content and spacer length, consistent with GC-biased gene

conversion, on the X, but not the Y (Fig. S7B, Table S15). Compared with those on the Y, palindromes on the X have higher GC content (p=3.83x10^{-15}, two sample t-test, Fig. S7H), providing more donor sites for GC-biased gene conversion.

Palindromes on the X displayed a high degree of conservation among species (Fig. 3C; Table S16). For instance, 27 out of a total of 28 palindromes were shared between bonobo and chimpanzee; 21 among African great apes; and 10 among all great apes. Even the more distantly related siamang shared six out of its 26 palindromes with the other species. Palindromes on the Y displayed a lower degree of conservation, with only closely related species sharing a substantial number of palindromes (Fig. 3D). Only two palindromes were shared among all African great apes, and none were shared among all great apes and between great apes and siamang. Many species-specific palindromes were detected on the Y, and even the closely related Bornean and Sumatran orangutans each acquired three and four new palindromes, respectively.

We estimated (Methods) that 22.8–55.9% of the length of non-human ape Y chromosomes consists of segmental duplications (SDs), compared to only 4.0–7.2% of the X chromosomes (Fig. 1C, Table S17). The *Pan* and *Pongo* lineages each showed two times higher percentage of their Y chromosomes occupied by SDs compared to the other ape lineages (Table S17). We found little evidence of lineage-specific SDs on the X (Table S17). Nevertheless, we observed a gain of 1.9 Mb, 0.8 Mb, and 2.2 Mb of interchromosomal SDs in the T2T vs. previous X chromosome assemblies[16,17,31] in bonobo, gorilla, and Sumatran orangutan, respectively.

Repetitive element composition changes quickly on the Y but is more stable on the X

We next produced comprehensive repeat annotations for both X and Y chromosomes across the ape lineage by integrating a combination of known repeats, repeat models identified in the analysis of human CHM13[40,41] and human T2T-Y[7], and *de novo* repeat curation (Table 1). Through these analyses, we identified 33 previously unknown satellites and characterized two variants of *DXZ4* repeats (Fig. S8, Table S18) and 13 previously uncharacterized composite repeats (Fig. S9, Table S19, Table S20). In total, 71–85% and 62–66% of Y and X chromosome length consisted of repeats, respectively (Fig. 4A, Table 1), compared to only 53% of the human T2T autosomal length[41]. PARs maintained a consistent repeat content and distribution both within a species and across the ape lineage (Ext. Data Fig. 1, Fig. S10, Table S21), including a high content of SINEs, *Alu*s in particular. High percentages and variable repetitive DNA content (Table 1, Table S22, Table S23, Table S18) and distributions (Ext. Data Fig. 1, Fig. S10, Table S21) were observed on the Y chromosomes of each species, substantially contributing to their length variation. While Y chromosome repeats were predominantly composed of satellites and simple/low complexity repeats (Fig. 4A), the TE content was significantly higher in the chrY X-degenerate than ampliconic regions (~65.6% vs. ~46.8%; p<0.001, Mann-Whitney U test; Fig. S10, Table S21), reflecting the absence of recombination in the former regions and frequent intrachromosomal recombination in the latter (recombination serves to remove many TEs[42]).

Chromosome X had a more consistent repeat content in each ape assembly (Fig. 4A, Table 1), composed mainly of retroelements (Table S22, Table S23, Table S18). It is highly enriched in L1s, consistent with their role in X chromosome inactivation[43]. The distributions of repeats along the X chromosomes were strikingly similar among species (Ext. Data Fig. 1), with notable exceptions in the expansion of alpha satellite arrays in non-centromeric regions in siamang[44,45], of the SAR/HSat1A satellite in non-human African apes, and of subtelomeric arrays of the pCht/StSat satellite in gorilla[46] (Fig. 4A, Ext. Data Fig. 1). The average TE content of X-ancestral regions on the X was significantly lower than that for X-degenerate regions on the Y (~59.3% vs. ~65.6%; p<0.001, Mann-Whitney U test; Fig. S10, Table S21) and significantly higher than that for Y ampliconic regions (~46.8%; p<0.001, Mann-Whitney U test), consistent with different recombination rates among these regions. The average TE content of ampliconic regions was significantly lower on the Y than on the X (~46.8% vs. ~53.8%; p<0.001, Mann-Whitney U test), suggesting more frequent recombination in such

regions on the Y.

Newly defined satellite sequences accounted for an average of 317 kb and 61 kb on each X and Y chromosome, respectively (Fig. S8), and many of them expanded in a lineage- and/or chromosome-specific manner (Table S18). For example, the bonobo-specific satellite Ariel flanked the PAR2 regions in an array with 318 units on the X and 134 units on the Y (Note S3). Similarly, the *Pan*-specific satellite Francisco, present on the X (and not on the Y), is found at a copy number >6 times higher in chimpanzee (4,277 copies) than in bonobo (686 copies). In general, variable TE types and satellite arrays expanded in a lineage-specific (LS) manner (Fig. 4A, Ext. Data Fig. 1, Table S24, Table S25), with the Y contributing more interspecies variation than the X (Fig. 4A). On average, a total of 5.72% of the Y chromosome was derived from LS repeats. The X and Y LS insertion patterns between closely related species were largely shared (Note S5).

The T2T assemblies of ape sex chromosomes have for the first time allowed us to explore the chromosomal distribution of motifs able to form non-B DNA structures—A-phased repeats, direct repeats, G-quadruplexes (G4s), inverted repeats, mirror repeats, short tandem repeats (STRs), and Z-DNA[47]—which have been implicated in numerous cellular processes, including replication and transcription[48]. Such motifs (Methods) covered from 6.3–8.7% of the X and from 10–24% of the Y (Methods; Table S26). Each non-B DNA motif type usually occupied a similar fraction (Table S26) and frequently was located in similar regions of the X chromosomes among species, with direct repeats frequently located at the subtelomeric regions and inverted repeats at the centromeric regions (Fig. S11). In contrast, the Y chromosomes exhibited a wide range of variation in content and location of different non-B DNA types (Table S26, Fig. S11). For instance, direct repeats expanded in the subtelomeric regions of the gorilla Y.

Non-B DNA was frequently enriched at satellites, suggesting their functional roles (Fig. S12, Table S27). For example, G4s were enriched at Gamma-satellite II (4.3–23-fold compared to the sex-chromosome average), where, due to their undermethylation[49], they might be important for keeping a transcriptionally permissive chromatin conformation previously demonstrated for this satellite[50]. The LSau satellite[51] exhibited a strong (16–27-fold) overrepresentation of G4s as well, where they might also function as mediators of epigenetic modifications[52] consistent with variable methylation levels at this satellite among apes[53]. The ACRO1 satellite[21] was enriched in A-phased repeats (2.9–7.5-fold) and G4s (2.8–6.2-fold), where the latter might promote genomic instability[54], frequently evident at the tips of acrocentric chromosomes. Across species, we also observed an enrichment of inverted repeats at alpha satellites, consistent with the suggested role of non-B DNA in centromere formation[55].

Lineage-specific evolution of sex-chromosome centromeres

We next uncovered the evolution of centromeres on ape sex chromosomes from these T2T assemblies (Methods). It was previously proposed that primates had experienced successive cycles of centromere remodeling where a new variant of alpha satellite (AS, with 171-bp repeat units) had emerged and expanded in the midst of the centromeres of all (or almost all) chromosomes of a progenitor species giving rise to a new taxon with centromeres different from the ones in the pre-existing taxa[56,57]. Thus, each major primate phylogenetic branch would have active centromeres corresponding to a different AS suprachromosomal family (SF). It was also envisioned that the vestigial layers of old, inactive centromeres, which flank the new, active centromere in the new taxon, would have degraded and shrunk (Fig. 4B)[58,59]. Consistent with these predictions, the active X chromosome centromere (cenX) was formed by ‘young’ SF1–3 in African apes, ‘older’ SF5 in orangutans, and yet older SF4 in the siamang. CenX was flanked by the older SF vestigial layers in all primates studied (e.g., by SF5, SF4, and SF6–11 in African apes; Fig. 4C) which testifies to its stable position in evolution. In contrast, cenY was almost devoid of the flanking layers (Fig. 4C) presumably because its position often changed in different primate lineages. Consistent with cenYs in *Homo* and *Pan* being an

exception to the expected pattern[7,60], we found their active arrays to be formed by older SF4 instead of younger SF1–3. However, such 'lagging' cenYs were not a common feature of other ape Y centromeres, which followed the expected pattern.

HORs (higher-order repeats, strings of AS monomers which reiterate with high identity; Table S28, Table S29, Note S6, Methods) appeared to be genus-specific except for cenX in *Homo* and *Pan*. The general architecture of the centromere was similar albeit not identical in two pairs of closely related species (Ext. Data Fig. 2, Fig. S13A–B). The chimpanzee active cenY array and adjacent non-AS regions were in reverse orientation as compared to bonobo (and humans; Note S6). The same HORs formed active arrays on the X, and also on the Y, in both species pairs; however the sizes of active arrays, their StV content (monomer order alteration in AS HORs[58]; Table S29), and the CDR positions (Centromere Dip Region; a signature methylation pattern that marks the kinetochore location[59,61]) within active arrays often differed between species (Fig. S13B, Note S6). All major HOR haplotypes (HORhaps[58,59]) were species-specific, indicating that the HORhap identity of a centromere had changed since any two species shared a common ancestor. It is presumed these changes had occurred through cycles of remodeling in the same manner as changes of SFs and HORs among more distantly related taxa, but more subtly (Ext. Data Fig. 2, Note S6).

SF1 expanded in cenX and cenY in the gorilla branch, and gorilla cenY contained a single active array made by new SFs (corresponding to a high density of CENP-B functional sites; Fig. S13A). In gorilla cenY, SF1 inserted into an AS-containing palindromic SD (Fig. S13G, Note S6). Gorilla also had an inverted cenX core (but not flanks) relative to human and *Pan*, with both inversion breakpoints in SF3 AS (Fig. S13E). The gorilla cenX appeared to have experienced three consecutive changes of centromere identity, one of which (SF01 to SF3) occurred in the common ancestor of African apes, and two others (SF3 to SF2 and SF2 to SF1) that were unique to gorilla (Fig. 4C, Fig. S13E, Note S6). Siamang centromere and telomere AS arrays[44] were composed of different SF4 ASs, where all three non-telomere arrays (one in cenX and two in cenY) were different from each other and all three telomere arrays (two in chrX and one in chrY) were similar to one another (Fig. S13F); therefore, siamang likely has chromosome-specific centromeres, similar to the other apes[44,57,62].

Ribosomal DNA repeats are present and active on some Y chromosomes

Ribosomal DNA (rDNA) arrays were present in the Y chromosome assemblies of siamang and Sumatran orangutan, consistent with prior literature[63,64], but not on the other sex chromosomes analyzed, suggesting frequent acquisition of rDNA during Y chromosome evolution. Individual UL-ONT reads encompassed the entire rDNA array with three copies on the Sumatran orangutan Y chromosome. In the assembly graph of the siamang Y chromosome, we identified a longer rDNA array that no individual UL-ONT reads could completely span. Using FISH coupled with a *k*-mer-based estimate of the total copy number derived from short-read sequencing data (Fig. 5, Methods), we estimated the copy number on the Y chromosome to be three in Sumatran orangutan, in agreement with the assembly, and ~16 in siamang (Fig. 5B, Fig. S14, Table S30). No rDNA signal was detected on the human, chimpanzee, gorilla, and Bornean orangutan Y chromosomes (Fig. 5A). Extending FISH analysis beyond the assembled genomes (Note S7), we confirmed these patterns and found rDNA on the Y chromosomes of white-cheeked and black crested gibbons. There were no rDNA repeats found on the X chromosomes in male karyotypes, making the Y copies sex-linked with the potential to increase dosage in males relative to females.

The transcriptional activity of the rDNA arrays on the Y chromosome in the Sumatran orangutan and siamang was confirmed by two methods. First, we analyzed the CpG methylation profiles in UL-ONT reads. While the intergenic spacer (IGS) is methylated regardless of activity status, the 45S rRNA gene is thought to be unmethylated when active and transcribed and methylated when inactive[65]. The three 45S genes on the Y chromosome in Sumatran orangutan, as well as the genes at the edges of the array on the siamang Y, were

unmethylated, indicating active transcription (Figs. 5E–F). Second, we used immunoFISH[66] with an antibody to the UBF transcription factor that binds to actively transcribed 45S genes. We did not detect the UBF signal on the Sumatran orangutan Y, likely because the signal falls below the detection limit; however, UBF was detected on the siamang Y (Figs. 5C–D), consistent with active RNA PolI transcription[67].

Protein-coding gene evolution

Our gene annotations (Table S31; Methods) indicated the presence of a high percentage of BUSCO genes on the X chromosomes (Table S32), and of most previously defined Y chromosome X-degenerate and ampliconic genes (Fig. 6). We manually curated Y chromosome genes (Methods) and validated the copy number of Y ampliconic genes with droplet digital PCR (ddPCR; Table S33, Table S34). On the X, gene density was ~2.5–5-fold higher in the ampliconic than X-ancestral regions (16–25 vs. 5.3–6.1 genes/Mb, respectively; Fig. 6A; Table S35). Gene density was higher still in palindromes (20-32 genes/Mb; Fig. 6A), particularly in palindromes shared among species (29–47 genes/Mb). Many genes located in such palindromes were housekeeping (e.g., *TMLHE*, *CENPVL1*, *H2AB2*, and *FAM156*; Fig. 3C; Table S35). Gene density was uniformly lower on the Y than on the X (Fig.6A), with a low density in both X-degenerate (2.0–4.5 genes/Mb) and ampliconic (2.8–5.7 genes/Mb) regions. We detected duplication and copy number expansion of autosomal genes, to Y palindromes of some genera (e.g., *KRT8* in Pan and *PTPN13* in gorilla; Table S35). Otherwise, Y palindromes largely contained previously known ampliconic genes (Fig. 3D; Table S35).

Ampliconic genes on the Y chromosome are present in multi-copy gene families, comprising *BPY2*, *CDY*, *DAZ*, *HSFY*, *PRY*, *RBMY*, *TSPY*, *VCY*, and *XKRY*[11] (Fig. 6B), products of which function in spermatogenesis (Table S35). Using detailed annotations of Y ampliconic genes in our T2T assemblies (Methods), we discovered a high overall rate of copy number changes, as well as episodes of significant lineage-specific gene family expansions and contractions (Fig. 6B, Note S8). For example, *RMBY* expanded in bonobo and contracted in chimpanzee and Sumatran orangutan; *CDY* expanded in Sumatran orangutan and contracted in Bornean orangutan; and *TSPY* underwent several significant copy number changes with a notable expansion in the human lineage. These results based on a single individual per species are largely consistent with prior ddPCR data generated for multiple individuals per species[68]. We discovered that some ampliconic genes have duplicated to multiple locations on the Y; in some cases they are present at several palindromes and/or located outside of palindromes (Table S36). We found no evidence of positive selection acting on ape Y ampliconic genes, as the nonsynonymous-to-synonymous rate ratio, d_N/d_S, was below 1 for all of them (Table S37). A significant signal of purifying selection, with d_N/d_S significantly less than 1, was detected for only three (*CDY*, *HSFY,* and *RBMY*) out of eight genes analyzed ($p \leq 0.05$, likelihood ratio tests, LRTs; Table S37).

TSPY, which is the only ampliconic gene family predominantly located in tandem arrays outside of palindromes (in all species but bonobo and siamang, Table S36), had a high copy number (≥18) in all species studied but gorilla and siamang (6 and 5 copies, respectively). A phylogenetic analysis of the protein-coding copies of *TSPY* identified species- and genus-specific clades (Ext. Data Fig. 3), suggesting either recent diversification or sequence homogenization. Independent diversification of gene copies predicts variable divergence points among genera, as well as copies shared among genera—neither pattern was observed in our data. Instead, our results are more consistent with the sequence homogenization hypothesis because all genus-specific clades had approximately the same divergence time, suggesting that they formed due to sequence homogenization with similar rates. Such homogenization might occur due to homologous recombination between palindrome arms (in bonobo and siamang) and between direct repeats[69].

While X-degenerate Y-linked gene content was generally well conserved, there were some notable exceptions (Fig. 6B, Note S9). X-degenerate Y-linked genes are those genes that do, or did, have functional copies on both the ancestral X and Y. We found that *TXLNGY*, for example, had pseudogenized in all studied apes,

despite originally being annotated as a functional gene in humans[11], potentially due to its expression in humans[70]. Similarly, *MXRA5Y* and *PRKY* were lost completely in siamang and both orangutans, and pseudogenized in nearly all African apes. Only three other genes showed evidence of pseudogenization or complete loss in at least one studied taxon; *USP9Y*, *TBL1Y*, and *EIF1AY*. The other 10 X-degenerate Y-linked genes were conserved in their presence across all studied apes suggesting their functional importance. Consistent with this prediction, we found evidence of purifying selection (with d_N/d_S significantly below 1, $p \leq 0.05$, likelihood ratio, LRT) for nine out of 13 genes analyzed (Table S37). We also found signal of branch heterogeneity for d_N/d_S for several genes, and identified stronger purifying selection ($p \leq 0.05$, LRT, Holm-Bonferroni correction) in the orangutan and human-gorilla branches for *USP9Y* (d_N/d_S=0.15 and d_N/d_S=0.17, respectively, vs. tree-average d_N/d_S=0.37) and a potential relaxation of purifying selection in the common ancestor of chimpanzee and bonobo for *NLGN4Y* (d_N/d_S=0.71 vs. tree-average d_N/d_S=0.15). Consistent with an observation for human and macaque genes[4], X-degenerate genes had a lower group-mean d_N/d_S as compared to that for ampliconic genes (0.38 vs. 0.65, joint model fit, LRT p-value$<10^{-10}$)

Previous studies indicated that *de novo* genes, defined as species- or genus-specific genes arising from noncoding sequences, play a role in fertility and frequently have testis-specific expression[71]. Thus, the Y chromosome is a promising candidate for *de novo* gene emergence. Using the new T2T assemblies, we were able to trace the emergence of two candidate *de novo* genes specific to ape Y chromosomes—one in bonobo and one in siamang (Note S10).

Divergent methylation patterns between the X and the Y chromosomes

Using long-read data mapped to these T2T assemblies, we analyzed 5mC DNA methylation (hereafter 'methylation') patterns across ape sex chromosomes—the Y and the active X, as our data originated from male cell lines. DNA methylation is functionally implicated in the regulation of gene expression, development, cell differentiation, and disease[72], yet it has been understudied for the sex chromosomes, especially for the Y. In females, the inactive X chromosome, which is transcriptionally less active and more heterochromatic than the active X, tends to have lower methylation than the active X chromosome[73–76]. We thus hypothesized that the Y chromosome, given its relative transcriptional inactivity and increased heterochromatin content, may have lower methylation than the active X. In line with this expectation, the Y (excluding PARs) exhibited lower methylation levels than the X (also excluding PARs; significant Wilcoxon rank-sum tests in all species but chimpanzee, p-values in Table S38; Ext. Data Fig. 4A).

We predicted high methylation levels at PARs because methylation had been previously noted to be elevated in regions with high recombination rates[77,78]. Consistent with this prediction, methylation was higher for PAR1 than the rest of the X chromosome in all species (Wilcoxon rank-sum test, p-values in Table S38; Ext. Data Fig. 4A). We did not detect significant methylation differences between each PAR2 and the rest of the X chromosome (Fig. S15A), possibly due to the recent emergence of PAR2. Ampliconic regions on the X and Y undergo intrachromosomal recombination and thus we expected them to be highly methylated. Methylation levels were indeed significantly higher in ampliconic than X-ancestral regions in chimpanzee, human, and Bornean orangutan X chromosomes (Wilcoxon rank-sum tests, p-values in Table S38, Ext. Data Fig. 4), but they were not significantly different between these two regions on the X of other species, and were in fact significantly lower in ampliconic than X-degenerate regions on the Y (Ext. Data Fig. 4). Thus, the relationship between methylation and recombination might be different for intra- vs. interchromosomal recombination.

We found that most groups of TEs, as well as genic sequences, follow the general pattern of highest methylation in PAR1, intermediate in non-PAR X, and lowest in non-PAR Y (Ext. Data Fig. 4B, p-values in Table S38). This same general pattern was observed in satellite regions (with the exception of human, which showed non-significant trends), despite their recent and frequent lineage-specific expansions. These patterns

suggest rapid evolution of methylation patterns on ape sex chromosomes. We observed sharp decreases of methylation levels near the transcription start sites for protein-coding genes (Fig. S15B-C), indicating the importance of promoter hypomethylation in the regulation of gene expression[79,80] on the sex chromosomes. Promoter methylation levels and the rank of gene expression levels were negatively correlated in all species, with a particularly strong pattern on the X chromosome, which harbors a large number of genes (Fig. S15D).

Improving the study of intraspecific ape diversity and selection

Our T2T assemblies enabled the first X and Y chromosome-wide analyses of great ape intraspecific diversity. Aligning short sequencing reads from 129 individuals across 11 subspecies (Table S39) to T2T and previous assemblies (Methods), we found a higher proportion of reads mapping to the former than the latter assemblies (Ext. Data Fig. 5A). In most cases, we observed lower mismatch rates between reads and references for the T2T vs. previous assemblies (Fig. S16A), demonstrating the superiority of the T2T assemblies as new reference genomes. After calling SNV and small indel variants considering either T2T or previous assemblies (Fig. S16B), we found that the new variant data set contained fewer homozygous variants, which can arise from structural errors in the reference genome[81,82], and largely restored the expected site frequency spectrum (SFS; Ext. Data Fig. 5B). However, eastern lowland and mountain gorillas still contained a substantial number of homozygous variants (Fig. S16C), highlighting the need for additional species- and subspecies-specific references. A detailed comparison on the chimpanzee Y revealed more variants due to the increased length and more uniform read distribution of the T2T assembly (Ext. Data Fig. 5C). In another example, we found a 33-fold reduction in variants in a segment of an ampliconic region on gorilla Y (Ext. Data Fig. 5D), likely due to a collapse of this segment in the previous assembly.

Across the X chromosome, the nucleotide diversity[83] of Sumatran orangutans was higher than that of Bornean orangutans (Ext. Data Fig. 5E), in agreement with prior work indicating a steady population decline and low effective population size in Bornean orangutans[84]. In the *Pan* lineage, central chimpanzees retained the highest genetic diversity, followed by eastern chimpanzees. Nigeria-Cameroon and western chimpanzees retained a relatively low diversity, potentially signaling historical population bottlenecks[85]. Gorillas exhibited a pattern similar to that of chimpanzees, with the western lowland gorillas retaining a higher genetic diversity than the eastern lowland and mountain gorillas, both of which have undergone a prolonged population decline[86]. In most subspecies studied, the Y exhibited a substantially lower diversity than the X (Ext. Data Fig. 5E), in line with previous studies in humans[87,88]. Among the great apes, bonobos displayed the highest nucleotide diversity on the Y. Within their respective species, the western lowland gorillas and central chimpanzees contained the highest Y chromosome diversity, consistent with the observations on the X. The diversity in PARs was higher than that on the X likely due to a higher effective population size, but in most cases had patterns similar to the ones observed for the X.

Of particular interest was putative selection on the Y, which can evolve rapidly due to different levels of sperm competition among ape species[11] (Table S1). To gain power, we combined all chimpanzee and, separately, all gorilla samples. Neutral expectations of nucleotide diversity[83] and Tajima's D values[89] (Note S11) were derived by simulating previously inferred demographic models for chimpanzees[85] and gorillas[90] under sex-chromosome-specific mutation rates and a range of male-to-female effective population size ratios. In gorillas, the observed Y/X diversity ratio was considerably lower than our simulated values even for very low male effective population sizes (Note S11). Similarly, in chimpanzees, the observed Y/X diversity ratio was inconsistent with neutrality except for very low male effective population size ($N_m < 0.25\ N_f$). Because male effective population size is reported to be high in chimpanzees[91], this suggests selection reduced diversity on the Y chromosome in both species, consistent with reports for humans[92]. The results of Tajima's D test suggested that purifying (i.e., background), rather than positive, selection predominantly drives the reduction in

diversity on the Y in both species (Note S11). We further analyzed diversity data on the X chromosomes to identify candidate regions of selection on the X (Note S11). Finally, incorporating diversity information, we evaluated selection acting on Y X-degenerate genes in chimpanzee and gorilla (Note S12). We found evidence of positive selection for *UTY,* a gene that has been implicated in immunity[93,94], in the chimpanzee lineage (q-value <0.05), which should be explored in further studies.

Discussion

Our complete, telomere-to-telomere assemblies have revealed the evolution of the great ape X and Y chromosomes in unprecedented detail. In sharp contrast to the X, the Y has undergone extremely rapid evolution in all ape species we studied—not just bonobo and chimpanzee, as was previously suggested[12,15]. The Y has experienced elevated rates of nucleotide substitutions, intrachromosomal rearrangements, and segmental duplications, likely due to a loss of recombination over most of its length. It also has reduced levels of DNA methylation, linked to the low expression levels observed for many Y chromosome genes[68].

Why, then, is the Y chromosome still present in apes despite constantly accumulating potentially deleterious mutations and repeats? Several studies have suggested that Y chromosomes are on their way towards being lost in mammals (e.g., [95,96]), but our analysis demonstrates that the Y harbors several protein-coding genes evolving under purifying selection in apes. Purifying selection was also suggested for Y chromosome genes in rhesus macaque[97]. Future studies are needed to investigate non-coding genes and regulatory elements on the Y, which may be essential for males and further contribute to selective pressure.

Palindromes in particular are thought to be critical for counterbalancing the degradation of the non-recombining Y chromosome[98] by enabling intrachromosomal recombination and gene conversion[9]. Thus, one might envision Y palindromes to be highly conserved, but we found that the sequence of these palindromes changes quickly across species. Rapid acquisition of new palindromes on the Y in different ape species might be due to random genetic drift, which is expected to be strong on the Y because of its small effective population size[98], and/or due to species-specific selection. Our analysis of ampliconic genes on the Y, which are primarily located in palindromes and play a role in spermatogenesis, did not provide evidence of species-specific selection. Instead we found a higher ratio of nonsynonymous-to-synonymous mutations for ampliconic vs. single-copy genes. This pattern is consistent with either relaxation of functional constraints or a higher rate of fixation of site-specific beneficial mutations due to gene conversion in ampliconic genes[4]. Future analyses of additional T2T primate genomes should distinguish between these possibilities. Notably, copies of some Y ampliconic genes were present at multiple locations on the Y, and not just within a single palindrome or tandem repeat. Such genetic redundancy represents an additional mechanism safeguarding genes on the non-recombining Y chromosome. The X chromosome also undergoes less recombination than the autosomes because, outside of PARs, it does not recombine in males. We find that it has utilized some of the same strategies to preserve its genetic content, including the presence of a large number of palindromes with a high density of housekeeping genes in all primates studied.

In addition to gene amplifications, a variety of lineage-specific satellite expansions were observed in the apes, with some specific to the Y (e.g. SAR/HSAT1A accumulation in gorilla Y) and some shared between X and Y (e.g. StSat/pCht expansions in gorilla and alpha satellite expansions in siamang). These observations prompt a question about the potential functionality of these satellites. Satellites on the *Drosophila* Y chromosome were shown to contribute to regulation of gene expression of autosomal genes[99]; it will be important to investigate whether the same phenomenon exists in apes. Moreover, divergence of sex chromosome satellite arrays was linked to reproductive isolation among *Drosophila* species[100]. The StSat/pCht repeats have been proposed to participate in telomere metabolism and meiotic telomere clustering[101,102], and these functions need to be

studied further because of the high representation of this satellite on the gorilla sex chromosomes. The enrichment of non-B DNA motifs within several satellites is also suggestive of functionality, since such structures may serve as binding sites for protein regulators[48] and may be involved in defining centromeres[55].

Our assemblies have also provided novel insights into alpha satellite organization in the apes. The patterns of HOR haplotypes in active arrays of closely related species were found to be species-specific and follow a layered expansion model where semi-symmetrical layers of older arrays surround the youngest and most homogeneous active core[59]. This signature pattern of an 'expanding centromere' is consistent with the recent 'kinetochore selection' hypothesis[58,59]. We have also clarified the previously unknown organization of alpha satellites in the lesser apes and shown that gibbon subtelomeric arrays are similar across chromosomes, while their centromeric arrays are chromosome-specific (Note S6).

Future work is needed to clarify the potential role of satellites in recombination. In some of the species studied here, subtelomeric satellites were found to be shared between X and Y and distal to the PAR. If recombination occurs within these satellites, our current annotation for the PARs will need to be expanded to include them. Additionally, the putative PAR2 sequence discovered in bonobo is flanked by an Ariel repeat that may serve as a cis-acting factor for increased double-strand break formation, as was found for a mo-2 minisatellite in mouse[103]. However, the bonobo PAR2 sequence was also found at the ends of several autosomes (Note S3), suggesting that it might act as a facilitator of recombination or represent a subtelomeric duplication[104] rather than a functional PAR. The presence of active rDNA arrays on the Y chromosomes of some primate species also hints at ectopic recombination between the Y chromosome and the short arms of the rDNA-bearing acrocentric chromosomes[7,105]. Copy number and transcriptional activity of rDNA units is known to be highly variable on human autosomes[106]; the analysis of additional primate individuals is needed to understand the prevalence and stability of these Y chromosome rDNAs.

Mapping intraspecific short-read data from multiple gorilla and chimpanzee individuals revealed intriguing patterns of diversity and illustrated the critical need for additional male great ape samples across all subspecies. Such studies will be useful for studying great apes' male- and female-specific dispersal and will greatly inform conservation efforts in these non-human ape species, all of which are endangered. As one example, we expected to observe signatures of selection on the Y chromosome in chimpanzees, as they experience high levels of sperm competition due to polyandrous mating[91], but our analyses supported selection not only in chimpanzees, but also in gorillas, which usually do not have polyandrous mating[91].

Finally, we expect these T2T assemblies to be pivotal for understanding disease-causing mutations and human-specific traits. The human X chromosome contains many genes important for cognition and reproduction[95], abnormal X chromosome gene dosage underlies female bias in autoimmune disorders[107], and X-linked mutations are responsible for 10% of Mendelian disorders[108], even though the X constitutes only ~5% of the human genome[6]. On the Y chromosome, deletions within the ampliconic regions have been previously linked to infertility[109,110]. Additional intraspecific studies, comparing the complete sex chromosomes of multiple individuals within each species (as was recently done for humans[111]), are now needed to reveal the full landscape of ape sex chromosome evolution and function.

Methods

Sequencing and assemblies

Sequencing. Whole-genome DNA sequencing (WGS) was performed using three different sequencing technologies. To obtain long and accurate reads, Pacific Biosciences (PacBio) HiFi sequencing was performed on a Sequel II with a depth of >60×. To obtain ultralong (>100-kb) reads, ONT sequencing was performed on a PromethION to achieve ≥100 Gb (≥29× depth). To assist with assemblies, paired-end short-read sequencing was performed on Hi-C (Dovetail Omni-C from Cantata Bio) libraries sequenced on Illumina NovaSeq 6000, targeting 400 M pairs of 150-bp reads (≥30× depth) per sample. For bonobo and gorilla parents, we generated paired-end short reads on an Illumina NovaSeq 6000 to achieve ≥518 M pairs of 151 bp reads (≥51× depth) for each sample. Full-length transcriptome sequencing was performed on testes tissue from specimens other than the T2T genome targets (Table S40) using PacBio Iso-Seq on up to three SMRT (8M) cells using a Sequel II.

Assemblies. The complete, haplotype-resolved assemblies of chromosomes X and Y were generated using a combination of Verkko[30] and expert manual curation. Haplotype-specific nodes in the Verkko graphs were labeled using parental-specific k-mers when trios were available (bonobo and gorilla) or Hi-C binned assemblies in the absence of trios (chimpanzee, orangutans, and siamang). Haplotype-consistent contigs and scaffolds were automatically extracted from the labeled Verkko graph, with unresolved gap sizes estimated directly from the graph structure (see Rautiainen et al.[30] for more details).

During curation, the primary component(s) of chromosomes X and Y were identified based on the graph topology as visualized in Bandage[112] and using MashMap[113] alignments of the assembly to the CHM13 human reference[21]. Several X and Y chromosomes were automatically completed by Verkko and required no manual intervention; for the remainder, manual interventions were employed (Table S5). Using available information such as parent-specific *k*-mer counts, depth of coverage, and node lengths, some artifactual edges could be removed and simple non-linear structures resolved. For more complex cases, ONT reads aligned through the graph were used to generate multiple candidate resolutions, which were individually validated to select the one with the best mapping support. Disconnected nodes due to HiFi coverage gaps were joined and gap-filled using localized, ONT-based Flye[114] assemblies. The resulting gapless, telomere-to-telomere (T2T) assemblies were oriented based on MashMap alignments to the existing reference genomes of the same or related species (Table S6); all chromosomes were oriented to start with PAR1.

To validate the T2T assemblies of chromosomes X and Y, we aligned all available read data (Table S3) to the assemblies to measure agreement between the assemblies and raw sequencing data. Specific alignment methods differed for the various data types (Supplemental Methods), but the general principles from McCartney et al.[115] were followed. Validation of the assemblies was done in multiple ways to assess assembly completeness and correctness. Assembly QV was calculated using Merqury[116]. Coverage analysis, erroneous *k*-mers, and haplotype-specific *k*-mers (for the two trios) were manually inspected using IGV[117].

Alignments

Multi-species whole-chromosome alignments. To estimate the substitution rates on the X and Y chromosomes, we used CACTUS[118] to generate seven-species multiple alignments, first for the X sequences, and separately for the Y sequences. Sequences were softmasked using repeat annotations (see below). We provided CACTUS with a guide tree, (((((bonobo,chimp),human),gorilla),(sorang,borang)),gibbon), but did not provide branch lengths.

Pairwise alignments. To compute the percentage of sequences aligned and to study structural variants and

segmental duplications, the pairwise alignment of the human chromosome X and Y was performed against each of chromosome X and Y of the six ape species using minimap2.24[119]. To support other analyses, lastz[120] was used to compute pairwise alignments of X and Y chromosomes for each species.

Nucleotide substitution analysis

Nucleotide substitution frequency analysis. Substitution rates were estimated (separately for the X and the Y chromosomes) for alignment blocks containing all seven species with the REV model implemented in PHYLOFIT[121].

Nucleotide substitution spectrum analysis. Substitution spectrum analysis was conducted using 13-way CACTUS[118] alignments, which, in addition to the seven studied species, include six ancestral species sequences reconstructed by CACTUS[118]. Triple-nucleotide sequences with 5' base identical among 13 sequences and 3' base identical among 13 sequences were used for downstream substitution spectrum analysis. For each branch, 96 types of triple-nucleotide substitutions were grouped into six types based on the middle base substitutions (C>A, C>G, C>T, T>A, T>C and T>G). To compare the distribution of substitution types between chromosome X and chromosome Y and PAR1, we applied *t*-test to the proportions of each substitution type per branch, using Bonferroni correction for multiple testing.

Segmental duplications and structural variants

Segmental duplications (SDs). The SD content in humans and non-human primates was identified using SEDEF (v1.1)[122] based on the analysis of genome assemblies soft-masked with TRF v.4.0.9[123], RepeatMasker[124], and Windowmasker (v2.2.22)[125]. The SD calls were additionally filtered to keep those with sequence identity >90%, length >1 kb, and satellite content <70%. Lineage-specific SDs were defined by comparing the putative homologous SD loci, defined as containing 10 kb syntenic sequence flanking the SD. The lineage-specific SDs of each species were, thus, identified based on non-orthologous locations in the genomes.

Structural variants. Structural variants were identified against the human reference genome, CHM13v2.0, via minimap (v2.24) pairwise alignment of ape chromosomes against the human chromosome X and Y[119,126]; 50-bp-300-kb sized SVs with PAV[127]. Larger events were identified and visually inspected using the Saffire SV variant calling pipeline (https://github.com/wharvey31/saffire_sv). The human-specific structural variants were identified by intersecting the variant loci of six ape species; deletions in the six ape species relative to human reference chromosome as putative human-specific insertion, and insertions as putative human-specific deletions. The phylogenetic branch of origin of each SV was predicted using maximum parsimony. As a limitation of this analysis, the SVs for branches including ancestors of the reference species, i.e. human ancestors (i.e. human-chimpanzee-bonobo, human-chimpanzee-bonobo-gorilla, and human-chimpanzee-bonobo-gorilla-orangutan common ancestors) were not computed.

Palindromes and ampliconic regions

Palindrome detection and grouping. We developed *palindrover* in order to screen the X and Y chromosomes for palindromes with ≥98% sequence identity, length ≥8 kb, and spacer ≤500 kb, only keeping candidates with <80% of repetitive content. After aligning the arms with lastz[120] (alignments with identity <85%, gaps >5%, <500 matched bases, or covering less than 40% of either arm, were discarded), we identified orthologous palindromes and grouped paralogous palindromes on the same chromosome.

Overview of the workflow for sequence class annotations. We annotated sequence classes following[11],

with modifications. First, PARs and satellite repeat tracks were created (by aligning X and Y chromosomes for PARs, and by merging adjacent (within 1 kb) RepeatMasker[124] annotation spanning >0.25 Mb). Next, ampliconic regions were identified as a union of palindromes and regions with high intrachromosomal similarity (i.e. similar to other locations within non-PAR, here identified as consecutive 5-kb windows mapping with ≥50% identity to the RepeatMasked chromosomes using blastn from BLAST+ v.2.5.0[128,129], excluding self-alignments, and spanning >90 kb). The remaining subregions on the Y were annotated as X-degenerate or ampliconic if overlapping respective genes. Subregions nested within two matching classes were annotated as such.

Satellite and repeat analysis

Satellite and repeat annotations. To identify canonical and novel repeats on chromosomes X and Y, we utilized the previously described pipeline[41], with modifications to include both the Dfam 3.6[130] and Repbase (v20181026)[131] libraries for each species during RepeatMasker[132] annotation. A subsequent RepeatMasker run was completed to include repeat models first identified in the analysis of T2T-CHM13, and the resulting annotations were merged. To identify and curate previously undefined satellites, we utilized additional TRF[123] and ULTRA[133] screening of annotation gaps >5 kb in length. To identify potential redundancy, satellite consensus sequences generated from gaps identified in each species were used as a RepeatMasker library to search for overlap in the other five analyzed primate species. Consensus sequences were considered redundant if there was a significant annotation overlap in the RepeatMasker output. Subsequently, final repeat annotations were produced by combining newly defined satellites and 17 variants of pCht/StSat derived from Cechova et al.[134] and merging resulting annotations. Newly defined satellites that could not be searched using RepeatMasker[132] due to complex variation were annotated using TRF[123] and manually added. Tandem composite repeats were identified using self-alignment dotplots and subsequently curated using BLAT[135] to identify unit lengths and polished using a strategy defined in[136]. Composite repeats were compiled in a distinct repeat annotation track from canonical repeat annotations.

Lineage-specific insertions were characterized by identifying unaligned regions from CACTUS alignments of the seven primate X and Y chromosomes with halAlignExtract[137]. Unaligned regions were filtered by length and for tandem repeats using TRF[123] and ULTRA[133]. RepeatMasker[132] was used to identify the content of the lineage-specific insertions using the approach described above.

Non-B DNA annotations. G-quadruplex motifs were annotated with Quadron[138], and other types of non-B DNA motifs—with gfa (https://github.com/abcsFrederick/non-B_gfa). To compute non-B DNA density, we used bedtools ‘coverage' command to count the number of overlaps between each 100 kb window and non-B DNA motifs. We used glm function implemented in R to perform simple and multiple logistic regression to evaluate the relationship between non-B DNA density and sequences gained by the new assemblies. The non-B DNA enrichment analysis for satellites is described in Supplemental Methods.

Centromere analysis. To analyze centromeres, we annotated alpha-satellites (AS) and built several tracks at the UCSC Genome Browser (https://genome.ucsc.edu/s/fedorrik/primatesX and https://genome.ucsc.edu/s/fedorrik/primatesY): (1) Suprachromosomal Family (SF) tracks using human-based annotation tools[59] and utilizing score/length thresholds of 0.7, 0.3, and zero; (2) AS-strand track; (3) Higher Order Repeat (HOR) track using species-specific tools specifically designed for this project (https://github.com/fedorrik/apeXY_hmm) and methods described in [59]; (4) Structural Variation (StV, i.e. altered monomer order) tracks in HORs; (5) CENP-B sites visualized by running a short match search with the sequence YTTCGTTGGAARCGGGA. Other methods are described in Supplemental Methods and Note S6.

Gene annotations and analysis

Gene annotations at the NCBI. The *de novo* gene annotations of the six primate assemblies were performed by the NCBI Eukaryotic Genome Annotation Pipeline as previously described for other genomes[139,140], between March 20 and May 31, 2023. The annotation of protein-coding and long non-coding genes was derived from the alignments of primate transcripts and proteins queried from GenBank and RefSeq, and same-species RNA-seq reads and PacBio Iso-Seq queried from the Sequence Read Archive to the WindowMasker[125] masked genome. cDNAs were aligned to the genomes using Splign[141], and proteins were aligned using ProSplign. The RNA-seq reads (Additional Data File 4), ranging from 673 million (*Pongo pygmaeus*) to 7.3 billion (*Pan troglodytes*) were aligned to the assembly using STAR[142], while the Iso-Seq reads (ranging from none for *Symphalangus syndactylus* to 27 million for *Gorilla gorilla*) were aligned using minimap2[119]. Short non-coding RNAs, rRNAs, and tRNAs were derived from RFAM[143] models searched with Infernal cmsearch[144] and tRNAscan-SE[145], respectively.

Gene annotations at the UCSC. Genome annotation was performed using the Comparative Annotation Toolkit (CAT)[146]. First, whole-genome alignments between the primate (gorilla, chimpanzee, bonobo, Sumatran orangutan, Bornean orangutan, and siamang) and human GRCh38, and T2T-CHM13v2 genomes were generated using CACTUS[118]. CAT then used the whole-genome alignments to project the UCSC GENCODEv35 CAT/Liftoff v2 annotation set from CHM13v2 to the primates. In addition, CAT was given Iso-Seq FLNC data to provide extrinsic hints to the Augustus PB (PacBio) module of CAT, which performs *ab initio* prediction of coding isoforms. CAT was also run with the Augustus Comparative Gene Prediction (CGP) module, which leverages whole-genome alignments to predict coding loci across many genomes simultaneously (i.e. gene prediction). CAT then combined these *ab initio* prediction sets with the human gene projections to produce the final gene sets and UCSC assembly hubs used in this project.

Curation and analysis of X-degenerate genes. For the Y chromosome, we collected annotations from the NCBI Eukaryotic Genome Annotation Pipeline (RefSeq), CAT, and Liftoff. We extracted X-degenerate gene annotations from each and mapped them onto the Y chromosome sequence for each in Geneious[147]. We identified that every gene was present and manually curated an annotation set with the most complete exonic complement across annotations. We extracted all CDS regions for each gene and aligned them. For the X chromosome, we extracted X-degenerate gene copies from the RefSeq annotations using gffread[148] and aligned them. All alignments were examined and curated by eye, and missing genes and exons were confirmed using BLAST[129]. All present genes were aligned to their orthologs and their gametologs, where we identified genes with significant deviations (truncations) relative to their X chromosome counterparts as pseudogenes. These alignments were also used to identify gene conversion events using GeneConv[149] and to detect selection.

Curation of ampliconic genes. We first collected annotations from the NCBI annotation pipeline, CAT, and Liftoff. To these annotations, we added mappings from human and species-specific gene sequences onto the latest assemblies and also included the full Iso-Seq transcripts. To combine these annotations, we first performed an interval analysis to find all annotated, mapped, or predicted copies, with one or more sources of evidence and then manually curated the final set of protein-coding and pseudogene copies for each of these genes (Additional Fille 5).

ddPCR ampliconic gene copy number validations. Copy numbers were determined with ddPCR using the protocols described in [13] and [68]. The sequences of the primers for bonobo, chimpanzee, gorilla, Bornean and Sumatran orangutan were from [68]. The primers for siamang were designed using Geneious Prime software[147] and are available in Table S34. ddPCR conditions are also described in Table S34.

Estimating rDNA copy number and activity by FISH and immuno-FISH. Chromosome spreads were prepared and labeled as described previously[66]. To estimate rDNA copy number and activity from FISH and immunoFISH images, individual rDNA arrays were segmented, the background-subtracted integrated intensity was measured for every array, and the fraction of the total signal of all arrays in a chromosome spread was calculated for each array. Similarly, the fraction of the total UBF fluorescence intensity was used to estimate the transcriptional activity of the chrY rDNA arrays. The fraction of the total rDNA fluorescence intensity was used to determine the number of rDNA copies per array. The total rDNA copy number in a genome was estimated from Illumina sequencing data based on *k*-mer counts. Full details are available in Supplemental Methods.

Gene-level selection using interspecific fixed differences. To detect selection from interspecific comparison of gene sequences, we started with alignments of X-degenerate or ampliconic genes, using one consensus sequence per species for ampliconic gene families. For these alignments, we inferred ML phylogeny with raxml-ng (GTR+G+I, default settings otherwise), and looked for evidence of gene-level episodic diversifying selection (EDS) using BUSTED with site-to-site synonymous rate variation and a flexible random effects branch-site variation for d_N/d_S[150,151]. Because all alignments were relatively short, we also fitted the standard MG94 + GTR model where d_N/d_S ratios were constant across sites and were either shared by all branches (global model) or estimated separately for each branch (local model). We tested for $d_N/d_S \neq 1$ using a likelihood ratio test (global model). To investigate branch-level variability in d_N/d_S, we used a version of the local model where all branches except one shared the same d_N/d_S ratio and the focal branch had its own d_N/d_S ratio; *p*-values from branch-level d_N/d_S tests were corrected using the Holm-Bonferroni procedure. Finally, to compare mean in global dN/dS between ampliconic and X-degenerate genes, we performed a joint MG94 + GTR model fit to all genes, with the null model that d_N/d_S is the same for all genes, and the alternative model that d_N/d_S are the same within group (ampliconic or X-degenerate), but different between groups. All analyses were run using[152].

Methylation analysis

CpG methylation calling. To generate CpG methylation calls, Meryl[116] was used to count *k*-mers and compute the 0.02% most frequent 15-mers in each ape draft diploid assembly. ONT and PacBio reads were mapped to the corresponding draft diploid assemblies with Winnowmap2[153] and filtered to remove secondary and unmapped reads. Modbam2bed (https://github.com/epi2me-labs/modbam2bed) was used to summarize modified base calls and generate a CpG methylation track viewable in IGV[154].

Methylation analysis. Using the processed long-read DNA methylation data to analyze large sequence classes (PAR1, Ampliconic regions, X-degenerate regions, X-ancestral regions), we split these regions into 100-kb bins and calculated mean methylation levels of all CpGs within each bin. For smaller sequence classes, such as specific repetitive elements, we generated mean methylation levels from individual elements themselves. For human data, we added another filtering step to remove regions where two long-read sequencing platforms yielded highly divergent results (mostly Yq12 region); non-human methylation data were concordant between the two sequencing platforms (Fig. S17) and thus were used in their entirety. Promoters were defined as regions 1 kb upstream of the transcription start site.

Diversity analysis

We collected short-read sequencing data from 129 individuals across 11 distinct great ape subspecies and aligned the reads to previous (using the previous reference of S. orangutan reference for B. orangutan data) and T2T sex chromosome assemblies. We next performed variant calling with GATK Haplotype Caller[155], conducted joint genotyping with GenotypeGVCFs[155], and removed low-confident variants. To further enhance the accuracy and completeness of variant detection, we adopted the masking strategy proposed by the

T2T-CHM13v2.0 human chrY study[7], in which PARs and/or Y chromosome were masked in a sex-specific manner. After generating karyotype-specific references for XX and XY samples, we realigned the reads of each sample to the updated references and called variants. The new variant set was validated reconstructing the Y chromosome phylogeny and estimating the time-to-most-recent common ancestor on it (Note S13). Using the complete variant call sets, we quantified the nucleotide diversity of each subspecies with VCFtools. For chromosome X, we assessed the diversity in PAR and ancestral regions. For chromosome Y, we computed the nucleotide diversity in X-degenerate regions.

Data Availability

The raw sequencing data generated in this study have been deposited in the Sequence Read Archive under BioProjects PRJNA602326, PRJNA902025, PRJNA976699-PRJNA976702, and PRJNA986878-PRJNA986879. The genome assemblies and NCBI annotations are available from GenBank/RefSeq (see Table S41 for accession numbers). The CAT/Liftoff annotations are available in a UCSC Genome Browser Hub: https://cgl.gi.ucsc.edu/data/T2T-primates-chrXY/. The reference genomes, alignments and variant calls are also available within the NHGRI AnVIL: https://anvil.terra.bio/#workspaces/anvil-dash-research/AnVIL_Ape_T2T_chrXY. The alignments generated for this project are available at: https://www.bx.psu.edu/makova_lab/data/APE_XY_T2T/ and https://public.gi.ucsc.edu/~hickey/hubs/hub-8-t2t-apes-2023v1/8-t2t-apes-2023v1.hal (with the following additional information: https://public.gi.ucsc.edu/~hickey/hubs/hub-8-t2t-apes-2023v1/8-t2t-apes-2023v1.README.md) Additional Data Files include human-specific structural variant coordinates (File 1), sequence class coordinates (File 2), palindrome coordinates (File 3), RNA-Seq and IsoSeq datasets used for gene annotations (File 4), and manual annotations of Y ampliconic genes (File 5).

Code Availability

The source code created to generate the results presented in this paper is publicly available on GitHub (https://github.com/makovalab-psu/T2T_primate_XY). All external scripts and programs are also linked through this GitHub repository.

Acknowledgements

From Kateryna Makova: This work was supported, in part, by the NIH awards R01GM130691 and R01GM136684 (both to KDM). We are grateful to Bernadette Weissensteiner and Kate Anthony who assisted with primate cell culture, to PSU Genomics Core Facility, PSU Sartorius Cell Culture Facility, and PSU College of Medicine Genome Sciences Core Facility for their technical assistance, and to San Diego Zoological Society, Coriell Institute, Smithsonian Institute, and University of Texas MD Anderson Cancer Center for providing samples and/or cell lines used in this study.

From Adam Phillippy: This work was supported, in part, by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health. This work utilized the computational resources of the NIH HPC Biowulf cluster (https://hpc.nih.gov).

From Evan Eichler: This work was supported, in part, by National Institutes of Health (NIH) grants HG002385 and HG010169 (to E.E.E.). E.E.E. is an investigator of the Howard Hughes Medical Institute.

From Françoise Thibaud-Nissen: The work of FT-N, DH and PM was supported by the National Center for Biotechnology Information of the National Library of Medicine (NLM), National Institutes of Health.

From Paul Medvedev: This material is based upon work supported by the National Science Foundation under grant nos. 2138585 and 1931531 (to P.M.). Research reported in this publication was also supported by the National Institutes of Health under Grant NIH R01GM146462 (to P.M.).

From Jennifer Gerton: This material was supported in part by NIH Grant R01CA266339 to JLG and by the Stowers Institute for Medical Research. Primary data related to the cytogenetic evaluation of the rDNA is deposited in ODR-XXXX (accession number in progress).

From Rachel O'Neill: The work of RO, GH, JS, PG, and SH was supported by the National Institutes of Health under Grant NIH R01GM123312 (to RO). The work utilized the computational resources of the Computational Biology Core and sequencing at the Center for Genome Innovation, both in the Institute for Systems Genomics.

From Zachary A. Szpiech: The work of ZAS and Sweetalana was supported, in part, by NIH Grant R35GM146926 (to ZAS).

From Christian D. Huber: The work of CDH was supported by NIH Grant R35GM146886 (to CDH). TML was supported by the NIH T32 GM102057 Computation, Bioinformatics, and Statistics (CBIOS) Training Program Grant.

From Soojin Yi: The work of SVY, DAH, and YEL was supported, in part, by the National Institutes of Health grant R01HG011641 (to SVY) and the National Science Foundation grant EF-2204761 (to SVY).

From Michael Schatz. This work was supported, in part, by NIH awards U01CA253481 and U24HG010263 to MCS.

From Ivan A. Alexandrov: This work was supported in part by the Center for Integration in Science of the Ministry of Aliyah, Israel (IAA).

From Melissa Wilson: The work of BJP and MAW was supported by the NIH award R35GM124827 to MAW.

From Charles Lee: This work was supported in part by National Institutes of Health (NIH) grants HG007497

(to CL and EEE).

We would like to acknowledge Francesca Chiaromonte, Tamara Goldfarb, Bernard de Massy, Terence D. Murphy, Morgan Park, Dylan J Taylor, Marta Tomaszkiewicz, and Allison Watwood for their assistance and/or advice.

Author Contributions

BDP performed computational validations, NCBI submissions, chimpanzee subspecies identification, biosample registration, figure generation, and overall project and consortium coordination. RSH generated alignments, identified pseudoautosomal boundaries and palindromes, and performed substitution analysis. MC classified assemblies into region classes, identified ampliconic regions, performed palindrome analysis, and contributed to palindrome analysis. GAH, PGSG, JMS, and SJH performed repetitive element annotation, manual curation, analyses and dfam submissions, JMS performed lineage-specific repeat analyses, and GAH generated tracks for figures, led by RJO. SN and SK performed sequence assemblies. GH and BP generated multi-species alignments. AS generated dotplots. SJS performed rDNA array copy number estimation, base calling, and alignment, and generated methylation tracks. DAY, WTH, and HJ performed segmental duplication and structural variation analyses. QL, AB, MCS, RCMC, MGT, CDH, TMLP, S, ZAS, PHa, CL, and SLKP performed diversity and selection analyses. KP, PHe, FTN, DH, PM, MAW, BJP, and MD performed gene annotations and analyses. KK performed non-B DNA analysis. XZ performed substitution spectrum analysis and assisted in figure preparation. DEC, KS, PCC, and AC performed DeepConsensus calling. MA and EB performed de novo gene analysis. CSC analyzed palindrome structure in orangutans. PHD and JLR provided HiFi data for bonobo. IAA, FR, VAS, VS, and KHM performed centromere analysis. SVY, DAH, and YHEL performed methylation analysis. TP, MB, and JLG performed rDNA analysis. AD and EP generated karyotypes. GAH, LuC and RJO confirmed the siamang karyotype. LDG and MV performed karyotype confirmation and FISH analysis on rDNA. HZ performed ddPCR and maintained cell culture. ACY, SYB, and GGB generated UL-ONT and Illumina sequences. SS and REG generated HiC libraries. KMM, APL, and GHG generated HiFi and IsoSeq PacBio sequences. AR, PM, and SJCC participated in project discussions, SJCC also collected gene ontology and mating system information, and AR performed methylation comparison between two sequencing platforms. LaC, LuC, and OAR provided samples. LuC also provided karyotype confirmation. BCMG coordinated project resources, maintained cell culture, performed ddPCR and RNA extractions. KDM, EEE, and AMP provided project leadership and coordination, and are co-leading the Primate T2T consortium. KDM wrote the manuscript with contributions from the other authors.

Competing Interests

EEE is a scientific advisory board (SAB) member of Variant Bio, Inc. RJO is a scientific advisory board (SAB) member of Colossal Biosciences, Inc. CL is a scientific advisory board (SAB) member of Nabsys, Inc. and Genome Insight, Inc.

Additional Information

Supplementary Information is available for this paper: Supplementary Figures, Supplementary Tables, Supplementary Methods, Supplementary Notes, and Additional Data Files. Correspondence and requests for materials should be addressed to Kateryna D. Makova (kdm16@psu.edu), Evan Eichler (eee@gs.washington.edu), and Adam Phillippy (adam.phillippy@nih.gov).

Table 1. Percentage of repetitive DNA across ape X and Y chromosomes

The percentages of bases annotated as canonical transposable elements, satellites, simple/low-complexity repeats, and other unknown, or CHM13-derived, repeats are listed for each X and Y chromosome.

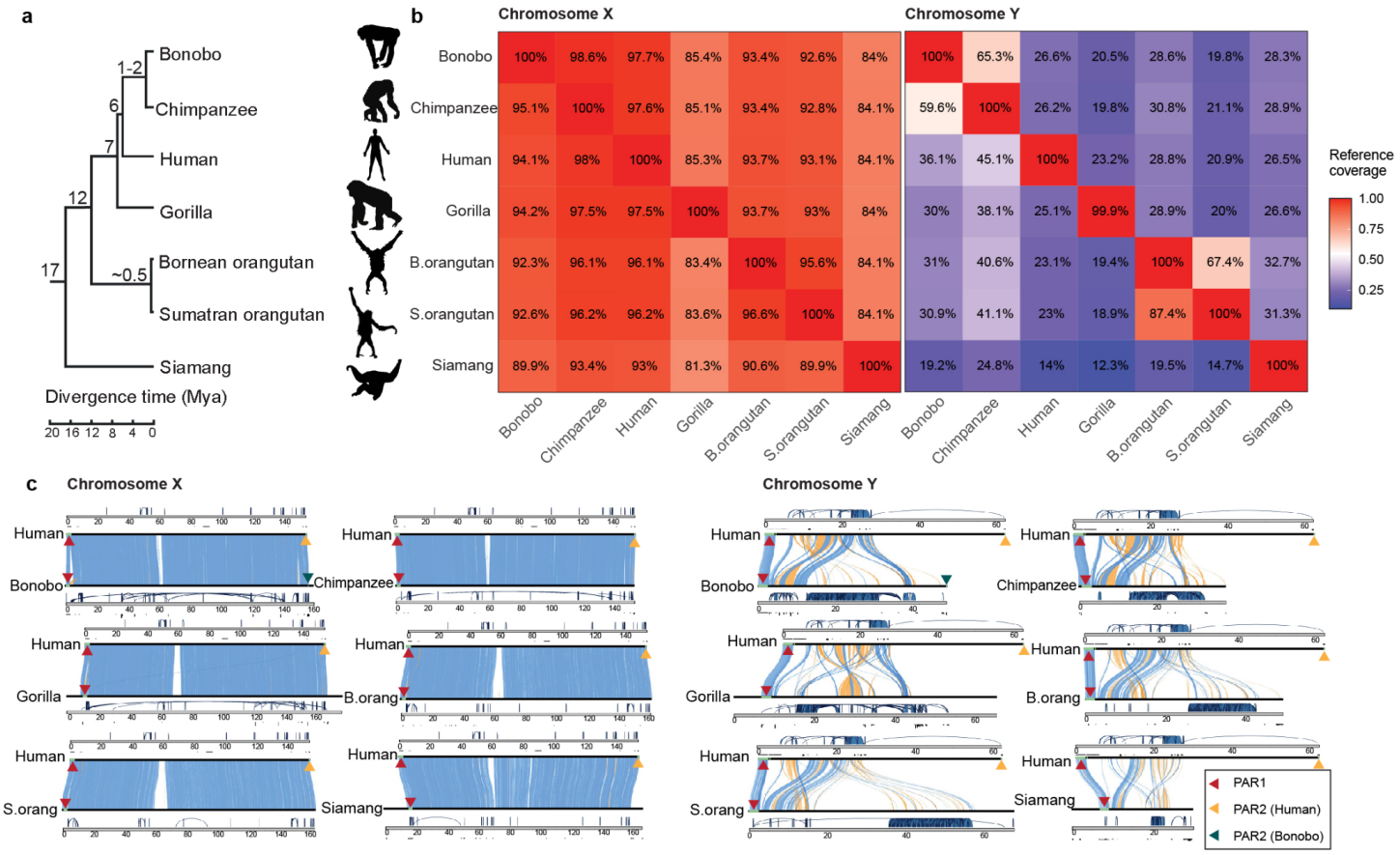
Repeat Class		Percentage of Chr X Assembly							
		CHM13 X	HG002 X	Bonobo X	Chimpanzee X	Gorilla X	Bornean orangutan X	Sumatran orangutan X	Siamang X
Canonical Transposable Elements	SINE	10.31%	10.30%	10.18%	10.30%	9.19%	9.88%	9.75%	9.97%
	Retroposon	0.15%	0.15%	0.17%	0.19%	0.15%	0.12%	0.11%	0.17%
	LINE	32.99%	32.99%	32.07%	32.94%	29.27%	33.15%	32.86%	28.69%
	LTR	11.07%	11.06%	11.08%	11.16%	9.88%	10.63%	10.53%	9.48%
	DNA	3.43%	3.42%	3.31%	3.41%	3.02%	3.28%	3.24%	3.11%
	Total TEs	**57.93%**	**57.93%**	**56.82%**	**57.99%**	**51.51%**	**57.07%**	**56.49%**	**51.41%**
Satellites and simple/low complexity repeats	Alpha Satellites	2.35%	2.35%	1.61%	1.39%	3.36%	4.35%	5.33%	11.01%
	StSat	0.00%	0.00%	2.08%	0.36%	9.62%	0.00%	0.00%	0.00%
	Newly Defined Satellites	0.21%	0.21%	0.30%	0.25%	0.17%	0.17%	0.16%	0.24%
	Other Satellites	0.10%	0.10%	0.09%	0.11%	0.09%	0.14%	0.13%	0.02%
	Simple/Low complexity repeats	1.53%	1.52%	1.51%	1.55%	1.38%	1.49%	1.47%	1.31%
Other	Other/ Unknown Repeats	0.07%	0.07%	0.07%	0.07%	0.06%	0.07%	0.07%	0.06%
	CHM13-Derived Repeats	0.01%	0.01%	0.04%	0.04%	0.03%	0.04%	0.05%	0.04%
	TOTAL MASKED	**62.20%**	**62.20%**	**62.52%**	**61.74%**	**66.22%**	**63.33%**	**63.70%**	**64.09%**

Repeat Class		Percentage of Chr Y Assembly							
		CHM13 Y	HG002 Y	Bonobo Y	Chimpanzee Y	Gorilla Y	Bornean orangutan Y	Sumatran orangutan Y	Siamang Y
Canonical Transposable Elements	SINE		7.02%	6.99%	7.84%	4.24%	7.34%	6.54%	7.04%
	Retroposon		0.03%	0.07%	0.11%	0.03%	0.07%	0.06%	0.17%
	LINE		10.34%	14.58%	17.01%	8.37%	19.85%	16.60%	12.87%
	LTR		7.39%	9.52%	11.68%	6.99%	12.64%	4.51%	8.73%
	DNA		0.74%	1.02%	1.19%	0.68%	0.95%	0.71%	1.05%
	Total TEs		**25.52%**	**32.19%**	**37.82%**	**20.31%**	**40.85%**	**28.42%**	**29.86%**
Satellites and simple/low complexity repeats	Alpha Satellites		0.71%	9.74%	5.08%	7.38%	5.70%	9.13%	30.67%
	StSat		0.00%	0.18%	0.98%	9.90%	0.00%	0.00%	0.00%
	Newly Defined Satellites		0.13%	0.24%	0.19%	0.05%	0.13%	0.11%	0.28%
	Other Satellites		19.62%	6.28%	7.15%	22.28%	3.98%	3.04%	2.93%
	Simple/Low complexity repeats		37.38%	23.90%	17.20%	24.33%	20.54%	30.16%	11.22%
Other	Other/ Unknown Repeats		0.03%	0.18%	0.13%	0.04%	0.10%	0.10%	0.40%
	CHM13-Derived Repeats		1.51%	3.26%	2.51%	1.04%	2.56%	1.69%	1.74%
	TOTAL MASKED		**84.89%**	**75.97%**	**71.06%**	**85.33%**	**73.86%**	**72.65%**	**77.10%**

Figures

Figure 1. Chromosome alignability and divergence

(**A**) The phylogenetic tree of the studied species (see text for references of divergence times). (**B**) Pairwise alignments of chromosomes X and Y (% of reference, as shown on the x-axis, covered by the query, as shown on the y-axis). (**C**) Alignment of the primate sex chromosome against the human T2T assembly[7,21]. Blue and yellow blocks indicate the direct or inverted alignments, respectively, between the chromosomes. Pseudoautosomal regions (PARs) are indicated by triangles (not to scale). (**D**) Phylogenetic trees of nucleotide sequences on the X and Y chromosomes using Progressive Cactus[118]. Branch lengths (substitutions per 100 sites) were estimated from multi-species alignment blocks including all seven species. (**E**) Substitution spectrum differences between chromosomes X and Y. Comparing the proportions of six single-base nucleotide substitution types among total nucleotide substitutions per branch between the two sex chromosomes (excluding PARs). The distribution of the proportion of each substitution type across phylogenetic branches is shown. The significance of differences was evaluated with a *t*-test and marked with * for $p<0.05$ and *** for $p<0.0005$. ‘B.orang’ and ‘B.orangutan’ stand for Bornean orangutan, and ‘S. orang’ and ‘S. orangutan’ stand for Sumatran orangutan.


a
Bonobo
Chimpanzee
Human
Gorilla
Bornean orangutan
Sumatran orangutan
Siamang
1-2
6
7
12
17
~0.5
Divergence time (Mya)
20 16 12 8 4 0
b
Chromosome X
Chromosome Y
Reference coverage
1.00
0.75
0.50
0.25
c
Chromosome X
Chromosome Y
PAR1
PAR2 (Human)
PAR2 (Bonobo)

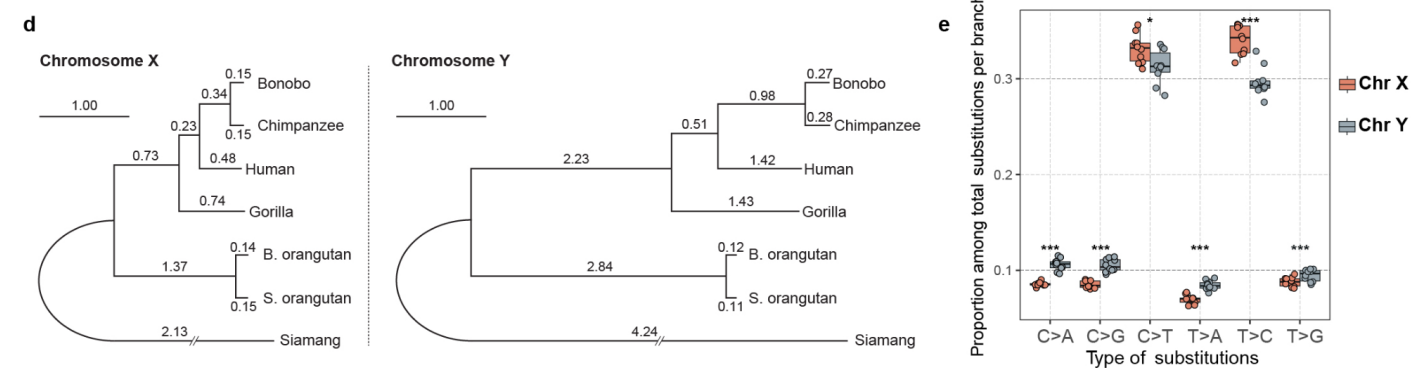

d
Chromosome X
Chromosome Y
e
Chr X
Chr Y
Proportion among total substitutions per branch
C>A C>G C>T T>A T>C T>G
Type of substitutions

Figure 2. Sequences gained, palindromes, sequence classes, and intrachromosomal similarity in the assemblies

Tracks for novel sequence relative to existing references (new in black), non-B DNA density (darker is more dense), palindromes (in black), and sequence classes (see color legend) are shown. The X and Y chromosomes are portrayed on different scales. No previous references existed for the Bornean orangutan or siamang, hence the solid black bars for the novel sequence tracks. No new sequence was added to the existing T2T human reference in this study and thus the human novel sequence tracks are empty (white). Self-similarity dot plots are also shown for the Y chromosomes (see percent identity legend). While these dot plots show the intrachromosomal similarity, the divergence between the Y chromosomes is also evident from the variable dot plot patterns. ‘B. orangutan’ and ‘S. orangutan’ stand for Bornean and Sumatran orangutan, respectively.

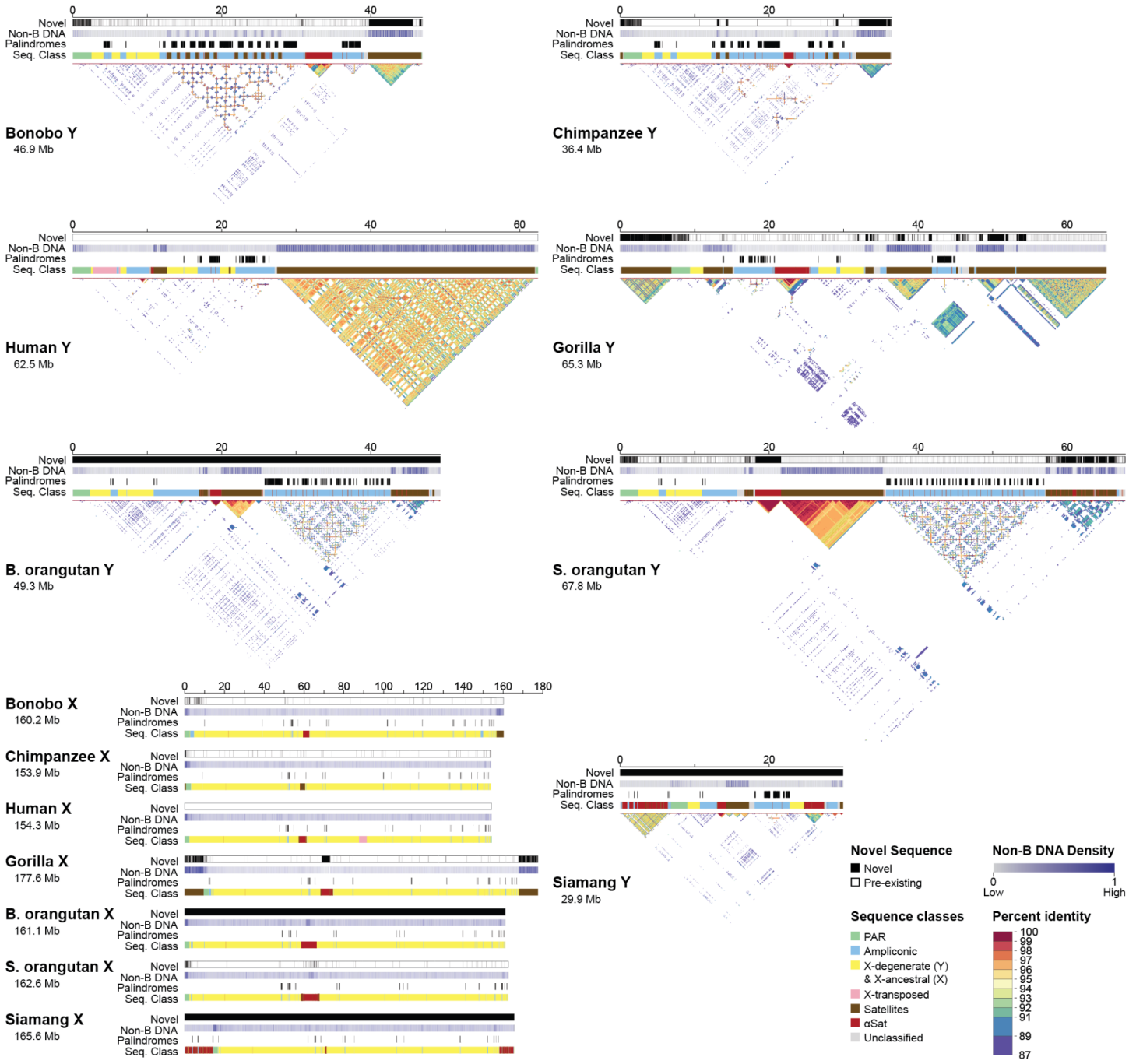

Bonobo Y
46.9 Mb
Chimpanzee Y
36.4 Mb
Human Y
62.5 Mb
Gorilla Y
65.3 Mb
B. orangutan Y
49.3 Mb
S. orangutan Y
67.8 Mb
Bonobo X
160.2 Mb
Chimpanzee X
153.9 Mb
Human X
154.3 Mb
Gorilla X
177.6 Mb
B. orangutan X
161.1 Mb
S. orangutan X
162.6 Mb
Siamang X
165.6 Mb
Siamang Y
29.9 Mb
Novel
Non-B DNA
Palindromes
Seq. Class
Novel Sequence
Novel
Pre-existing
Non-B DNA Density
Low
High
Sequence classes
PAR
Ampliconic
X-degenerate (Y) & X-ancestral (X)
X-transposed
Satellites
αSat
Unclassified
Percent identity

Figure 3. Conservation of ampliconic regions and palindromes across species

(**A**) A comparison of ampliconic regions on the X chromosomes and (**B)** Y chromosomes between species with similarities highlighted using a dot plot analysis. Ampliconic regions were extracted and concatenated independently for each species, and visualized with gepard[156] using a window size of 100. (**C**) Palindromes on the X chromosome: all palindromes are shown, but shared palindromes are connected by edges. An edge in one color from one species always connects to only one other species. Circular genomic coordinates are accompanied by color maps of sequence classes. Genes located in palindromes shared across species are shown. The number of genes following each other in a sequence without being interrupted by other genes is shown in parentheses. (**D**) The same for the Y chromosomes, with the only difference being that we limited the plot to ampliconic gene families (*BPY*2, *CDY*, *DAZ*, *HSFY*, *PRY*, *RBMY*, *TSPY*, *VCY*) located in palindromes.

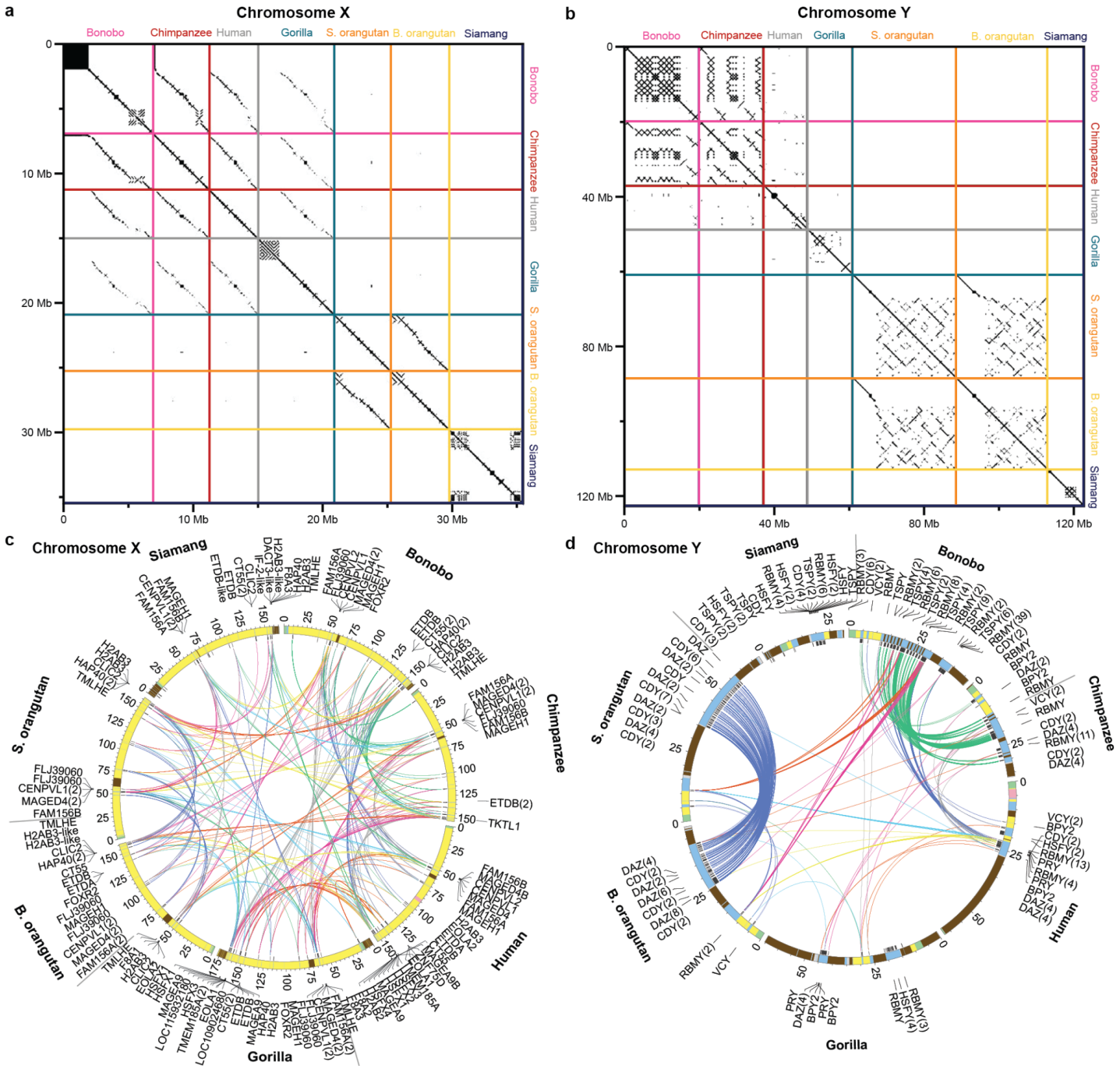

a
Chromosome X
Bonobo
Chimpanzee
Human
Gorilla
S. orangutan
B. orangutan
Siamang
0
10 Mb
20 Mb
30 Mb
b
Chromosome Y
40 Mb
80 Mb
120 Mb
c
Chromosome X
Siamang
Bonobo
Chimpanzee
Human
Gorilla
B. orangutan
S. orangutan
ETDB(2)
TKTL1
d
Chromosome Y
Siamang
Bonobo
Chimpanzee
Human
Gorilla
B. orangutan
S. orangutan

Figure 4. Repeats and centromeres on ape sex chromosomes

(**A**) Overall repeat annotations (left), lineage-specific repeat expansions (center), and major satellites (right) across each of the ape sex chromosomes. Overall repeat annotations (left) are depicted as a percentage of total nucleotides. Each repeat class is defined by color, with gray representing non-repetitive DNA. Previously uncharacterized human repeats derived from the CHM13 genome analyses are demarcated in teal, adding 0.02% to 2.84% of annotations in each of the non-human apes. Newly defined satellites (Methods), depicted in light orange, account for an average of 344 kb and 91 kb on each ape X and Y chromosome, respectively. The number of bases comprising lineage-specific repeat expansions (middle) are shown in the same colors as the overall repeat annotations, except that non-repetitive DNA (gray) is omitted. The number of bases on each X and Y chromosome comprising canonical satellites are shown, with each satellite represented by a different color according to the included key. Of note, StSat/pCht and SAR/HSat1A satellites have undergone expansion on gorilla, bonobo, and chimpanzee X and Y chromosomes. Alpha satellites, present in all species, form large subterminal expansions in siamang gibbon. (**B**)The left panel shows the primate phylogenetic tree with active alpha satellite (AS) suprafamilies (SFs) specified. Chromosome-specific organization indicates that the active centromere in each chromosome has a different higher order repeat (HOR). In pan-chromosomal organization, all centromeres have similar repeats. The right panel shows the generalized centromeres for each branch (not to scale) with SF composition of the active core indicated in the middle and of the dead flanking layers on the sides. Each branch has one or few SFs fewer than in African apes, but may have a number of branch-specific layers not shared with the human lineage (shown by hues of the same color). The African ape centromere cores are shown as horizontal bars of SF1-SF3 to represent that each chromosome has only one SF, and the SF differs with each chromosome. (**C**) The UCSC Genome Browser tracks showing the SF composition of centromere cores (not to scale) and the flanks for cenYs and cenXs. CenX is always surrounded by stable vestigial layers, which represent the remnants of dead ancestral centromeres, while cenY has a ‘naked’ centromere devoid of standard monomeric layers. Thin gray lines under the tracks show overlaps with the segmental duplications tracks, which are abundant among cenY flanks. In gorilla cenX, SF3 (cyan) was replaced by SF2 (purple) and then by SF1 (pink). The colors for all AS SFs, applicable to panels B and C, are listed in the included key. See details in Note S6.

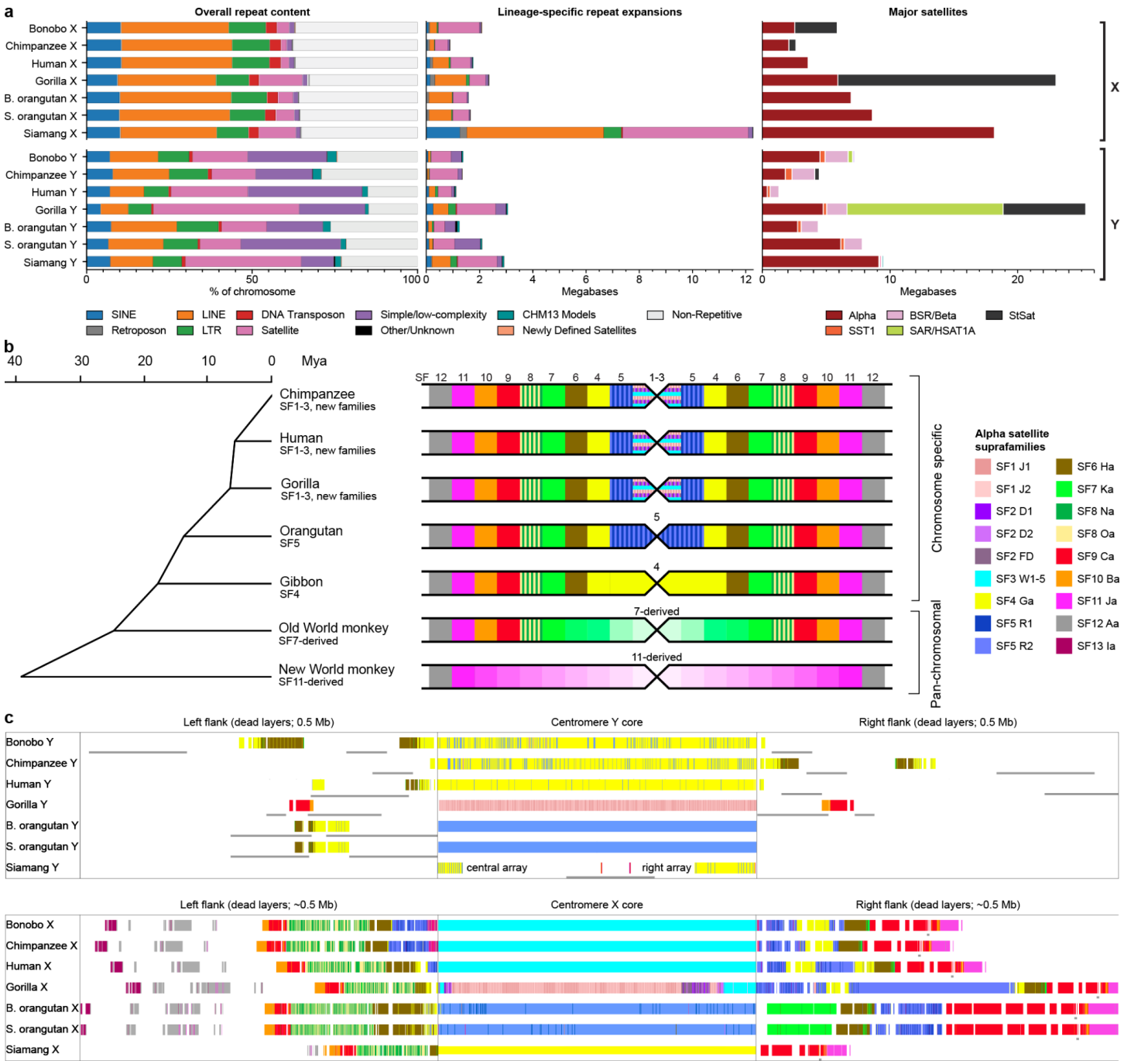

a
Overall repeat content
Lineage-specific repeat expansions
Major satellites
Bonobo X
Chimpanzee X
Human X
Gorilla X
B. orangutan X
S. orangutan X
Siamang X
Bonobo Y
Chimpanzee Y
Human Y
Gorilla Y
B. orangutan Y
S. orangutan Y
Siamang Y
X
Y
% of chromosome
Megabases
Megabases
SINE
LINE
DNA Transposon
Simple/low-complexity
CHM13 Models
Non-Repetitive
Retroposon
LTR
Satellite
Other/Unknown
Newly Defined Satellites
Alpha
BSR/Beta
StSat
SST1
SAR/HSAT1A
b
Mya
Chimpanzee
SF1-3, new families
Human
SF1-3, new families
Gorilla
SF1-3, new families
Orangutan
SF5
Gibbon
SF4
Old World monkey
SF7-derived
New World monkey
SF11-derived
7-derived
11-derived
Chromosome specific
Pan-chromosomal
Alpha satellite suprafamilies
SF1 J1
SF1 J2
SF2 D1
SF2 D2
SF2 FD
SF3 W1-5
SF4 Ga
SF5 R1
SF5 R2
SF6 Ha
SF7 Ka
SF8 Na
SF8 Oa
SF9 Ca
SF10 Ba
SF11 Ja
SF12 Aa
SF13 Ia
c
Left flank (dead layers; 0.5 Mb)
Centromere Y core
Right flank (dead layers; 0.5 Mb)
central array
right array
Left flank (dead layers; ~0.5 Mb)
Centromere X core
Right flank (dead layers; ~0.5 Mb)

Figure 5. Estimation of rDNA copy number and activity on chromosome Y arrays

(**A**) Gallery view of Y chromosomes from species used in this study. Chromosomes were labeled by FISH with BAC probes containing rDNA (BAC RP11-450E20, green) and *SRY* (BAC RP11-400O10, red). DNA was counter-stained with DAPI. rDNA signal is present on the distal ends of the q-arms of Y chromosomes in Sumatran orangutan and siamang. (**B**) Quantification of rDNA copy number on chrY in siamang and Sumatran orangutan. Chromosome spreads were labeled by FISH with probes for rDNA and *SRY* as in panel A. The rDNA copy number on chrY was calculated from its fraction of the total fluorescent intensity of the rDNA signals on all chromosomes and the Illumina sequencing estimate of the total copy number of rDNA repeats in the genome (339 copies in siamang and 814 copies in Sumatran orangutan). The box plot shows mean values with standard deviations of chrY rDNA from 20 chromosome spreads. The rounded average rDNA arrays on chrY were 16 copies for siamang and 3 copies for Sumatran orangutan. (**C**) A representative image of siamang chrY labeled by immuno-FISH with rDNA probe (green) and the antibody against rDNA transcription factor UBF (magenta). The chrY rDNA array is positive for the UBF signal. (**D**) Quantification of siamang chrY rDNA and UBF expressed as the fraction of the total fluorescent intensity of all rDNA-containing chromosomes in a chromosome spread. The box plot shows means with standard deviations from 20 spreads. ChrY rDNA arrays contain on average ~10% of the total chromosomal UBF signal. Siamang (**E**) and Sumatran orangutan (**F**) read-level plots showing ONT methylation patterns at the chrY rDNA locus and surrounding regions. The coverage track shows the depth of sequencing coverage across the rDNA array, and the methylation track displays the methylation status of individual cytosines. Only reads >100 kb that are anchored in unique sequence outside the rDNA array and span at least two 45S units are shown. Unmethylated and methylated cytosines are shown in blue and red, respectively.

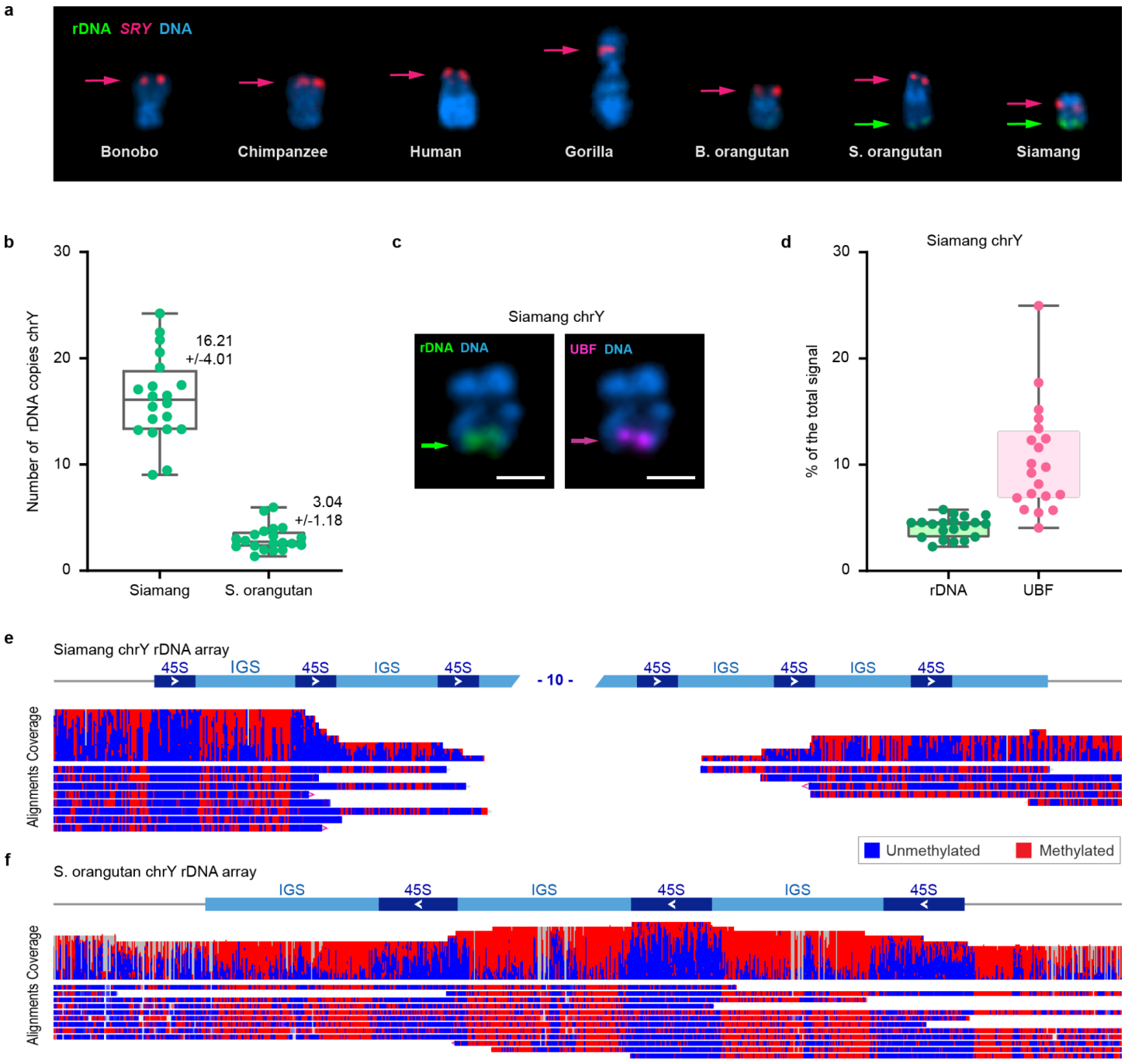
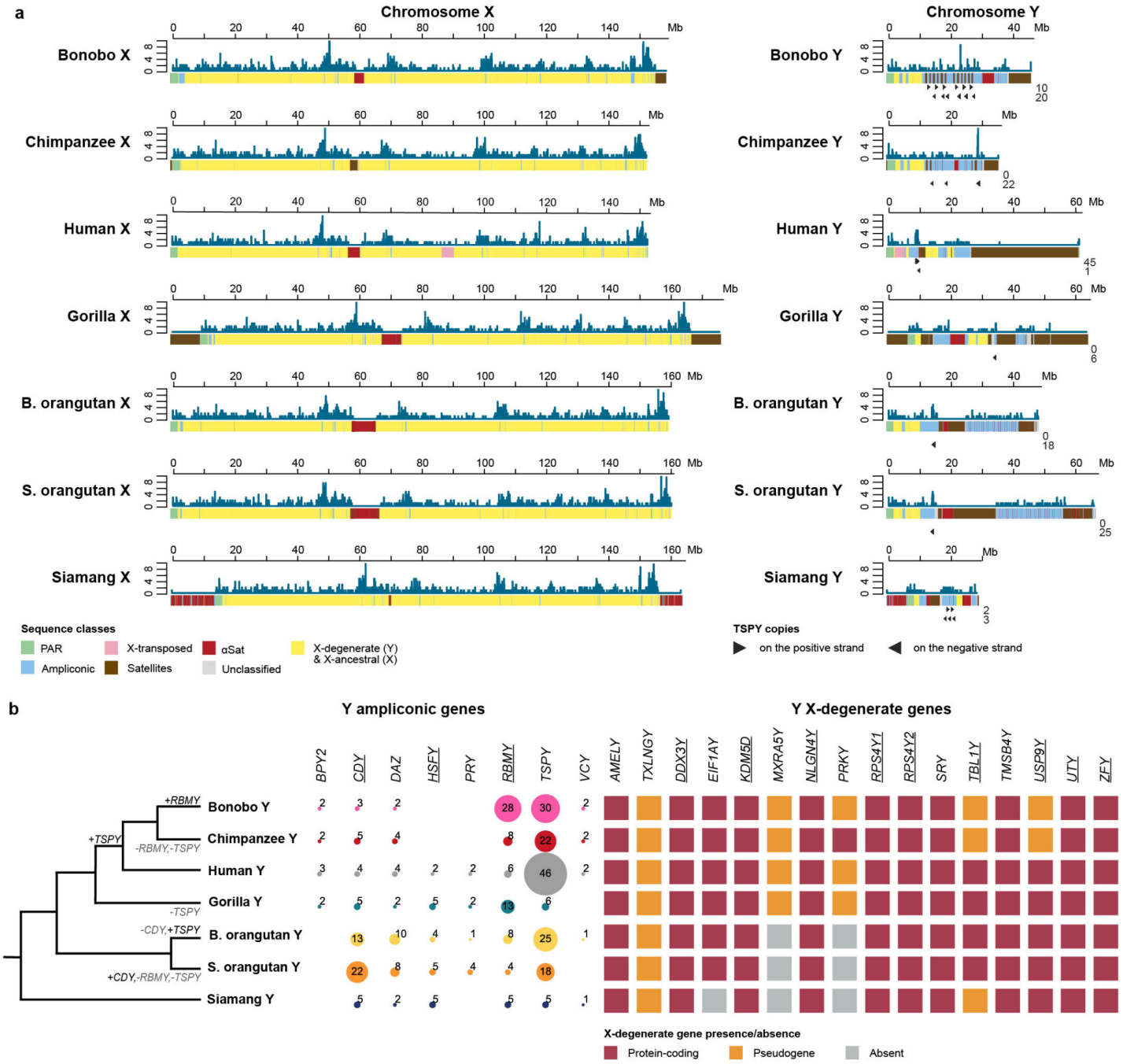

a
rDNA SRY DNA
Bonobo
Chimpanzee
Human
Gorilla
B. orangutan
S. orangutan
Siamang
b
Number of rDNA copies chrY
30
20
10
0
16.21
+/-4.01
3.04
+/-1.18
Siamang
S. orangutan
c
Siamang chrY
rDNA DNA
UBF DNA
d
Siamang chrY
% of the total signal
30
20
10
0
rDNA
UBF
e
Siamang chrY rDNA array
45S
IGS
45S
IGS
45S
- 10 -
45S
IGS
45S
IGS
45S
Alignments Coverage
Unmethylated
Methylated
f
S. orangutan chrY rDNA array
IGS
45S
IGS
45S
IGS
45S
Alignments Coverage

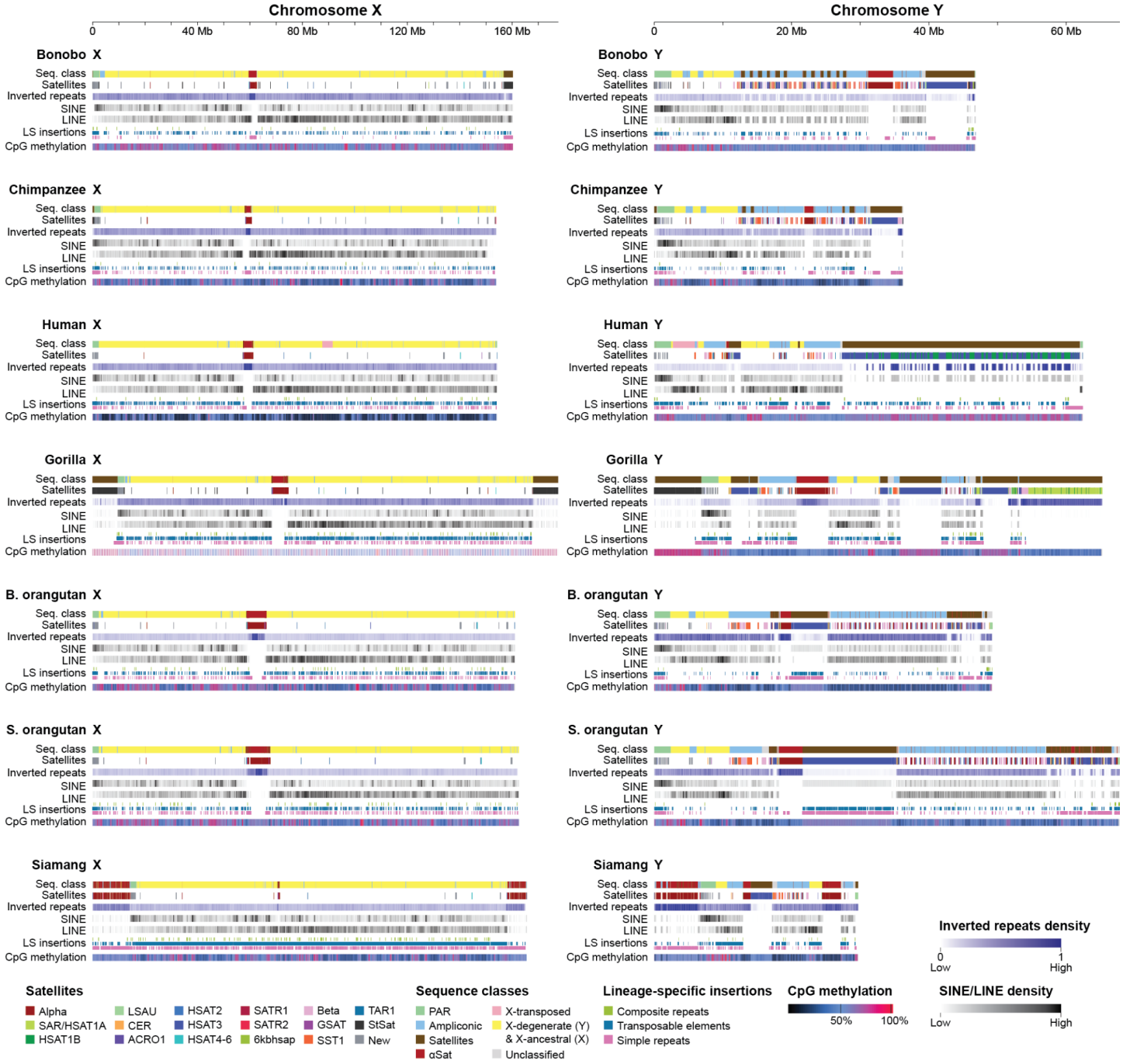
Figure 6. Gene evolution

(A) Density (number of genes per 100 kb, shown on the y-axis) of protein-coding genes along the X and Y chromosomes with respect to sequence classes visualized on the x-axis for each chromosome. X and Y chromosomes are drawn to the same scale. The *TSPY* copies are shown below the Y chromosomes as black arrows pointing in the direction of DNA strands carrying gene copies, with the total number of copies per strand indicated. **(B)** Copy number (noted by number and shown by circle size) or absence of ampliconic genes, and presence, pseudogenization, and absence (i.e., deletion) of X-degenerate genes, on the Y chromosome. *XKRY* was found to be a pseudogene in all species studied and thus is not shown. The *RBMY* gene family harbored two distinct gene variants, each present in multiple copies, within both orangutan species (Fig. S18). Significant gains and losses in ampliconic gene copy number (Note S8) are shown on the phylogenetic tree. Genes showing signatures of purifying selection (Methods) are underlined.


a
Chromosome X
Chromosome Y
Bonobo X
Chimpanzee X
Human X
Gorilla X
B. orangutan X
S. orangutan X
Siamang X
Bonobo Y
Chimpanzee Y
Human Y
Gorilla Y
B. orangutan Y
S. orangutan Y
Siamang Y
Mb
Sequence classes
PAR
Ampliconic
X-transposed
Satellites
αSat
Unclassified
X-degenerate (Y) & X-ancestral (X)
TSPY copies
on the positive strand
on the negative strand
b
Y ampliconic genes
Y X-degenerate genes
BPY2
CDY
DAZ
HSFY
PRY
RBMY
TSPY
VCY
AMELY
TXLNGY
DDX3Y
EIF1AY
KDM5D
MXRA5Y
NLGN4Y
PRKY
RPS4Y1
RPS4Y2
SRY
TBL1Y
TMSB4Y
USP9Y
UTY
ZFY
+RBMY
Bonobo Y
+TSPY
Chimpanzee Y
-RBMY,-TSPY
Human Y
Gorilla Y
-TSPY
-CDY,+TSPY
B. orangutan Y
S. orangutan Y
+CDY,-RBMY,-TSPY
Siamang Y
X-degenerate gene presence/absence
Protein-coding
Pseudogene
Absent

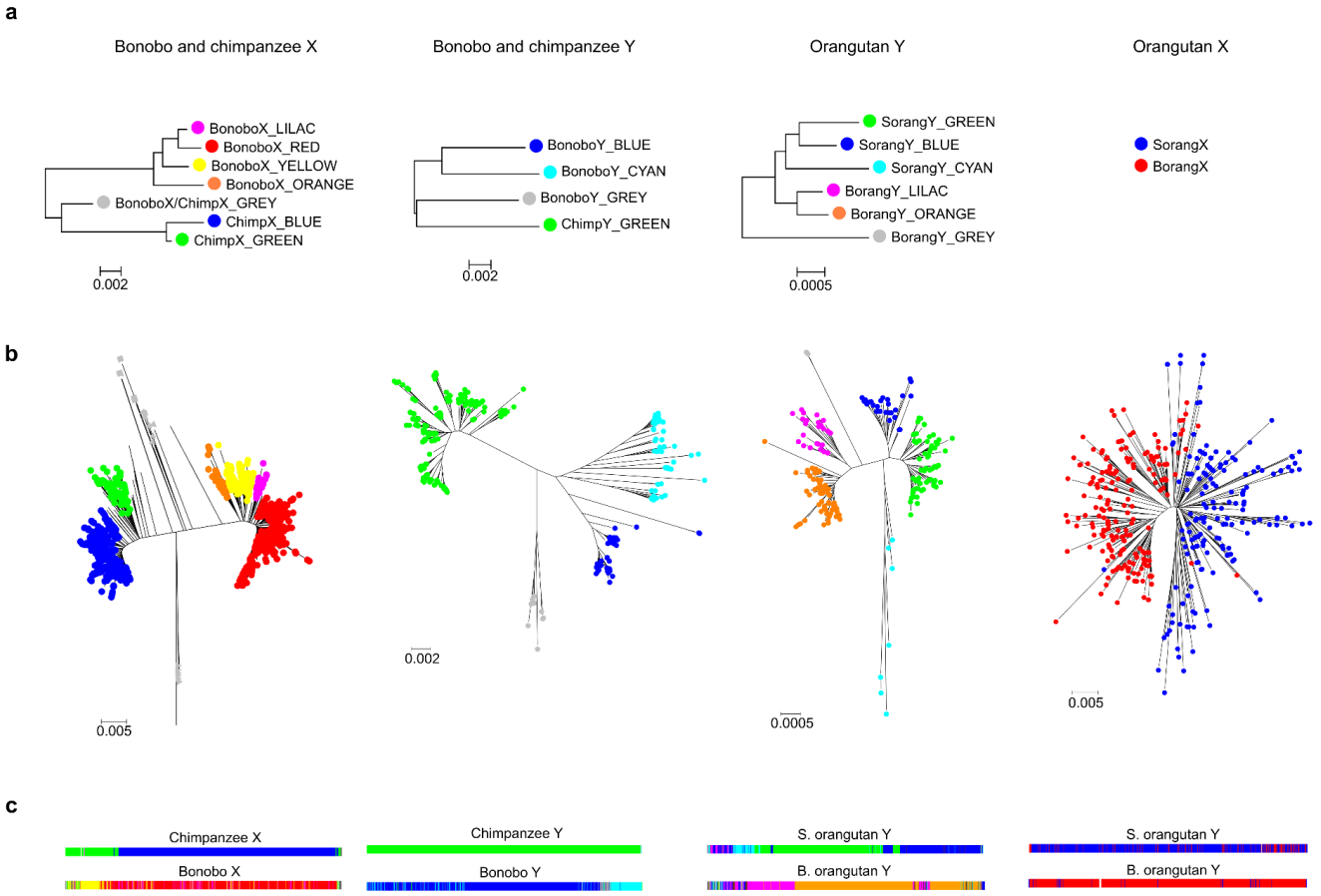
Extended Data Figure 1. Repeats and satellites on the X and Y chromosomes

Repeats and satellites shown with sequence class annotations and CpG methylation for chromosomes X and Y. The scales are different between chromosomes X and Y. The tracks for each species are: (1) sequence class annotation, (2) satellites, (3) inverted repeats, (4) SINEs, (5) LINEs, (6) lineage-specific (LS) insertions of composite repeats (green), transposable elements (blue), and satellites, simple repeats, and low-complexity repeats (pink), and (7) CpG methylation. The inverted repeats, SINEs, and LINEs tracks are plotted in blocks with darker colors representing a higher density. CpG methylation is also displayed on a gradient between dark blue (low methylation) and magenta (high methylation) based on the percentage of supporting aligned ONT reads. The remaining tracks (sequence class, satellites, and LS insertions) are displayed as presence/absence (color/no color). The class and satellite tracks are discrete, whereas the LS insertions are plotted as mini tracks to avoid overplotting where >1 label applies.


Chromosome X
0
40 Mb
80 Mb
120 Mb
160 Mb
Chromosome Y
0
20 Mb
40 Mb
60 Mb
Bonobo X
Bonobo Y
Chimpanzee X
Chimpanzee Y
Human X
Human Y
Gorilla X
Gorilla Y
B. orangutan X
B. orangutan Y
S. orangutan X
S. orangutan Y
Siamang X
Siamang Y
Seq. class
Satellites
Inverted repeats
SINE
LINE
LS insertions
CpG methylation
Inverted repeats density
0
Low
1
High
Satellites
Alpha
SAR/HSAT1A
HSAT1B
LSAU
CER
ACRO1
HSAT2
HSAT3
HSAT4-6
SATR1
SATR2
6kbhsap
Beta
GSAT
SST1
TAR1
StSat
New
Sequence classes
PAR
Ampliconic
Satellites
αSat
X-transposed
X-degenerate (Y) & X-ancestral (X)
Unclassified
Lineage-specific insertions
Composite repeats
Transposable elements
Simple repeats
CpG methylation
50%
100%
SINE/LINE density
Low
High

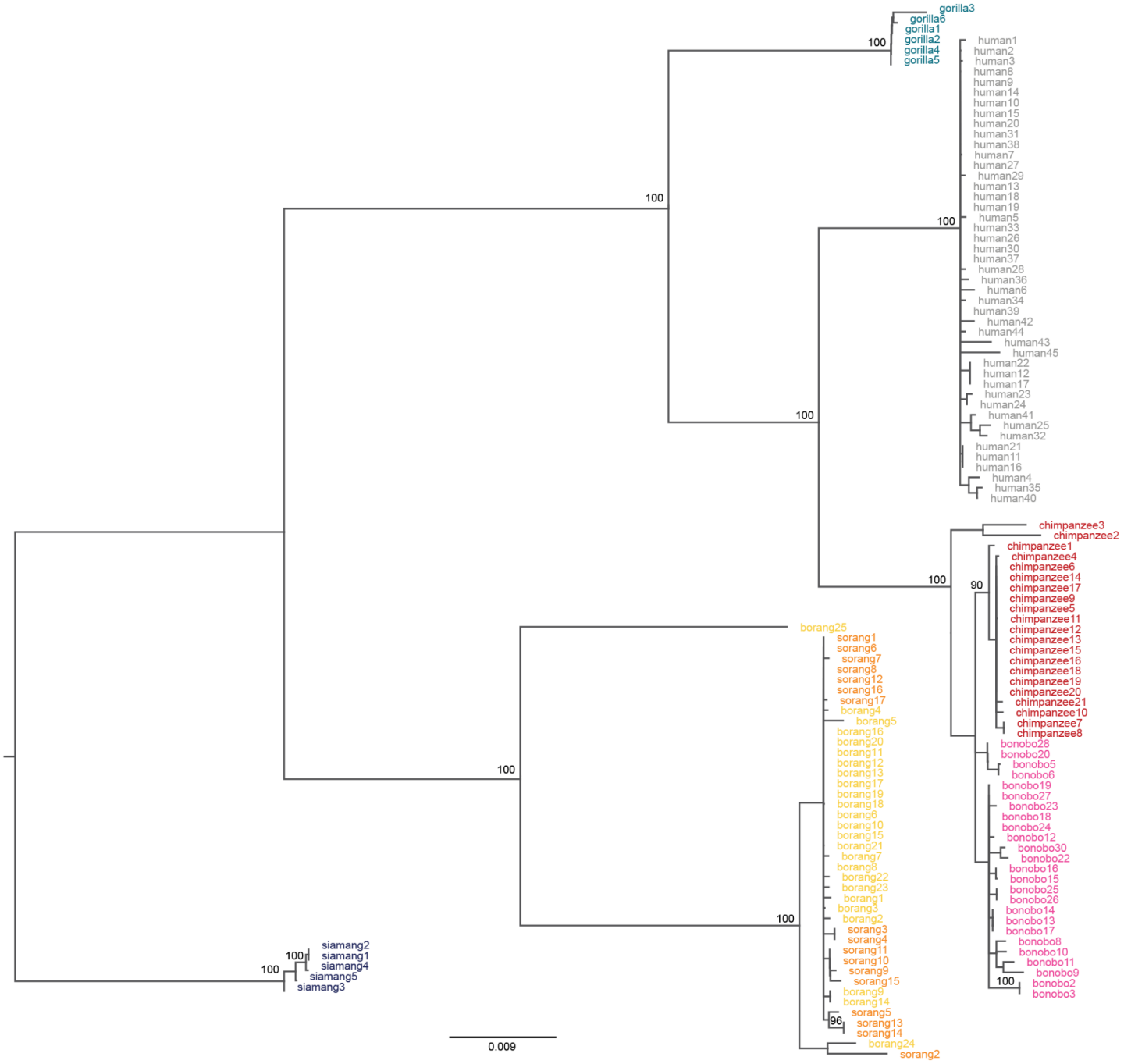
Extended Data Figure 2. Alpha satellite higher order repeat (HOR) haplotypes are species-specific in *Pongo* and *Pan* (except for the few distal HOR copies)

(**A**) Consensus HOR haplotype (HORhap) trees, (**B**) HOR trees, and (**C**) HORhap UCSC Genome Browser annotation tracks for active alpha satellite arrays of chromosomes X and Y in two *Pan* and two *Pongo* species. The details of building the trees and the tracks are in Note S6. Each colored branch in each HOR tree represents a HORhap. All branches in HOR trees are species-specific, except for the GREY branch in Pan cenX tree, for which mixing of chimpanzee (square markers) and bonobo (triangle markers) HORs in the GREY cluster was observed. All branches were used to obtain HORhap consensus sequences and HMMs further used in HMMER-based HORhap classification tool[59] to produce HORhap annotations of the active HOR arrays shown in the Browser tracks. The larger branches with shorter twigs correspond to the younger large active HORhaps, and the smaller branches with longer twigs correspond to the older and smaller side arrays. The arrays corresponding to smaller branches with yet longer twigs are the oldest and often cannot be seen in the tracks; they were located towards the periphery of the arrays. The *Pongo* X tree had a 'star-like' shape and did not have obvious HORhaps; HORs colored by species indicate almost no mixing between species. Analysis of species-specific consensus sequences (this tree is not shown, as there were just two sequences) showed two consistent differences. Thus, we concluded that the species did not share the same HORhaps, but no significant divides could be seen in the tree due to the short HOR length (a 4-mer). The length of the divide in the tree depends on HOR length; with the same degree of divergence there will be more differences between longer HORs (Table in Fig. S13C). The age of the HORhaps is also confirmed by consensus trees where the oldest GREY twigs branch out closer to the root and are nearly equidistant to the active HOR branches of respective species. Thus, GREY HORs likely resemble the HORs that existed in a common ancestor of both species. Note that only in *Pan* cenX such sequences have survived in both species. The younger and more derived consensus HORhaps branch out farther from the root. The values of intra-array divergence, which further confirm the age of the HORhaps, are shown in Note S6. Thus, all but the oldest HORhaps are species-specific, and indicate considerable evolution that occurred after the species diverged.


a
Bonobo and chimpanzee X
BonoboX_LILAC
BonoboX_RED
BonoboX_YELLOW
BonoboX_ORANGE
BonoboX/ChimpX_GREY
ChimpX_BLUE
ChimpX_GREEN
0.002
Bonobo and chimpanzee Y
BonoboY_BLUE
BonoboY_CYAN
BonoboY_GREY
ChimpY_GREEN
0.002
Orangutan Y
SorangY_GREEN
SorangY_BLUE
SorangY_CYAN
BorangY_LILAC
BorangY_ORANGE
BorangY_GREY
0.0005
Orangutan X
SorangX
BorangX
b
0.005
0.002
0.0005
0.005
c
Chimpanzee X
Bonobo X
Chimpanzee Y
Bonobo Y
S. orangutan Y
B. orangutan Y
S. orangutan Y
B. orangutan Y

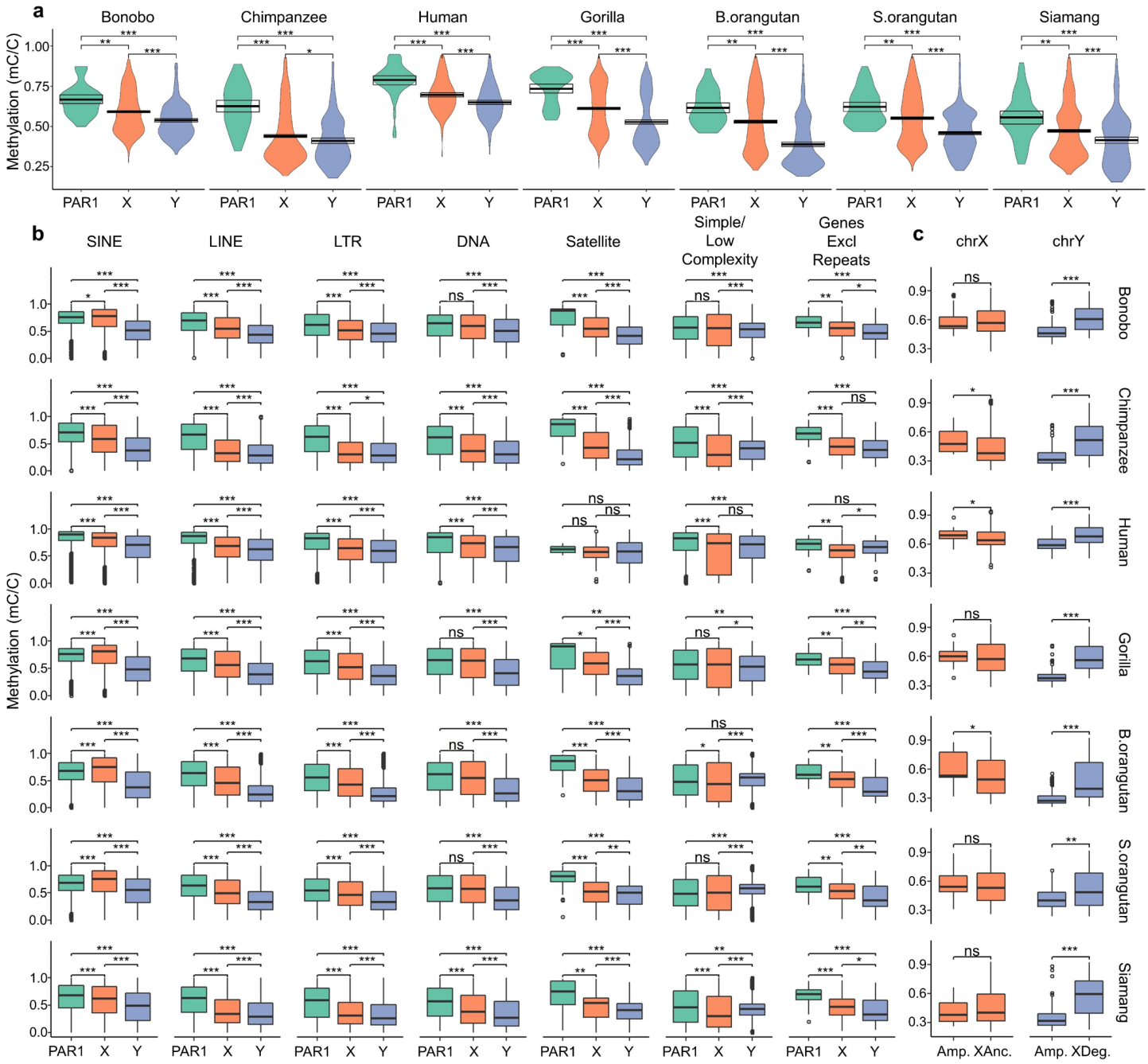
Extended Data Figure 3. Phylogenetic analysis of the *TSPY* gene family

Phylogenetic analysis (Methods) of the protein-coding copies of the *TSPY* gene family in great apes, using siamang as an outgroup, uncovered genus-specific clustering suggesting homogenization among copies. Bootstrap values ≥90% are shown. All but one *TSPY* protein-coding copies in the Bornean orangutan are located in an array with an average distance of 25.2 kb between individual copies, while one copy (id 25, not clustering with the other orangutan *TSPY* copies on the tree) is located 126 kb downstream from the last copy in the array. Five truncated copies of the TSPY gene in bonobo were excluded from the analysis. All sequences were aligned using the Clustal Omega algorithm and the tree was constructed using a neighbor-joining method with the Tamura-Nei genetic distance model as implemented in Geneious Prime[147]. Bootstrap resampling was done with 500 replicates. Bootstrap values higher than 90 were kept in the plot.


gorilla3
gorilla6
gorilla1
gorilla2
gorilla4
gorilla5
100
human1
human2
human3
human8
human9
human14
human10
human15
human20
human31
human38
human7
human27
human29
human13
human18
human19
human5
human33
human26
human30
human37
human28
human36
human6
human34
human39
human42
human44
human43
human45
human22
human12
human17
human23
human24
human41
human25
human32
human21
human11
human16
human4
human35
human40
100
100
100
chimpanzee3
chimpanzee2
chimpanzee1
chimpanzee4
chimpanzee6
chimpanzee14
chimpanzee17
chimpanzee9
chimpanzee5
chimpanzee11
chimpanzee12
chimpanzee13
chimpanzee15
chimpanzee16
chimpanzee18
chimpanzee19
chimpanzee20
chimpanzee21
chimpanzee10
chimpanzee7
chimpanzee8
100
90
bonobo28
bonobo20
bonobo5
bonobo6
bonobo19
bonobo27
bonobo23
bonobo18
bonobo24
bonobo12
bonobo30
bonobo22
bonobo16
bonobo15
bonobo25
bonobo26
bonobo14
bonobo13
bonobo17
bonobo8
bonobo10
bonobo11
bonobo9
bonobo2
bonobo3
100
borang25
sorang1
sorang6
sorang7
sorang8
sorang12
sorang16
sorang17
borang4
borang5
borang16
borang20
borang11
borang12
borang13
borang17
borang19
borang18
borang6
borang10
borang15
borang21
borang7
borang8
borang22
borang23
borang1
borang3
borang2
sorang3
sorang4
sorang11
sorang10
sorang9
sorang15
borang9
borang14
sorang5
96
sorang13
sorang14
borang24
sorang2
100
100
siamang2
siamang1
siamang4
siamang5
siamang3
100
100
0.009

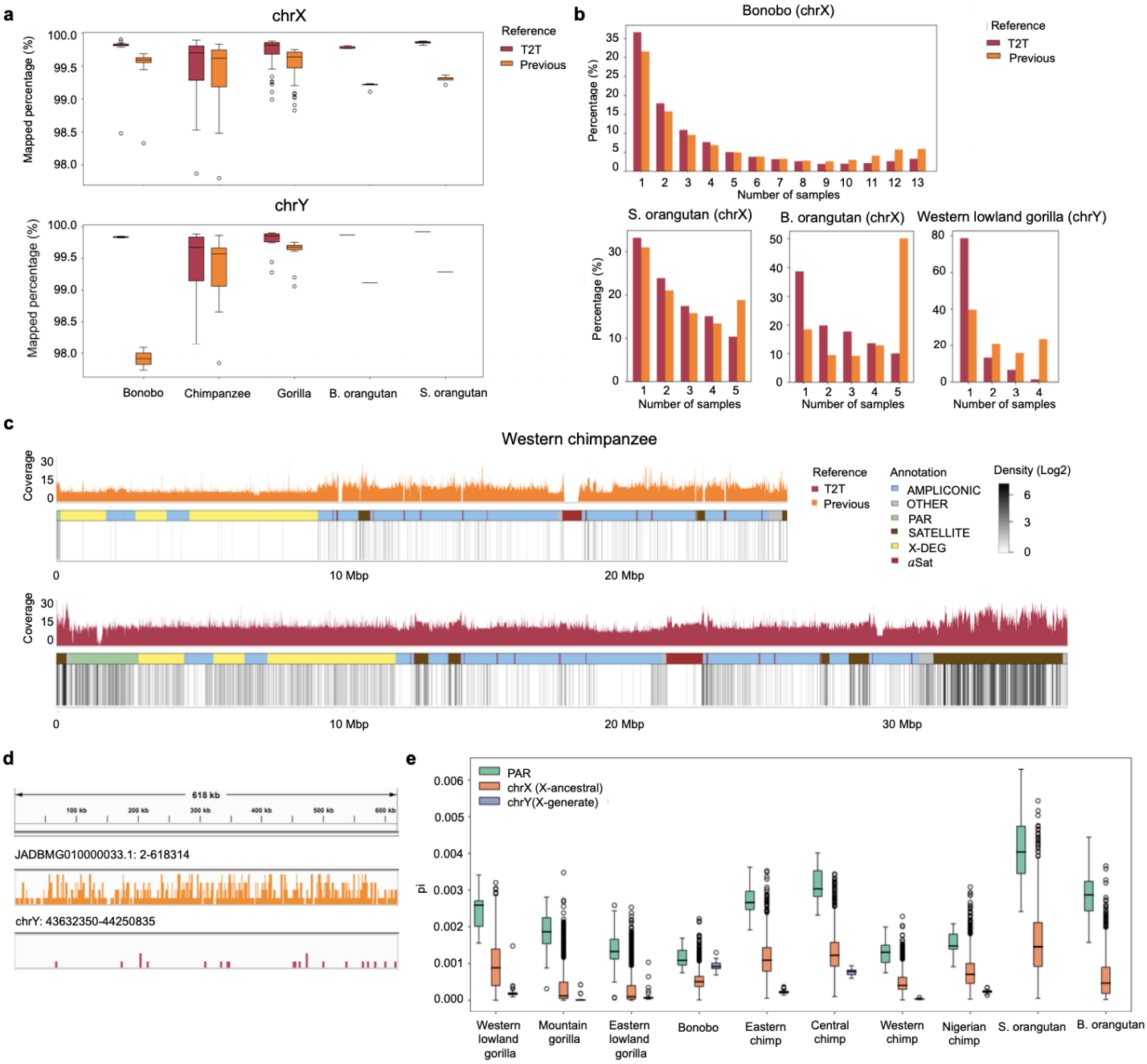
Extended Data Figure 4. Methylation patterns

(**A**) DNA methylation levels in Pseudoautosomal region 1 (PAR1; teal), non-PAR chromosome X (orange), and non-PAR chromosome Y (periwinkle). *p*-values were determined using the Wilcoxon rank-sum tests. * $p<0.05$; ** $p<10^{-3}$; *** $p<10^{-6}$. (**B**) Differences in DNA methylation levels between different repeat categories as well as protein-coding genes (after excluding repetitive sequences). (**C**) Differences in methylation levels between ampliconic and X-ancestral/X-degenerate regions in the X and the Y chromosomes (in 100-kb bins).



Extended Data Figure 5. T2T assemblies facilitate short-read mapping and enable the analysis of genetic diversity in great apes

(**A**) The percentage of short reads mapped to T2T vs. previous sex chromosome assemblies (using the previous reference assembly of Sumatran orangutan for Bornean orangutan data). (**B**) Allele frequencies (y-axis) of variants called from reads mapped to T2T vs. previous assemblies. (**C**) Coverage and variant density (in log2 values of densities per 10 kb) distribution across previous (shown in the reverse orientation) and T2T assemblies for western chimpanzee. Peak variant densities were observed at 5.9 for previous chrY, and at 7.6 for T2T chrY. (**D**) Distributions of variant allele frequencies on JADBMG010000033.1 (positions 2 to 618,314, upper), a contig from a previous chrY assembly, and T2T chrY (positions 43,632,350 to 44,250,835, bottom), for western lowland gorilla, visualized using IGV. (**E**) Nucleotide diversity (pi)[83] in pseudoautosomal regions (PARs), X-ancestral regions of chromosome X, and X-degenerate (X-DEG) regions of chromosome Y. ‘Chimp’ stands for chimpanzee. ‘Nigerian chimp’ stands for Nigeria-Cameroon chimpanzee.


a
chrX
Mapped percentage (%)
Reference
T2T
Previous
chrY
Bonobo
Chimpanzee
Gorilla
B. orangutan
S. orangutan
b
Bonobo (chrX)
Percentage (%)
Number of samples
S. orangutan (chrX)
B. orangutan (chrX)
Western lowland gorilla (chrY)
c
Western chimpanzee
Coverage
10 Mbp
20 Mbp
30 Mbp
Annotation
AMPLICONIC
OTHER
PAR
SATELLITE
X-DEG
αSat
Density (Log2)
d
618 kb
JADBMG010000033.1: 2-618314
chrY: 43632350-44250835
e
PAR
chrX (X-ancestral)
chrY(X-generate)
pi
Western lowland gorilla
Mountain gorilla
Eastern lowland gorilla
Bonobo
Eastern chimp
Central chimp
Western chimp
Nigerian chimp
S. orangutan
B. orangutan

References

1. Veyrunes, F. *et al.* Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. *Genome Res.* **18**, 965–973 (2008).
2. Bellott, D. W. *et al.* Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* **508**, 494–499 (2014).
3. Sinclair, A. H. *et al.* A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *Nature* **346**, 240–244 (1990).
4. Betrán, E., Demuth, J. P. & Williford, A. Why Chromosome Palindromes? *International Journal of Evolutionary Biology* vol. 2012 1–14 Preprint at https://doi.org/10.1155/2012/207958 (2012).
5. Lahn, B. T. & Page, D. C. Four evolutionary strata on the human X chromosome. *Science* **286**, 964–967 (1999).
6. Miga, K. H. *et al.* Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020).
7. Rhie, A. *et al.* The complete sequence of a human Y chromosome. *Nature* (2023) doi:10.1038/s41586-023-06457-y.
8. Rozen, S. *et al.* Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* **423**, 873–876 (2003).
9. Trombetta, B. & Cruciani, F. Y chromosome palindromes and gene conversion. *Hum. Genet.* **136**, 605–619 (2017).
10. Tomaszkiewicz, M., Medvedev, P. & Makova, K. D. Y and W Chromosome Assemblies: Approaches and Discoveries. *Trends Genet.* **33**, 266–282 (2017).
11. 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).
12. Hughes, J. F. *et al.* Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* **463**, 536–539 (2010).
13. 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).
14. Zhou, Y. *et al.* Eighty million years of rapid evolution of the primate Y chromosome. *Nat Ecol Evol* (2023) doi:10.1038/s41559-022-01974-x.
15. Cechova, M. *et al.* Dynamic evolution of great ape Y chromosomes. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 26273–26280 (2020).
16. Gordon, D. *et al.* Long-read sequence assembly of the gorilla genome. *Science* **352**, aae0344 (2016).
17. Mao, Y. *et al.* A high-quality bonobo genome refines the analysis of hominid evolution. *Nature* **594**, 77–81 (2021).
18. Gläser, B. *et al.* Simian Y chromosomes: species-specific rearrangements of DAZ, RBM, and TSPY versus contiguity of PAR and SRY. *Mamm. Genome* **9**, 226–231 (1998).
19. Hallast, P. & Jobling, M. A. The Y chromosomes of the great apes. *Hum. Genet.* **136**, 511–528 (2017).
20. Glazko, G. V. & Nei, M. Estimation of divergence times for major lineages of primate species. *Mol. Biol. Evol.* **20**, 424–434 (2003).
21. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
22. Locke, D. P. *et al.* Comparative and demographic analysis of orang-utan genomes. *Nature* **469**, 529–533 (2011).
23. Mattle-Greminger, M. P. *et al.* Genomes reveal marked differences in the adaptive evolution between orangutan species. *Genome Biol.* **19**, 193 (2018).
24. Brand, C. M., White, F. J., Rogers, A. R. & Webster, T. H. Estimating bonobo () and chimpanzee () evolutionary history from nucleotide site patterns. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2200858119 (2022).
25. Cahill, J. A., Soares, A. E. R., Green, R. E. & Shapiro, B. Inferring species divergence times using pairwise sequential Markovian coalescent modelling and low-coverage genomic data. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **371**, (2016).
26. Wegmann, D. & Excoffier, L. Bayesian inference of the demographic history of chimpanzees. *Mol. Biol. Evol.* **27**, 1425–1435 (2010).
27. Stone, A. C. *et al.* More reliable estimates of divergence times in Pan using complete mtDNA sequences and accounting for population structure. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **365**, 3277–3288 (2010).

28. Carbone, L. *et al.* Gibbon genome and the fast karyotype evolution of small apes. *Nature* **513**, 195–201 (2014).
29. Anderson, M. J. & Dixson, A. F. Sperm competition: motility and the midpiece in primates. *Nature* **416**, 496 (2002).
30. Rautiainen, M. *et al.* Verkko: telomere-to-telomere assembly of diploid chromosomes. *bioRxiv* (2022) doi:10.1101/2022.06.24.497523.
31. Kronenberg, Z. N. *et al.* High-resolution comparative analysis of great ape genomes. *Science* **360**, (2018).
32. Weissensteiner, M. H. *et al.* Accurate sequencing of DNA motifs able to form alternative (non-B) structures. *Genome Res.* **33**, 907–922 (2023).
33. Makova, K. D. & Li, W.-H. Strong male-driven evolution of DNA sequences in humans and apes. *Nature* **416**, 624–626 (2002).
34. Li, W. H., Yi, S. & Makova, K. Male-driven evolution. *Curr. Opin. Genet. Dev.* **12**, 650–656 (2002).
35. Bergeron, L. A. *et al.* Evolution of the germline mutation rate across vertebrates. *Nature* **615**, 285–291 (2023).
36. Jónsson, H. *et al.* Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519–522 (2017).
37. Agarwal, I. & Przeworski, M. Signatures of replication timing, recombination, and sex in the spectrum of rare variants on the human X chromosome and autosomes. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 17916–17924 (2019).
38. Ezawa, K., OOta, S. & Saitou, N. Genome-Wide Search of Gene Conversions in Duplicated Genes of Mouse and Rat. *Mol. Biol. Evol.* **23**, 927–940 (2006).
39. Hallast, P., Balaresque, P., Bowden, G. R., Ballereau, S. & Jobling, M. A. Recombination dynamics of a human Y-chromosomal palindrome: rapid GC-biased gene conversion, multi-kilobase conversion tracts, and rare inversions. *PLoS Genet.* **9**, e1003666 (2013).
40. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
41. Hoyt, S. J. *et al.* From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science* **376**, eabk3112 (2022).
42. Rizzon, C., Marais, G., Gouy, M. & Biémont, C. Recombination rate and the distribution of transposable elements in the Drosophila melanogaster genome. *Genome Res.* **12**, 400–407 (2002).
43. Chow, J. C. *et al.* LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. *Cell* **141**, 956–969 (2010).
44. Koga, A., Hirai, Y., Hara, T. & Hirai, H. Repetitive sequences originating from the centromere constitute large-scale heterochromatin in the telomere region in the siamang, a small ape. *Heredity* **109**, 180–187 (2012).
45. Cellamare, A. *et al.* New insights into centromere organization and evolution from the white-cheeked gibbon and marmoset. *Mol. Biol. Evol.* **26**, 1889–1900 (2009).
46. Ventura, M. *et al.* Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee. *Genome Res.* **21**, 1640–1649 (2011).
47. Makova, K. D. & Weissensteiner, M. H. Noncanonical DNA structures are drivers of genome evolution. *Trends Genet.* **39**, 109–124 (2023).
48. Wang, G. & Vasquez, K. M. Dynamic alternative DNA structures in biology and disease. *Nat. Rev. Genet.* **24**, 211–234 (2023).
49. Halder, R. *et al.* Guanine quadruplex DNA structure restricts methylation of CpG dinucleotides genome-wide. *Mol. Biosyst.* **6**, 2439–2447 (2010).
50. Kim, J.-H. *et al.* Human gamma-satellite DNA maintains open chromatin structure and protects a transgene from epigenetic silencing. *Genome Res.* **19**, 533–544 (2009).
51. Meneveri, R., Agresti, A., Rocchi, M., Marozzi, A. & Ginelli, E. Analysis of GC-rich repetitive nucleotide sequences in great apes. *J. Mol. Evol.* **40**, 405–412 (1995).
52. Mukherjee, A. K., Sharma, S. & Chowdhury, S. Non-duplex G-Quadruplex Structures Emerge as Mediators of Epigenetic Modifications. *Trends Genet.* **35**, 129–144 (2019).
53. Meneveri, R. *et al.* Molecular organization and chromosomal location of human GC-rich heterochromatic blocks. *Gene* **123**, 227–234 (1993).
54. Bacolla, A. & Wells, R. D. Non-B DNA conformations as determinants of mutagenesis and human disease. *Mol. Carcinog.* **48**, 273–285 (2009).

55. Kasinathan, S. & Henikoff, S. Non-B-Form DNA Is Enriched at Centromeres. *Mol. Biol. Evol.* **35**, 949–962 (2018).
56. Shepelev, V. A., Alexandrov, A. A., Yurov, Y. B. & Alexandrov, I. A. The evolutionary origin of man can be traced in the layers of defunct ancestral alpha satellites flanking the active centromeres of human chromosomes. *PLoS Genet.* **5**, e1000641 (2009).
57. Alexandrov, I., Kazakov, A., Tumeneva, I., Shepelev, V. & Yurov, Y. Alpha-satellite DNA of primates: old and new families. *Chromosoma* **110**, 253–266 (2001).
58. Miga, K. H. & Alexandrov, I. A. Variation and Evolution of Human Centromeres: A Field Guide and Perspective. *Annu. Rev. Genet.* **55**, 583–602 (2021).
59. Altemose, N. *et al.* Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022).
60. Hughes, J. F., Skaletsky, H. & Page, D. C. ALRY-MAJOR:PT: Major repeat unit of chimpanzee alpha repetitive DNA from the Y chromosome centromere - a consensus. *Direct submission to Repbase Update* https://www.girinst.org/ (2004).
61. Gershman, A. *et al.* Epigenetic patterns in a complete human genome. *Science* **376**, eabj5089 (2022).
62. Jørgensen, A. L., Laursen, H. B., Jones, C. & Bak, A. L. Evolutionarily different alphoid repeat DNA on homologous chromosomes in human and chimpanzee. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 3310–3314 (1992).
63. Greve, G. *et al.* Y-Chromosome variation in hominids: intraspecific variation is limited to the polygamous chimpanzee. *PLoS One* **6**, e29311 (2011).
64. Ledbetter, D. H. NOR-bearing Y chromosome in a primate, Hylobates (Symphalangus) syndactylus. *Cytogenet. Cell Genet.* **29**, 250–252 (1981).
65. Hori, Y., Shimamoto, A. & Kobayashi, T. The human ribosomal DNA array is composed of highly homogenized tandem clusters. *Genome Res.* **31**, 1971–1982 (2021).
66. Potapova, T. A. *et al.* Superresolution microscopy reveals linkages between ribosomal DNA on heterologous chromosomes. *J. Cell Biol.* **218**, 2492–2513 (2019).
67. Sanij, E. *et al.* UBF levels determine the number of active ribosomal RNA genes in mammals. *J. Cell Biol.* **183**, 1259–1274 (2008).
68. Vegesna, R. *et al.* Ampliconic Genes on the Great Ape Y Chromosomes: Rapid Evolution of Copy Number but Conservation of Expression Levels. *Genome Biol. Evol.* **12**, 842–859 (2020).
69. Bonito, M. *et al.* New insights into the evolution of human Y chromosome palindromes through mutation and gene conversion. *Hum. Mol. Genet.* **30**, 2272–2285 (2021).
70. Wang, J. *et al.* Sex-specific gene expression in the blood of four primates. *Genomics* **113**, 2605–2613 (2021).
71. Rivard, E. L. *et al.* A putative de novo evolved gene required for spermatid chromatin condensation in Drosophila melanogaster. *PLoS Genet.* **17**, e1009787 (2021).
72. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 (2012).
73. Hellman, A. & Chess, A. Gene body-specific methylation on the active X chromosome. *Science* **315**, 1141–1143 (2007).
74. Sharp, A. J. *et al.* DNA methylation profiles of human active and inactive X chromosomes. *Genome Res.* **21**, 1592–1600 (2011).
75. Singh, D. *et al.* Koala methylomes reveal divergent and conserved DNA methylation signatures of X chromosome regulation. *Proc. Biol. Sci.* **288**, 20202244 (2021).
76. Sun, D., Maney, D. L., Layman, T. S., Chatterjee, P. & Yi, S. V. Regional epigenetic differentiation of the Z Chromosome between sexes in a female heterogametic system. *Genome Res.* **29**, 1673–1684 (2019).
77. Sigurdsson, M. I., Smith, A. V., Bjornsson, H. T. & Jonsson, J. J. HapMap methylation-associated SNPs, markers of germline DNA methylation, positively correlate with regional levels of human meiotic recombination. *Genome Res.* **19**, 581–589 (2009).
78. Zeng, J. & Yi, S. V. Specific modifications of histone tails, but not DNA methylation, mirror the temporal variation of mammalian recombination hotspots. *Genome Biol. Evol.* **6**, 2918–2929 (2014).
79. Weber, M. *et al.* Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.* **39**, 457–466 (2007).
80. Elango, N. & Yi, S. V. DNA methylation and structural and functional bimodality of vertebrate promoters.

Mol. Biol. Evol. **25**, 1602–1608 (2008).
81. Aganezov, S. *et al.* A complete reference genome improves analysis of human genetic variation. *Science* **376**, eabl3533 (2022).
82. Audano, P. A. *et al.* Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **176**, 663–675.e19 (2019).
83. Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. U. S. A.* **76**, 5269–5273 (1979).
84. Kuhlwilm, M. *et al.* Evolution and demography of the great apes. *Curr. Opin. Genet. Dev.* **41**, 124–129 (2016).
85. de Manuel, M. *et al.* Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science* **354**, 477–481 (2016).
86. Xue, Y. *et al.* Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science* **348**, 242–245 (2015).
87. Oetjens, M. T., Shen, F., Emery, S. B., Zou, Z. & Kidd, J. M. Y-chromosome structural diversity in the bonobo and chimpanzee lineages. *bioRxiv* (2015) doi:10.1101/029702.
88. Wilson Sayres, M. A., Lohmueller, K. E. & Nielsen, R. Natural selection reduced diversity on human y chromosomes. *PLoS Genet.* **10**, e1004064 (2014).
89. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
90. Pawar, H. *et al.* Ghost admixture in eastern gorillas. *Nat Ecol Evol* **7**, 1503–1514 (2023).
91. Vigilant, L. & Bradley, B. J. Genetic variation in gorillas. *Am. J. Primatol.* **64**, (2004).
92. Wilson Sayres, M. A., Lohmueller, K. E. & Nielsen, R. Natural selection reduced diversity on human y chromosomes. *PLoS Genet.* **10**, e1004064 (2014).
93. Maan, A. A. *et al.* The Y chromosome: a blueprint for men's health? *Eur. J. Hum. Genet.* **25**, 1181–1188 (2017).
94. Vogt, M. H. J. *et al.* UTY gene codes for an HLA-B60–restricted human male-specific minor histocompatibility antigen involved in stem cell graft rejection: characterization of the critical polymorphic amino acid residues for T-cell recognition. *Blood* **96**, 3126–3132 (2000).
95. Graves, J. A. M. Sex chromosome specialization and degeneration in mammals. *Cell* **124**, 901–914 (2006).
96. Graves, J. A. M., Koina, E. & Sankovic, N. How the gene content of human sex chromosomes evolved. *Curr. Opin. Genet. Dev.* **16**, 219–224 (2006).
97. Hughes, J. F. *et al.* Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* **483**, 82–86 (2012).
98. Charlesworth, B. & Charlesworth, D. The degeneration of Y chromosomes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **355**, 1563–1572 (2000).
99. Lemos, B., Branco, A. T. & Hartl, D. L. Epigenetic effects of polymorphic Y chromosomes modulate chromatin components, immune response, and sexual conflict. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 15826–15831 (2010).
100. Ferree, P. M. & Barbash, D. A. Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in Drosophila. *PLoS Biol.* **7**, e1000234 (2009).
101. Novo, C. *et al.* The heterochromatic chromosome caps in great apes impact telomere metabolism. *Nucleic Acids Res.* **41**, 4792–4801 (2013).
102. Bass, H. W. *et al.* Evidence for the coincident initiation of homolog pairing and synapsis during the telomere-clustering (bouquet) stage of meiotic prophase. *J. Cell Sci.* **113 (Pt 6)**, 1033–1042 (2000).
103. Acquaviva, L. *et al.* Ensuring meiotic DNA break formation in the mouse pseudoautosomal region. *Nature* **582**, 426–431 (2020).
104. Sudmant, P. H. *et al.* Evolution and diversity of copy number variation in the great ape lineage. *Genome Res.* **23**, 1373–1382 (2013).
105. Guarracino, A. *et al.* Recombination between heterologous human acrocentric chromosomes. *Nature* **617**, 335–343 (2023).
106. van Sluis, M., van Vuuren, C., Mangan, H. & McStay, B. NORs on human acrocentric chromosome p-arms are active by default and can associate with nucleoli independently of rDNA. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 10368–10377 (2020).

107. Jiwrajka, N. & Anguera, M. C. The X in seX-biased immunity and autoimmune rheumatic disease. *J. Exp. Med.* **219**, (2022).
108. Ross, M. T. *et al.* The DNA sequence of the human X chromosome. *Nature* **434**, 325–337 (2005).
109. Lange, J. *et al.* Isodicentric Y chromosomes and sex disorders as byproducts of homologous recombination that maintains palindromes. *Cell* **138**, 855–869 (2009).
110. Lange, J. *et al.* Intrachromosomal homologous recombination between inverted amplicons on opposing Y-chromosome arms. *Genomics* **102**, 257–264 (2013).
111. Hallast, P. *et al.* Assembly of 43 human Y chromosomes reveals extensive complexity and variation. *Nature* **621**, 355–364 (2023).
112. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350–3352 (2015).
113. Jain, C., Koren, S., Dilthey, A., Phillippy, A. M. & Aluru, S. A fast adaptive algorithm for computing whole-genome homology maps. *Bioinformatics* **34**, i748–i756 (2018).
114. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
115. Mc Cartney, A. M. *et al.* Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *Nat. Methods* **19**, 687–695 (2022).
116. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
117. Robinson, J. T., Thorvaldsdóttir, H., Wenger, A. M., Zehir, A. & Mesirov, J. P. Variant Review with the Integrative Genomics Viewer. *Cancer Res.* **77**, e31–e34 (2017).
118. Armstrong, J. *et al.* Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246–251 (2020).
119. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
120. Harris, R. S. *Improved Pairwise Alignment of Genomic DNA*. (2007).
121. Siepel, A. & Haussler, D. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.* **21**, 468–488 (2004).
122. Numanagic, I. *et al.* Fast characterization of segmental duplications in genome assemblies. *Bioinformatics* **34**, i706–i714 (2018).
123. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
124. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **Chapter 4**, 4.10.1–4.10.14 (2009).
125. Morgulis, A., Gertz, E. M., Schäffer, A. A. & Agarwala, R. WindowMasker: window-based masker for sequenced genomes. *Bioinformatics* **22**, 134–141 (2006).
126. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).
127. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, (2021).
128. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
129. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
130. Storer, J., Hubley, R., Rosen, J., Wheeler, T. J. & Smit, A. F. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob. DNA* **12**, 2 (2021).
131. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
132. Tempel, S. Using and understanding RepeatMasker. *Methods Mol. Biol.* **859**, 29–51 (2012).
133. Olson, D. & Wheeler, T. ULTRA: A Model Based Tool to Detect Tandem Repeats. *ACM BCB* **2018**, 37–46 (2018).
134. Cechova, M. *et al.* High satellite repeat turnover in great apes studied with short- and long-read technologies. *Mol. Biol. Evol.* (2019) doi:10.1093/molbev/msz156.
135. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
136. Storer, J. M., Hubley, R., Rosen, J. & Smit, A. F. A. Curation Guidelines for de novo Generated Transposable Element Families. *Curr Protoc* **1**, e154 (2021).
137. Hickey, G., Paten, B., Earl, D., Zerbino, D. & Haussler, D. HAL: a hierarchical format for storing and

analyzing multiple genome alignments. *Bioinformatics* **29**, 1341–1342 (2013).
138. Sahakyan, A. B. *et al.* Machine learning model for sequence-driven DNA G-quadruplex formation. *Sci. Rep.* **7**, 14535 (2017).
139. Rhie, A. *et al.* Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).
140. Pruitt, K. D. *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* **42**, D756–63 (2014).
141. Kapustin, Y., Souvorov, A., Tatusova, T. & Lipman, D. Splign: algorithms for computing spliced alignments with identification of paralogs. *Biol. Direct* **3**, 20 (2008).
142. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
143. Kalvari, I. *et al.* Non-Coding RNA Analysis Using the Rfam Database. *Curr. Protoc. Bioinformatics* **62**, e51 (2018).
144. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
145. Chan, P. P. & Lowe, T. M. tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. *Methods Mol. Biol.* **1962**, 1–14 (2019).
146. Fiddes, I. T. *et al.* Comparative Annotation Toolkit (CAT)-simultaneous clade and personal genome annotation. *Genome Res.* **28**, 1029–1038 (2018).
147. Kearse, M. *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
148. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare. *F1000Res.* **9**, (2020).
149. Sawyer, S. Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* **6**, 526–538 (1989).
150. Murrell, B. *et al.* Gene-wide identification of episodic selection. *Mol. Biol. Evol.* **32**, 1365–1371 (2015).
151. Wisotsky, S. R., Kosakovsky Pond, S. L., Shank, S. D. & Muse, S. V. Synonymous Site-to-Site Substitution Rate Variation Dramatically Inflates False Positive Rates of Selection Analyses: Ignore at Your Own Peril. *Mol. Biol. Evol.* **37**, 2430–2439 (2020).
152. Kosakovsky Pond, S. L. *et al.* HyPhy 2.5-A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. *Mol. Biol. Evol.* **37**, 295–299 (2020).
153. Jain, C., Rhie, A., Hansen, N. F., Koren, S. & Phillippy, A. M. Long-read mapping to repetitive reference sequences using Winnowmap2. *Nat. Methods* **19**, 705–710 (2022).
154. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
155. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
156. Krumsiek, J., Arnold, R. & Rattei, T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**, 1026–1028 (2007).