Dynamic evolution of great ape Y chromosomes

Monika Cechova[1#], Rahulsimham Vegesna[2#], Marta Tomaszkiewicz[1], Robert S. Harris[1], Di Chen[1], Samarth Rangavittal[2], Paul Medvedev[3*], Kateryna D. Makova[1*]

[1]Department of Biology, Penn State University, University Park, PA 16802
[2]Intercollege Graduate Program in Bioinformatics and Genomics, Penn State University, University Park, PA 16802
[3]Departments of Computer Science and Engineering, and of Biochemistry and Molecular Biology, Penn State University, University Park, PA 16802
[#]These authors contributed equally
*To whom correspondence should be addressed at kdm16@psu.edu or pzm11@psu.edu

Abstract

The mammalian male-specific Y chromosome plays a critical role in sex determination and male fertility. However, because of its repetitive and haploid nature, it is frequently absent from genome assemblies and remains enigmatic. The Y chromosomes of great apes represent a particular puzzle: their gene content is more similar between human and gorilla than between human and chimpanzee, even though human and chimpanzee shared a more recent common ancestor. To solve this puzzle, here we constructed a dataset including Ys from all extant great ape genera. We generated assemblies of bonobo and orangutan Ys, from short and long sequencing reads, and aligned them with the publicly available human, chimpanzee and gorilla Y assemblies. Analyzing this dataset, we found that the genus *Pan*, including chimpanzee and bonobo, experienced accelerated substitution rates. Additionally, *Pan* also exhibited elevated gene death rates. These observations are consistent with high levels of sperm competition in *Pan*. Furthermore, we inferred that the great ape common ancestor already possessed multi-copy sequences homologous to most human and chimpanzee palindromes. Nonetheless, each species also acquired distinct ampliconic sequences. We also detected increased chromatin contacts between and within palindromes (from Hi-C data), likely facilitating gene conversion and structural rearrangements. Moreover, our ENCODE data analysis suggested that Y palindromes exist to promote gene conversion preventing degradation of not only genes, as is commonly believed, but also gene regulatory sites. Our results highlight the dynamic mode of Y chromosome evolution, and open avenues for studies of male-specific dispersal in endangered great ape species.

Introduction

The mammalian male-specific sex chromosome—the Y—is vital for sex determination and male fertility, and is a useful marker for population genetics studies. It carries *SRY*, a gene encoding the testis-determining factor that initiates male sex determination (1). The human Y also harbors azoospermia factor regions, deletions of which can result in infertility (2). Y chromosome sequences have been used to analyze male dispersal (3) and to unravel sex bias during hybridization (4, 5) in natural populations. Thus, the Y is important biologically and its sequences have critical practical implications. Moreover, study of the Y is needed to obtain a complete picture of mammalian genome evolution. Yet, due to its repetitive structure and haploid nature, the Y has been sequenced and assembled for only a handful of mammalian species (6).

Among great apes, the Y has so far been assembled only in human (7), chimpanzee (8), and gorilla (9). A comparative study of these Y assemblies (9) uncovered some unexpected patterns which could not be explained with the data from three species alone. Despite a recent divergence of these species (~7 million years ago, MYA (10)), their Y chromosomes differ enormously in size and gene content, in sharp contrast with the stability of the rest of the genome. For instance, the chimpanzee Y is only half the size of the human Y, and the percentage of gene families shared by these two chromosomes (68%) that split ~6 MYA (10) is similar to that shared by human and chicken autosomes that split ~310 MYA (8, 11). Puzzlingly, in terms of shared genes and overall architecture, the human Y is more similar to the gorilla Y than to the chimpanzee Y even though human and chimpanzee have a more recent common ancestor (9). Y chromosomes from additional great ape species should be sequenced to understand whether high interspecific variability in gene content and architecture is characteristic of all great ape Ys, and not just of the three Ys previously assembled.

All great ape Y chromosomes studied thus far include *pseudoautosomal regions (PARs)*, which recombine with the X chromosome, and *male-specific* X-degenerate, ampliconic, and heterochromatic regions, which evolve as a single linkage unit (7). The *X-degenerate regions* are composed of segments with different levels of homology to the X chromosome—*strata*, which correspond to stepwise losses of recombination on the Y. Because of lack of recombination, X-degenerate regions in particular are expected to accumulate gene-disrupting mutations, however this has not been previously examined in detail. The *ampliconic regions* consist of repetitive sequences that have >50% identity to each other (7) and contain *palindromes*—inverted repeats (separated by a spacer) up to several megabases long whose arms are >99.9% identical (12). Palindromes are thought to evolve to allow intrachromosomal (Y-Y) gene conversion which rescues the otherwise non-recombining male-specific regions on the Y from deleterious mutations (13). We presently lack knowledge about how conserved palindrome sequences are across great apes. The *heterochromatic regions* are rich in satellite repeats (14). In general, X-degenerate regions are more conserved, whereas ampliconic regions are prone to rearrangements, and heterochromatic regions evolve very rapidly, among species (8, 9, 14). However, with the exception of a recent study of heterochromatic regions (14), the evolution of great ape Y chromosomes outside of human, chimpanzee, and gorilla has not been explored.

Known male-specific Y chromosome protein-coding genes are located in either the X-degenerate or ampliconic regions. *X-degenerate genes* (16 on the human Y) are single-copy, ubiquitously expressed genes that fulfill housekeeping functions (15). Multi-copy *ampliconic genes* (nine gene families on the human Y, eight of which—all but *TSPY*—are located in palindromes) are expressed only in testis and function during spermatogenesis (7). Notably, some human Y chromosome genes are deleted or pseudogenized in other great apes, and thus are not essential for all species (9, 16, 17). To illuminate genes essential for male reproduction in non-human great apes, all of which are endangered species, one needs to conduct a comprehensive, cross-species analysis of Y gene content evolution, yet such an analysis has been missing.

In the present study we performed a detailed comparative analysis of the Y chromosome in five species representing all four great ape genera: the human (*Homo*) lineage diverged from the chimpanzee (*Pan*), gorilla (*Gorilla*), and orangutan (*Pongo*) lineages ~6 MYA, ~7 MYA, and ~13 MYA, respectively (10), and the bonobo and chimpanzee lineages (both belonging to the genus *Pan*), which diverged ~0.77-1.8 MYA (18, 19). We produced draft assemblies of the bonobo and Sumatran orangutan Ys, and combined them with the publicly available human, chimpanzee and gorilla Y assemblies to construct great ape Y multi-species alignments. This comprehensive data set enabled us to answer several pivotal questions about evolution of great ape Y chromosomes. First, we assessed lineage-specific substitution rates on the Y, and identified species experiencing significant rate acceleration. Second, we evaluated the conservation of palindromic sequences across great apes at the sequence level, and examined chromatin interactions within Y ampliconic regions. Third, we determined interspecific gene content turnover on the Y. Our results highlight the highly dynamic nature of great ape Y chromosome evolution and contradict the view that these chromosomes represent an evolutionary dead end (20).

Results

Assemblies

To obtain Y chromosome assemblies for all major great ape lineages, we augmented publicly available human, chimpanzee, and gorilla Y assemblies (7–9, 21) by producing draft bonobo and Sumatran orangutan (henceforth called ‘orangutan’) Y assemblies (**Fig. S1**, see Methods for details). For the latter two species, we generated and assembled (22) deep-coverage short sequencing reads from male individuals, and identified putative Y contigs by mapping them against the corresponding female reference assemblies (23). These contigs were then scaffolded with mate-pair reads (24). The orangutan Y assembly was further improved by merging (25) with another high-quality assembly generated with 10×Genomics technology (26). The bonobo Y assembly was improved by additional scaffolding with long Y-enriched Pacific Biosciences (PacBio) reads (27, 28). We improved the continuity of the gorilla Y assembly by merging two previously published assemblies (9, 21). To remove PARs, we filtered each species-specific Y assembly against the corresponding female reference genomes. The resulting bonobo, orangutan and gorilla assemblies of Y male-specific regions (henceforth called ‘Y assemblies’) were of high quality, as evidenced by their high degree of homology to the human and chimpanzee Y chromosomes (**Fig. S2**) and by the presence of sequences of

most expected (16) homologs of human Y genes (**Fig. S3**). They also were of sufficient continuity (**Table S1**), particularly taking the highly repetitive structure of the Y into account.

Ampliconic and X-degenerate scaffolds

To determine which scaffolds are ampliconic and which are X-degenerate in our bonobo, gorilla, and orangutan Y assemblies (such annotations are already available in the human and chimpanzee Y assemblies (7, 8)), we developed a classifier which combines the copy count in the assemblies with mapping read depth information from whole-genome sequencing of male individuals (**Note S1**). This approach was needed as Y ampliconic regions can often be collapsed in assemblies based on next-generation sequencing data (6). Using this classifier, we identified 12.5 Mb, 9.8 Mb, and 14.5 Mb of X-degenerate scaffolds in bonobo, gorilla, and orangutan, respectively. The length of ampliconic regions was more variable: 10.8 Mb in bonobo, 4.2 Mb in gorilla, and 2.2 M in orangutan. Due to potential collapse of repeats, we might have underestimated the true lengths of ampliconic regions. However, their length estimates are expected to reflect their complexity. For instance, the complexity might be low in the orangutan Y, which is consistent with a high read depth we find in orangutan Y ampliconic scaffolds (**Fig. S4**) and with previous cytogenetic studies demonstrating long repetitive arrays harboring genes in orangutan (29).

Alignments

Using ProgressiveCactus (30), we aligned the sequences of the Y chromosomes from five great ape species—the generated-here assemblies of orangutan and bonobo Ys, the improved-by-us gorilla Y assembly, and the publicly available chimpanzee (8) and human (7) Y assemblies. The resulting multi-species alignment allowed us to identify species-specific sequences, sequences shared by all species, as well as sequences shared by some but not all species (**Fig. S5, Table S2A**). These results were confirmed by pairwise alignments (**Table S2B**). For instance, the bonobo Y assembly had the highest percentage of its sequence aligned to the chimpanzee Y (52.7% and 65.6% from multi-species and pairwise alignments, respectively) than to any other species, in line with the recent divergence of these two species (18, 19). It also harbored a large portion of species-specific sequences (41.8%). As was shown previously (9), the gorilla Y had the highest percentage of its sequence aligning to the human Y (75.7% and 89.6% from multi-species and pairwise alignments, respectively). In terms of sequence identity (**Tables S2C-D**), as expected, the chimpanzee and bonobo Ys were most similar to each other (99.2% and ~98% from multi-species and pairwise alignments, respectively), while the orangutan Y had the lowest identity to any other great ape Y chromosomes (~93-94% and ~92% from multi-species and pairwise alignments, respectively). From multi-species alignments (**Table S2C**), the human Y was most similar in sequence to the chimpanzee or bonobo Ys (97.9% and 97.8%, respectively), less similar to the gorilla Y (97.2%), and the least similar to the orangutan Y (93.6%), in agreement with the accepted phylogeny of these species (10). The pairwise alignments confirmed this trend (**Table S2D**). These results argue against incomplete lineage sorting at the male-specific Y chromosome locus in great apes.

Substitution rates on the Y

We next asked whether the chimpanzee Y chromosome, whose architecture and gene content differ drastically from the human and gorilla Y chromosomes (9), has experienced an elevated

substitution rate. Using our multi-species Y chromosome alignment, we estimated substitution rates along the branches of the great ape phylogenetic tree (**Fig. 1A**; see Methods for details). A similar analysis was performed using an alignment including autosomes (**Fig. 1B**). A higher substitution rate on the Y chromosome than on the autosomes, i.e. male mutation bias (31), was found for each branch of the phylogeny (**Fig. 1**, **Note S2**). Notably, the Y-to-autosomal substitution rate ratio was higher in the *Pan* lineage, including the chimpanzee (1.76) and bonobo (1.76) lineages and the lineage of their common ancestor (1.91), than in the human lineage (1.48). These trends did not change after correcting for ancestral polymorphism (**Note S2**). To analyze these data in more detail, we used a test akin to the relative rate test (32) and addressed whether the *Pan* lineage experienced more substitutions than the human lineage (**Table S3**). Using gorilla as an outgroup, we observed a significantly higher number of substitutions that occurred between chimpanzee and gorilla than between human and gorilla. The ratio of these two numbers was 1.006 (significantly different from 1, $p<1\times10^{-5}$, χ^2-test) for autosomes, but was as high as 1.090 ($p<1\times10^{-5}$, χ^2-test) for the Y. Similarly, we observed a higher number of substitutions that occurred between bonobo and gorilla than between human and gorilla. The ratio of these two numbers was 1.029 ($p<1\times10^{-5}$, χ^2-test) for autosomes, but was as high as 1.114 ($p<1\times10^{-5}$, χ^2-test) for the Y. In both cases, these ratios were significantly higher for the chromosome Y than for the autosomes ($p<1\times10^{-5}$ in both cases, χ^2-test on contingency table). Thus, while the *Pan* lineage experienced an elevated substitution rate at both autosomes and the Y, this elevation was particularly strong on the Y.

Gene content evolution

Utilizing sequence assemblies and testis expression data (17), we evaluated gene content and the rates of gene birth and death on the Y chromosomes of five great ape species. First, we examined the presence/absence of homologs of human Y chromosome genes (16 X-degenerate genes and nine ampliconic gene families, a total of 25 gene families; for multi-copy gene families, we were not studying copy number variation, but only presence/absence of a family in a species; **Fig. S6**). Such data were previously available for the chimpanzee Y, in which seven out of 25 human Y gene families became pseudogenized or deleted (8), and for the gorilla Y, in which only one gene family (*VCY*) out of 25 is absent (9). Here, we compiled the data for bonobo and orangutan. From the 25 gene families present on the human Y, the bonobo Y lacked seven (*HSFY, PRY, TBL1Y, TXLNGY*, *USP9Y*, *VCY*, and *XKRY*) and the orangutan Y lacked five (*TXLNGY*, *CYorf15A*, *PRKY*, *USP9Y*, and *VCY*). Second, we annotated putative new genes in our bonobo and orangutan Y assemblies (**Note S3**). Our results suggest that the bonobo and orangutan Y chromosomes, similar to the chimpanzee (8) and gorilla (9) Ys, do not harbor novel genes. As a result, we obtained the complete information about gene family content on the Y chromosome in five great ape species.

Using this information, we reconstructed gene content at ancestral nodes and asked whether the rates of gene birth and death varied across the great ape phylogeny. For this analysis, we employed the evolutionary model developed by Iwasaki and Takagi (33) and used the macaque Y chromosome as an outgroup (34). Because X-degenerate and ampliconic genes might exhibit different trends, we analyzed them separately (**Fig. 2, Table S4**). Considering gene births, none were observed for X-degenerate genes, and only one (*VCY*, in the human-chimpanzee-bonobo common ancestor) was observed for ampliconic genes, leading to overall low gene birth rates. Considering gene deaths, three ampliconic gene families and three X-generate genes were lost by the chimpanzee-bonobo common ancestor, leading to death rates of 0.095 and 0.049

events/MY, respectively. Bonobo lost an additional ampliconic gene, whereas chimpanzee lost an additional X-degenerate gene, leading to death rates of 0.182 and 0.080 events/MY, respectively. In contrast, no deaths of either ampliconic or X-degenerate genes were observed in human and gorilla. Orangutan did not experience any deaths of X-degenerate genes, but lost four ampliconic genes. Its ampliconic gene death rate (0.021 events/MY) was still lower than that in the bonobo or in the bonobo-chimpanzee common ancestor. To summarize, across great apes, the *Pan* genus exhibited the highest death rates for both X-degenerate and ampliconic genes.

Conservation of human and chimpanzee palindrome sequences

Did the palindromes now present on the human Y (P1-P8) and chimpanzee Y (C1-C19) evolve before or after the great ape lineages split? To answer this question, we identified the proportions of human and chimpanzee palindrome sequences that aligned to bonobo, orangutan and gorilla Ys in our multi-species alignments (**Fig. 3A**, **Table S5**). Among human palindromes, P5 and P6 were the most conserved (covered by 89-97% of other great ape Y assemblies), whereas the majority of P3 sequences were human-specific (covered by only 31-38% of other great ape Y assemblies). Nevertheless, the common ancestor of great apes most likely already had substantial lengths of sequences homologous to P1, P2, and P4-P8, and some sequences of P3 (**Fig. 3B**). Chimpanzee palindromes C17, C18, and C19 are homologous to human palindromes P8, P7, and P6, respectively (8). Therefore below we focused on the other chimpanzee palindromes and, following (8), divided them into five homologous groups: C1 (C1+C6+C8+C10+C14+C16), C2 (C2+C11+C15), C3 (C3+C12), C4 (C4+C13), and C5 (C5+C7+C9) (**Table S6**). The palindromes in the C3, C4, and C5 groups had substantial proportions (usually 70-90%) of their sequences covered by alignments with other great ape Ys (**Fig. 3A**). In contrast, most of C2 sequences (85%) were shared with bonobo, and a substantial proportion of C1 sequences was chimpanzee-specific. Nonetheless, the common ancestor of great apes likely already had large amounts of sequences homologous to group C3, C4, and C5 palindromes, and also some sequences homologous to group C1 and C2 palindromes (**Fig. 3B**).

To determine whether the bonobo, orangutan, and gorilla sequences homologous to human or chimpanzee palindromes were multi-copy (i.e. present in more than one copy), and thus could form palindromes, in the common ancestor of great apes, we obtained their read depths from whole-genome sequencing of their respective males (**Fig. 3A**; see Methods). This approach was used because we expect that some palindromes were collapsed in our Y assemblies and, hence, the copy number within the assembly may be unreliable. Additionally, we used the data on the homology between human and chimpanzee palindromes summarized from the literature (7–9) (**Table S7**). Using maximum parsimony reconstruction, we concluded (**Note S4**) that sequences homologous to P4, P5, P8, C4, and partial sequences homologous to P1, P2, C2, and C3 were multi-copy in the common ancestor of great apes (**Fig. 3B**). Sequences homologous to P3, P6, and C1 were multi-copy in the human-gorilla common ancestor, those homologous to P7 were multi-copy in the human-chimpanzee common ancestor, and those homologous to C5 were multi-copy in the bonobo-chimpanzee common ancestor (**Fig. 3B**; **Note S4**).

Species-specific multi-copy sequences in bonobo, gorilla, and orangutan

In addition to finding sequences homologous to human and/or chimpanzee palindromes, we detected 9.5 Mb, 1.6 Mb, and 3.5 Mb of species-specific sequences in our bonobo, gorilla, and orangutan Y assemblies (**Fig. S5**). By mapping male whole-genome sequencing reads to these sequences (see Methods), we found that 81%, 56%, and 32% of them in bonobo, gorilla, and orangutan had copy number of 2 or above (**Table S8**). Thus, large portions of Y species-specific sequences are multi-copy and might harbor species-specific palindromes.

What drives conservation of gene-free palindromes P6 and P7?

Palindromes are hypothesized to evolve to allow ampliconic genes to withstand high mutation rates on the Y via gene conversion in the absence of interchromosomal recombination (12, 13). Two human palindromes—P6 and P7—do not harbor any genes, however the large proportions of their sequences are present and are multi-copy (**Fig. 3A, Table S7**) in most great ape species we examined (with exceptions of P7 absent from bonobo and of single-copy P6 and P7 in orangutan). In fact, P6 is the most conserved human palindrome (**Fig. 3A**) and was present in the multi-copy state in the human-gorilla common ancestor (**Fig. 3B**). We hypothesized that conservation of P6 and P7 might be explained by their role in regulation of gene expression.

Using ENCODE data (35), we identified candidate open chromatin and protein-binding sites in P6 and P7 (**Fig. S7**). In P6, we found markers for open chromatin (DNase I hypersensitive sites) and histone modifications H3K4me1 and H3K27ac, associated with enhancers (36), in human umbilical vein endothelial cells (HUVEC). In P7, we found markers for CAMP-responsive element binding protein 1 (CREB-1), participating in transcription regulation (37), in a human liver cancer cell line HepG2. Interestingly, we did not identify any open-chromatin or enhancer marks in P6 and P7 in testis, suggesting that the sites we found above regulate genes expressed outside of this tissue.

Frequent chromatin interactions between and within palindromes

Because ampliconic regions on the Y undergo gene conversion and Non-Allelic Homologous Recombination (NAHR) (38), we hypothesized that these processes are facilitated by increased chromatin interactions. To evaluate this possibility, we studied chromatin interactions on the Y chromosome utilizing a statistical approach specifically developed for handling Hi-C data originating from repetitive sequences (39). We used publicly available human and chimpanzee Hi-C data generated for induced pluripotent stem cells (iPSCs) (40). For human, we additionally used data generated for HUVEC (41). We found prominent chromatin contacts formed both between and within palindromes located inside ampliconic regions on the Y chromosome (**Fig. 4**). In fact, the contacts in the human palindromic regions were significantly overrepresented when compared with the expectation based on the proportion of the Y occupied by palindromes ($p<0.001$, permutation test with palindromic/non-palindromic group categories; **Table S9**, **Fig. S8**), suggesting biological importance. Notably, we observed similar patterns for two different human cell types, as well as for both human and chimpanzee iPSCs (**Fig. 4, Fig. S9**).

We also hypothesized that arms from the same palindrome interact with each other via chromatin contacts. Our analysis of human Hi-C data from iPSC cells (40) suggests that

palindrome arms are indeed co-localized (**Fig. 4C**). P1 and P5, the largest palindromes in the human genome (7), exhibited a particularly prominent folding pattern (red vertical lines in **Fig. 4C**). These results suggest that, in addition to the enrichment in the local interactions expected to be present in the Hi-C data (42), homologous regions of the two arms of a palindrome interact with each other with high frequency.

Discussion

The Y chromosome assemblies of bonobo and orangutan we present here, together with the human, chimpanzee, and gorilla assemblies produced previously (7–9), have allowed us to conduct a comparative analysis of great ape Y chromosomes. We found that, even though the sequences of great ape Y chromosomes follow the expected species phylogeny, their evolution is more dynamic than that of the autosomes and chromosome X. Still, interestingly, many core features of the Y chromosome, including multi-copy gene families and ampliconic sequences, were likely present in the common ancestor of great apes.

Substitution rates. Higher substitution rates on the Y chromosome than on the autosomes, which we found across the great ape phylogeny, confirm another study (43) and are consistent with male mutation bias, likely caused by a higher number of germline cell divisions in the male than the female germline (31). Higher autosomal substitution rates we detected in the *Pan* than *Homo* lineage corroborate yet another study (44) and can be explained by shorter generation time in *Pan (45)*. A higher Y-to-autosomal substitution ratio (i.e. stronger male mutation bias) in the *Pan* than *Homo* lineage, as observed by us here, could be due to several reasons. *First*, species with sperm competition produce more sperm and thus undergo a greater number of replication rounds than species without sperm competition. This can increase the chance of replication errors, resulting in more mutations (46), and might lead to higher substitution rates on the Y chromosome than autosomes (31). Consistent with this explanation, both chimpanzee and bonobo have polyandrous mating patterns, experience sperm competition, and exhibit strong male mutation bias, as compared with a largely monoandrous mating pattern and no sperm competition for human and gorilla (47), who exhibit weaker male mutation bias (**Note S2**). On the other hand, a dispersed mating pattern (48) is associated with limited sperm competition and thus should lead to weak male mutation bias in orangutan; however, our results (**Note S2**) contradict this expectation by suggesting a strong bias in orangutan. *Second*, a shorter spermatogenic cycle can also increase the number of replication rounds per time unit and can elevate substitution rates on the Y, leading to stronger male mutation bias. In agreement with this explanation, the spermatogenic cycle is shorter in chimpanzee (14 days) than in human (16 days) (49, 50); the data are limited for other great apes. *Third*, selection could increase Y mutation rates and, consequently, male mutation bias in some species (e.g., in species of the genus *Pan*). Indeed, because the Y carries genes important for male fertility, selective elevation of mutation rates on the Y could generate sperm with more diverse characteristics; such variability might provide a distinct reproductive advantage for an individual and a species. Currently we do not have evidence of selection operating on the Y in *Pan* (**Note S5**). *Fourth*, differences in male vs. female generation times may affect male mutation bias (51); however, our data do not support this explanation. A stronger male mutation bias would be expected in *Pan* than in *Homo* if the ratio of male-to-female generation times were respectively higher (if this were the case, then *Pan* would experience more rounds of replication in the male germline,

leading to stronger male mutation bias). However, the opposite is true: sex-specific generation times are similar for chimpanzee (24-25 years), whereas generation time is higher in males than in females in human (32 vs. 26 years) (52).

Note that the precise estimates of male mutation bias in closely related species are affected by ancient genetic polymorphism (43) and thus, even though we corrected for this effect (**Note S2**), our results should be taken with caution because of incomplete data on the sizes of ancestral great ape populations (53). In addition to phylogenetic analyses, male mutation bias in great apes was investigated from pedigree studies, which should not have been affected by ancient genetic polymorphism. One such study detected significantly higher male mutation bias in chimpanzee than in human (54), in agreement with our results, while another study found no significant differences in the magnitude of male mutation bias among human, chimpanzee, gorilla, and orangutan (55). In general, current pedigree studies in non-human great apes are based on the analysis of only a handful of trios per species (54, 55). Increasing sample sizes in such studies should lead to a higher power to detect potential differences in male mutation bias among species.

Evolution of palindromic and other multi-copy sequences. We found that multi-copy sequences are abundant in great ape Ys and that many of them were present in the common ancestor of great apes. In fact, substantial portions of most human palindromes (five out of eight), and of most chimpanzee palindrome groups (three out of five), were likely multi-copy (and thus potentially palindromic) in the common ancestor of great apes, suggesting conservation over >13 MY of evolution. Moreover, two of the three rhesus macaque palindromes are conserved with human palindromes P4 and P5 (34), indicating conservation over >25 MY. On the other hand, our study also found species-specific amplification or loss of palindromes and other repetitive sequences, indicating that Y ampliconic sequence evolution is highly dynamic. A vivid example of this pattern is an extensive amplification of sequences homologous to P1, P2, P4, P5, and C3-C4 in orangutan. We also found that sizeable proportions of species-specific sequences in the bonobo, gorilla and orangutan Ys are multi-copy. Future studies will establish which of them form palindromes as opposed to tandem repeats. Regardless, repetitive sequences constitute a biologically significant component of great ape Y chromosomes and their multi-copy state might be selected for.

Ampliconic sequences are thought to have evolved multiple times in diverse species (reviewed in (56)) to enable intrachromosomal (Y-Y) gene conversion. Y-Y gene conversion can compensate for the degeneration in the absence of interchromosomal recombination on the Y by removing deleterious mutations (57, 58), can decrease the loss of less mutated alleles by drift, can lead to concerted evolution of repetitive sequences (13), and can increase the fixation rate of beneficial mutations (59). Yet, despite its critical importance for the Y, how Y-Y gene conversion occurs mechanistically is not well understood. Our analysis of Hi-C data demonstrated that ampliconic sequences and palindrome arms co-localize on the Y in both human and chimpanzee, potentially facilitating Y-Y gene conversion. Our study is the first one to apply novel, repeat-aware software (39) to Hi-C data from the Y, which allowed us to comprehensively examine intra-Y chromatin interactions; most previous investigations excluded the Y from such analyses because of its repetitive nature. The high frequency of chromatin interactions in amplicones on the Y may facilitate not only Y-Y gene conversion, but also NAHR. This process is frequently used to explain rapid evolution of the ampliconic gene families' copy number (60), as well as frequent structural rearrangements (61, 62), some of them leading to

spermatogenic failure, sex reversal, and Turner syndrome (63). Therefore, our findings contribute to our understanding of the mechanisms behind the dynamic evolutionary processes on the Y and have potential applications to human health.

Previous studies (e.g., reviewed in (12, 13, 56)) focused on the role of Y-Y gene conversion in preserving Y ampliconic gene families, which are critical for spermatogenesis and fertility (7), and suggested that this phenomenon constitutes the major adaptive role of palindromic sequences. Our findings suggest that, instead of genes, some human palindromes (P6 and P7) possess regulatory regions of expression of genes transcribed outside of testis and thus likely located on non-Y chromosomes, and such regulatory regions might drive conservation of gene-free palindromes. This observation should be examined in more detail in the future, but can potentially shift a paradigm in our understanding of Y chromosome functions. Indeed, our results imply that, in addition to carrying genes important for male sex determination and spermatogenesis, the Y chromosome participates in regulating gene expression in the genome. This echoes findings in *Drosophila* (e.g., (64)) and in the mouse Y chromosome, which contains a small, gene-free region that interacts with the rest of the genome (65).

Gene content evolution. We found that the gene content in the common ancestor of great apes likely was the same as is currently found in gorilla, and included eight ampliconic and 16 X-degenerate genes (**Fig. 2**). Analyzing the data on ampliconic gene content (**Fig. 2**), palindrome sequence (**Fig. 3B**), and ampliconic gene copy number (Fig. 2 in (17)) evolution jointly, we can infer which ampliconic genes were present in the multi-copy state in the common ancestor of great apes. Our results suggest that the common ancestor of great apes had sequences homologous to P1, P2, P4, P5 and P8 in multi-copy state (**Fig. 3B**), which in human carry *DAZ*, *BPY2*, *CDY*, *HSFY*, *XKRY*, and *VCY* (6). Except for *VCY*, which was acquired by the human-chimpanzee common ancestor, the remaining five genes were likely present as multi-copy gene families in the common ancestor of great apes, because three of them (*DAZ*, *BPY2*, and *CDY*) are present as multi-copy in all great ape species (17), and the other two (*HSFY* and *XKRY*)—in all great ape species but chimpanzee and bonobo (17), in which they were lost (**Fig. 2**).

Our comprehensive analysis of great ape Y gene content has allowed us to investigate rates of gene birth and death on the Y. We discovered that there is only one gene family that was born throughout the whole great ape phylogeny—*VCY* was acquired by the common ancestor of human and chimpanzee. As a result, except for this branch, we found uniformly low rates of gene birth. A low rate of *ampliconic* gene birth contradicts predictions of high birth rate made in previous studies for such genes (56), but suggests that great ape radiation does not provide sufficient time for gene acquisition by ampliconic regions. For new genes to survive on the Y chromosome, they should be beneficial to males. Ampliconic regions on the Y chromosomes of several other mammals indeed acquired such genes (66, 67).

We expected to observe a high death rate for X-degenerate genes, but a low death rate for ampliconic genes, because the former genes do not undergo Y-Y gene conversion and thus should accumulate deleterious mutations, whereas the latter genes are multi-copy (all copies need to be deleted or pseudogenized for the gene family to die) and can be rescued by Y-Y gene conversion. Unexpectedly, the rates of gene death did not differ between ampliconic and X-degenerate genes. Indeed, ~44.4% of ampliconic gene families were either deleted or pseudogenized, as compared with also ~43.8% of X-degenerate genes, across the great ape Y phylogenetic tree. These values were 67% and 50%, respectively, if we included the macaque

Y. While our data did not support our hypothesis, other findings suggest that death of ampliconic genes is a gradual process. Indeed, we have recently shown that ampliconic gene families dead in some great ape species have reduced copy number in other species (17, 60), lowering the chances for Y-Y gene conversion. These observations suggest that such genes are on the way to become non-essential and are at death's door in great apes.

The rates of gene death varied among great ape species. In particular, we observed high rates of death in the *Pan* lineage—in the common ancestor of bonobo and chimpanzee, but also in the bonobo and chimpanzee lineages. Thus, the evolutionary forces driving such a high rate of gene death have likely been operating in the *Pan* lineage continuously since its divergence from the human lineage. What evolutionary forces could explain this observation? *First*, gene-disrupting or gene deletion mutations could be hitchhiking in haplotypes with positively selected mutations. Positive selection might be acting in the *Pan* lineage due to its polyandrous mating pattern and sperm competition. No gene deaths in the human and gorilla lineages, experiencing no sperm competition, and low gene death rates in orangutan, experiencing limited sperm competition, are consistent with this explanation. Our tests for positive selection acting at protein-coding genes did not produce significant results, however they indicated significantly elevated nonsynonymous-to-synonymous rate ratios for four X-degenerate genes in bonobo, chimpanzee, and/or their common ancestor (**Note S5**). Note that our power to detect positive selection from phylogenetic data collected for closely related species might be low. *Second*, the *Pan* Y could have undergone stronger drift leading to fixation of variants lacking genes and gene families, which were already in the process of becoming non-essential. Predominant dispersal by females, and the associated low male effective population size, observed in bonobo and chimpanzee (68), as compared with predominant dispersal by males in gorilla and orangutan (3, 69), are consistent with this explanation. The intraspecific sequence variation (i.e. diversity) data on the Y chromosome in great apes are needed to differentiate between these two explanations.

Future directions. While our comparative study of great ape Y chromosome has enabled us to address several important topics in their evolution, many unanswered questions remain. Future studies should include sequencing of the Y chromosome for multiple individuals per species, and such data are expected to provide a resolution between the evolutionary scenarios driving high substitution and gene death rates in the *Pan* lineage. The analysis of morphological and genetic (e.g., in ampliconic gene copy number) variability in sperm in different great ape species can provide complementary information.

In this study we uncovered many principal features of ampliconic sequence and gene evolution, opening opportunities for new inquiries. Future investigations should focus on deciphering the sequences of different copies and isoforms of ampliconic genes (70), which should allow one to examine natural selection operating at them in detail. Sequence amplification and increase in gene copy number (17) in orangutan should be examined further as well. Because chromatin organization depends on the tissue of origin (41, 71), the high prevalence of intra-ampliconic and intra-palindromic contacts we found in two somatic tissues should be confirmed in sperm and testis. Additionally, comparing chromatin organization and evolution of palindromes on the Y vs. X chromosomes should aid in understanding the unique role repetitive regions might play on the Y.

From a more applied perspective, the bonobo and orangutan Y assemblies presented here are useful for developing genetic markers to track male-specific dispersal in these endangered

species. This is of utmost importance because both species experience population decreases due to habitat loss. Therefore our results are expected to be of great utility to conservation genetics efforts aimed at restoring these populations.

Materials and Methods

Short- and long-read data to generate Y chromosome assemblies

We used human and chimpanzee Y sequences from version hg38 and panTro6 assemblies, respectively (7, 8). In both cases, PARs were masked (we used human coordinates 10,001-2,781,479 and 56,887,903-57,217,415 for PAR1 and PAR2, respectively, as provided by Ensembl (https://useast.ensembl.org/info/genome/genebuild/human_PARS.html) and 26,208,347-26,341,346 for PAR in chimpanzee, identified in house using LASTZ (72) alignments). For the gorilla Y chromosome, there are two published assemblies: the one produced from short- and long-read data by our group (9), and the one recently generated using PacBio reads pre-processed with Pacasus to remove within-read duplications (21). We merged the two assemblies to obtain a more continuous and complete version; the merging was performed with metassembler v.1.5 (25). Since metassembler additionally uses information from mate-pair reads, we used the previously-generated GY19 library (9) and set insert sizes to range from 298 to 6,444 bp in the configuration file.

For both orangutan and bonobo (**Fig. S1**), we first generated short, paired-end reads for a male individual. Sumatran orangutan male (ID 1991-51) genomic DNA provided by the Smithsonian Institution was extracted from testis using DNeasy Blood and Tissue Kit (Qiagen). Bonobo male (ID PR00251) genomic DNA was extracted using the same kit from the fibroblast cell line provided by the San Diego Zoological Society. Illumina paired-end PCR-free libraries with insert size of ~1,000 bp were constructed from both DNA samples and sequenced on HiSeq2500 using 251-bp paired-end sequencing protocol (approximately 600 million reads were generated for each library). The reads were assembled using DISCOVAR (22). To identify putative Y-chromosomal contigs, we processed each of these two male assemblies with DiscoverY (23) using the male-to-female abundance threshold of 0.4 and using "-abundance-min 4". The resulting contigs were scaffolded with filtered mate-pair 250-bp reads, generated for the same individuals using Nextera Mate Pair Sample Preparation Kit (with median insert size of 8 kb), using BESST v.2.2.8 (24). Filtering of mate pairs was performed as follows. All mate-pair reads were mapped to the respective (bonobo or orangutan) whole-genome male DISCOVAR assembly using bwa mem v.0.7.17-r1188 (73). From these alignments, reads putatively originating from Y chromosome were retained and used for scaffolding. The next step, aimed at further improving the continuity of Y assemblies, differed between orangutan and bonobo. The orangutan Y assembly was merged (using metassembler v.1.5 (25)) with an assembly generated for orangutan male AG06213 with 10×Genomics (**Table S10**). To aid metassembler, one can provide mate-pair reads; we used mate-pair reads from individual 1991-51 and set the insert size range from 5,088 to 12,683 bp in the configuration file. High-molecular-weight genomic DNA from Sumatran orangutan male (ID AG06213) was extracted from fibroblast cell line (provided by L. Carrel, PSU) using MagAttract HMW DNA kit (Qiagen) and used to construct a 10x Genomics library. The resulting reads were assembled with Supernova assembler v.1.0.2. The bonobo Y assembly was augmented with Pacasus-corrected (default

parameters) (21) flow-sorted Y-enriched PacBio reads from a bonobo male Ppa_MFS using SSPACE-LongRead (v.1-1) (27), and additional scaffolding and assembly were performed with PBJelly (PBSuite_14.9.9, parameters <blasr>-minMatch 8 -minPctIdentity 85 -bestn 1 -nCandidates 20 -maxScore -500 -nproc 6 -noSplitSubreads</blasr>) (28).

The subsequent processing steps were the same for the orangutan Y and bonobo Y assemblies. For the orangutan, we first called variants from the first male (ID 1991-0051) relative to the assembly we obtained after merging with 10×Genomics data. The variants were called with FreeBayes (v.1.3.1) (74) using settings "--ploidy 1 --min-alternate-fraction 0.6", and filtered for quality (biopet-vcffilter --minQualScore 30). We then used bcftools to 'polish' the assembly by consensus calling (bcftools consensus -H 1, **Table S11**). Using the same approach, we polished the bonobo Y assembly using the variants from bonobo male ID PR00251.

To remove the pseudoautosomal region (PAR) and potential residual autosomal contamination, all scaffolds in gorilla, bonobo and orangutan Y assemblies were filtered against corresponding gorilla (gorGor4), bonobo (panPan2), and orangutan (ponAbe3) female reference genomes. We mapped the assembly scaffolds to the female reference using blasr v.5.3.3 (75) with parameters: "blasr --bestn 1 --nproc 32 --hitPolicy randombest --header --minPctSimilarity 95". Scaffolds that had at least 95% sequence identity to the female genome were excluded from the assemblies.

Multi-species alignments

To align the Y chromosomes of five great ape species, several alignment algorithms were evaluated: SibeliaZ (76), Multiz-TBA (77), Multiz-roast (77), and ProgressiveCactus (30). The alignments were evaluated both by comparing the proportion of aligned base pairs and visually, i.e. by comparing the assembled coverage in non-human great ape species sequences aligned to the human X-degenerate genes. ProgressiveCactus showed the most promise and was chosen for all further analyses. We soft-masked repeats prior to the alignment of assemblies using RepeatMasker (78) with the following settings: *RepeatMasker -pa 63 -xsmall -species Primates ${assembly}.rmsk.fa* (78). To generate multiple sequence alignment with ProgressiveCactus, we used default parameters and did not provide a guide tree. The alignment was converted from hal (79) to MAF format using hal2maf function (parameterized with, for example, --noAncestors --refGenome hg_Y --maxRefGap 100 --maxBlockLen 10000). The role of the MAF reference sequence was given to each species in turn, to create a separate MAF file for each, as well as for the ProgressiveCactus-inferred ancestral sequence.

We split the MAF file by species subset, producing a separate MAF file containing all blocks that only involve the particular species in that subset. The MAF file of blocks containing all five species was converted to multi-fasta by MAF to FASTA tool, run at https://usegalaxy.org/ (Galaxy Version 1.0.1), using an option "One Sequence per Species". Sequence identity was calculated from the multi-fasta file using the custom script multi_fasta_to_pairwise_identity.py.

Independently, sequence identity distributions were derived from pairwise alignments as follows. For each pair of species (s1, s2) LASTZ (72) was used to align s1's Y assembly and s2's. Masking was disabled, allowing alignment of duplicated elements. Substitution scores were

identical to those used by us in Tomaszkiewicz et al. (9). The exact LASTZ command line was: “lastz s1.chrY[unmask] s2.chrY[unmask] W=12 Q=human_primate.scores.” Alignments were then post-processed as follows. For each base in s1, the highest identity alignment covering that base was chosen, and base counts were collected for identities at 0.1% intervals. Average sequence identity was computed from these binned counts, excluding any unaligned bases.

Percentage of sequence aligned in multiple alignment was computed as follows. For the portion of species s1 aligning to s2, all MAF blocks containing s1 and s2 (as well as, possibly, other species) were extracted. These were reduced to the aligning intervals in s1 by computing the mathematical union of all s1 intervals in any of the extracted blocks. The combined length of these intervals is the number of bases in s1 that have an alignment to s2. This length was divided by the number of non-N bases in s1.

Percentage of sequence aligned in pairwise alignments was computed as follows. For the portion of species s1 aligning to s2, the s1-to-s2 LASTZ alignment was reduced to the aligning intervals in s1 by computing the mathematical union of all s1-intervals in any alignment block. The combined length of these intervals is the number of aligning bases in s1. This length was divided by the number of non-N bases in s1.

Substitution rate analysis

To estimate the substitution rates on the Y chromosome, we used the ancestor-based five-species alignment described in the previous section. To provide conservative estimates of substitution rates, we removed duplicates from the alignment using mafDuplicateFilter from the mafTools suite (version 0.1), (80), similar to (32). This step replaces multiple sequences from the same species by the sequence closest to the consensus of an alignment block. We also converted all alignments blocks to the positive strand of the ancestral sequence (maf_flip_for_ref.py). Again, to obtain conservative estimates, we only retained alignment blocks where all five species were present, thus largely restricting our analysis to X-degenerate regions. This step was performed using script *parse_cactus* (available on GitHub) and the python libraries *bx.align.maf* and *bx.align.tools*. The final filtered alignment was then used to pick the best-fitted substitution model using jModelTest (81, 82). Taking into account AIC and BIC of the various models tested with jModelTest and the availability of models implemented in *phylofit* (83), we decided to proceed with GTR (also called REV) model (84) with variable substitution rates (--nrates=4). Using this model and our filtered alignment, we ran *phylofit* (83) with the following settings: phyloFit -E -Z -D 42 --subst-mod REV --nrates 4 --tree "(((hg_Y,panTro_Y),gorGor_Y),ponAbe_Y)". The estimated branch lengths were used for all the downstream analysis. The output was then visualized using R script and library *ape* (85)(86).

Gene content analysis

Analysis of genes homologous to human Y genes. To retrieve the bonobo and orangutan Y chromosome X-degenerate and ampliconic genes, we aligned the scaffolds from bonobo and Sumatran orangutan Y chromosome assemblies to species-specific or closest-species-specific reference coding sequences of X-degenerate and ampliconic genes. Next, we visualized the

alignment results in Integrative Genomics Viewer (IGV) (v.2.3.72), and protein-coding consensus sequences were retrieved and checked for ORFs. However information about *RPS4Y2* and *MXRA5Y* remained missing. We used gene predictions (see next paragraph) and testis-specific transcriptome assemblies (17) from the same publication to study the presence of the *RPS4Y2* and *MXRA5Y* genes (**Fig. S6**).

Predicting novel genes in bonobo and orangutan assemblies. We used AUGUSTUS (87) (--species=human --softmasking=on --codingseq=on) to predict genes in the Y chromosome assemblies of bonobo and orangutan. From the list of predicted genes we retained those that have a start and stop codon. Using blastp from BLAST (2.9.0; -db uniprot_sprot.pep -max_target_seqs 1 -evalue 1e-5 -num_threads 10) (88) we annotated the predicted gene sequences. Based on blast annotation, we classified all the genes which are not annotated by blastp as candidate *de novo* genes. From these genes, we retained those that had >90% sequence identity to the sequence in the UniProt (89) protein database and were covered by least 90% of the gene sequence in the blast output. To make sure that the predicted genes are on the Y chromosome and do not represent misassemblies, we performed an extra filtering step where we used only those genes which are present on contigs which align to either human or chimpanzee Y chromosomes. The resulting gene annotations which are not found on the human Y chromosome were assigned as candidate novel genes translocated to the Y. The longest predicted transcript sequences for genes annotated as Y-chromosome-specific were obtained without the above filters for validation of missing gene content in bonobo and orangutan. The gene predictions which were annotated as X homologs (*TXLNG*, *PRKX*, *NLGNX*, etc.) of Y chromosome genes were also classified as Y genes.

Reconstruction of gene content evolution in great apes. Once we obtained the complete Y chromosome gene content in great apes, we converted the gene content into binary values which represent the presence or absence of a gene in a species. The complete deletion or pseudogenization of gene/gene families is represented by ‘zero’ and the presence of an uninterrupted gene is represented by ‘one’. Using the model developed by Iwasaki and Takagi (33), we reconstructed the evolutionary history of (separately ampliconic and X-degenerate) genes across great apes. The phylogenetic tree of great apes (90), along with the table representing the presence or absence of genes (**Fig. S6**), was used as an input to generate the rate of gene birth and gene death for each branch of the tree, and the reconstructed gene content at each internal node of the tree. To obtain the rate in the units of events per million years, the rate of gene birth, as well as the rate of gene death, was divided by the length of the branches (in millions years) (33).

Palindrome analysis

Human and chimpanzee palindrome arm sequence conservation. The ProgressiveCactus output file was converted to MAF (see the preceding paragraph) twice. The first time we set --refGenome to ‘human’, which results in a human-centric MAF file (where the first line represents sequences from the human Y) and the second time we set --refGenome parameter to ‘chimpanzee’ to obtain a chimpanzee-based MAF file. The coordinates of the human and chimpanzee palindromes (PanTro4) were obtained from a previous publication (9). Chimpanzee palindrome coordinates were converted to the PanTro6 version using the liftOver utility from the

UCSC Genome Browser (91). For few chimpanzee palindromes for which liftOver failed to convert the coordinates, we generated a dotplot of chimpanzee Y to itself using Gepard (92) and identified the locations of palindromes from the dotplot. To overcome redundancy of sequence between the arms of palindrome, we further obtained the coordinates of the first arm of the palindrome for human (**Table S12A**) and chimpanzee (**Table S12B**), from which the palindrome coverage was calculated.

With the use of AlignIO from Biopython (93), we parsed the MAF files to obtain all the alignment blocks which overlapped given palindrome arm coordinates for each species and summed the number of sites covered by these parsed alignment blocks. For each alignment block within the palindrome arm, we considered only those sequences which had less than 5% gaps and counted the number of aligned nucleotides per species at sites where the palindrome arm is unmasked (softmasked with Repeat Masker (78)). Next we calculated the number of nonrepetitive characters in the whole palindrome arm and used it to calculate the percentage of non-repetitive alignment within the palindrome arm per species.

Palindrome sequence read depth in bonobo, gorilla, and orangutan. We used a pre-established pipeline used to identify human and chimpanzee palindrome sequences in Y assemblies (9). The bonobo, gorilla, and orangutan Y contigs were broken into non-overlapping 1-kb windows and each window was aligned to human (hg38) and (separately) chimpanzee (panTro6) Y chromosome sequences using LASTZ (version 1.04.00) (--scores=human_primate.q, --seed=match12, --markend) (72) separately. The substitution scores were set identical to those used for published LASTZ alignments of primates (94). The best alignment for each window was retrieved from the LASTZ output. All the windows with alignments of >800 bp were considered as homologs to human or chimpanzee Y palindromes and used in the downstream analysis.

The whole-genome male sequencing reads from orangutan, gorilla, and bonobo (**Table S10**) were mapped to the human and (separately) chimpanzee Ys using bwa mem (version 0.7.17-r1188)(73) and the resulting output files were sorted and indexed using samtools (version 1.9) (95). Using computeGCBias and correctGCBias functions from deepTools (version 3.3.0) (96), we performed GC correction of the sequencing read depths. Finally, using bedtools (v2.27.1)(97) 'coverage' function, we obtained the read depth of the windows homologous to human and chimpanzee palindrome sequences (of whole palindromes) identified as described in the previous paragraph. We compared the read depths of palindromic sequence windows to that of windows overlapping X-degenerate genes.

Finding the copy number for ampliconic regions that are species-specific. From the multiple sequence alignment of the Y chromosomes of great apes (species-centric), we extracted the alignment blocks which were unique to a particular species and not aligned to the other Y chromosomes. We filtered the species-specific blocks which were >100 bp in size and obtained their GC-corrected read depth from male whole-genome sequencing data. To calculate the copy number of the resulting blocks we divided their read depth by the median read depth of windows which overlap X-degenerate genes from the palindrome copy number analysis (read depth of single-copy regions).

Search for regulatory factor binding sites in human palindromes P6 and P7. We checked for the presence of sites specific to DNA binding proteins on human palindromes P6 and P7,

which could imply the presence of functionality related to gene regulation. Initially we used the ENCODE track (http://genome.ucsc.edu/ENCODE/) at the UCSC Genome Browser (91) to search for epigenetic modifications in the palindrome regions (98). In particular, we used the track from the Bernstein Lab at the Broad Institute containing H3K27Ac and H3K4Me1 data on seven cell lines from ENCODE, from which we identified human umbilical vein endothelial cells (HUVEC) to have signals in palindrome P6. Later we extended our search to ENCODE data portal (https://www.encodeproject.org/) from where we downloaded the BAM files to look for the presence of peaks (**Table S13**)(35). We used the "Search by region" page under Data in the ENCODE data portal (https://www.encodeproject.org/region-search/) and GRCh38 setting to search for files which are related to the coordinates of palindromes P6 (chrY:16159590-16425757) and P7 (chrY:15874906-15904894). The resulting files were visualized in the UCSC Genome Browser to identify signals in the peaks, and from this information we identified signals in human liver cancer cell line HepG2 for P7. The current ENCODE data processing pipeline by default filters out reads with low mapping quality (i.e. mapping in multiple places of the genome), as a result we did not find any peaks/signals in the majority of the datasets (https://www.encodeproject.org/pipelines/). To overcome this limitation, we manually downloaded the unfiltered BAM files (HUVEC: H3K27Ac, H3K4Me1 and DNase-seq; Testis:H3K27Ac, H3K4Me1 and Dnase-seq; and HepG2: CREB1; Table S13), which include low-mapping-quality reads. We performed peak calling using macs2 (version2.1.4; -f BAM --broad --broad-cutoff 0.05) (99) and the control samples were used whenever available. We used integrative genomics viewer (IGV) (version 2.4.19)(100) to visualize the data.

Hi-C analysis. We used mHi-C (39) to process the Hi-C data generated by (40). Both human and chimpanzee samples originated from pluripotent stem cells (iPSCs), processed with protocol using *Mbo I* restriction enzyme (with the GATCGATC recognition site) and sequenced on Illumina Hi-Seq 4000. The sequencing reads were downloaded from SRA (accessions SRR1658709 and SRR8187262 for human and chimpanzee, respectively) and subsampled to the first 100 million reads. The original mHi-C scripts were modified to match the reference genome and other settings; the resolution was set to 10,000, cutsite was set to 0, and normalization method was set to KR (101) with uniPrior (built based on unique bin pairs within valid interaction distance, see https://github.com/keleslab/mHiC). Additionally, we analyzed a biological sample extracted from umbilical vein endothelial cells of a newborn male, cultured in cell line HUVEC (CC-2517), further subjected to the *in situ* Hi-C protocol with *Mbo I* restriction enzyme and sequenced on Illumina HiSeq 2000 (41). For this analysis, we used the first 50 million reads of SRR1658570 dataset.

In order to calculate the enrichment of ampliconic interactions, we defined these regions separately in human and chimpanzee. In human, we used the corresponding coordinates of eight palindromes (9). Because chimpanzee Y is composed of a higher number of intertwined amplicons and 19 palindromes, the continous ampliconic part spanning the *q* arm of this chromosome, as well as a large part of *p* arm adjacent to the centromere, was analyzed as a single unit.

Data and code availability

Sequencing data, assemblies and alignments are available under Bioproject PRJNA602326.
Code is available at https://github.com/makovalab-psu/great-ape-Y-evolution.

Acknowledgements

We are grateful to Laura Carrel (PSU), Oliver Ryder (San Diego Zoological Society), Malcolm Ferguson-Smith (University of Cambridge), and the Smithsonian Institution for their assistance; to Sven Warris for running Pacasus, to Kristoffer Sahlin for his advice on the parameter choice for BESST, and to Francesca Chiaromonte and Ana Kenney for advice on the statistical analysis of Hi-C data. This research was supported by the National Institute Of General Medical Sciences of the National Institutes of Health (NIH) under Award Number R01GM130691 (to KDM). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Additionally, this study was supported by the funds made available through the Clinical and Translational Sciences Institute, Institute for CyberScience, and Eberly College of Science—at Penn State. Additional support was provided under grants from the Pennsylvania Department of Health using Tobacco Settlement and CURE Funds. The department specifically disclaims any responsibility for any analyses, responsibility or conclusions. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. This study was also supported by PSU-NIH funded CBIOS Predoctoral Training Program (RV is a trainee), and the National Science Foundation (NSF) awards DBI-1356529, IIS-1453527, and CCF-1439057 (to PM).

References

1. P. Berta, *et al.*, Genetic evidence equating SRY and the testis-determining factor. *Nature* **348**, 448–450 (1990).

2. V. S. Vineeth, S. S. Malini, A Journey on Y Chromosomal Genes and Male Infertility. *Int. J. Hum. Genet.* **11**, 203–215 (2011).

3. M. I. Douadi, *et al.*, Sex-biased dispersal in western lowland gorillas (Gorilla gorilla gorilla). *Mol. Ecol.* **16**, 2247–2259 (2007).

4. C. Roos, *et al.*, Nuclear versus mitochondrial DNA: evidence for hybridization in colobine monkeys. *BMC Evol. Biol.* **11**, 77 (2011).

5. A. J. Tosi, K. M. Detwiler, T. R. Disotell, Y-chromosomal Markers Suitable for Noninvasive Studies of Guenon Hybridization. *Int. J. Primatol.* **26**, 685–696 (2005).

6. M. Tomaszkiewicz, P. Medvedev, K. D. Makova, Y and W Chromosome Assemblies: Approaches and Discoveries. *Trends Genet.* **33**, 266–282 (2017).

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

8. J. F. Hughes, *et al.*, Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* **463**, 536–539 (2010).

9. M. Tomaszkiewicz, *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).

10. G. V. Glazko, M. Nei, Estimation of divergence times for major lineages of primate species. *Mol. Biol. Evol.* **20**, 424–434 (2003).

11. International Chicken Genome Sequencing Consortium, Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695–716 (2004).

12. S. Rozen, *et al.*, Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* **423**, 873–876 (2003).

13. E. Betrán, J. P. Demuth, A. Williford, Why Chromosome Palindromes? *International Journal of Evolutionary Biology* **2012**, 1–14 (2012).

14. M. Cechova, *et al.*, High satellite repeat turnover in great apes studied with short- and long-read technologies. *Mol. Biol. Evol.* (2019) https:/doi.org/10.1093/molbev/msz156.

15. D. W. Bellott, *et al.*, Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* **508**, 494–499 (2014).

16. P. Hallast, M. A. Jobling, The Y chromosomes of the great apes. *Hum. Genet.* **136**, 511–528 (2017).

17. Rahulsimham Vegesna, Marta Tomaszkiewicz, Oliver A. Ryder, Rebeca Campos-Sánchez, Paul Medvedev, Michael DeGiorgio, and Kateryna D. Makova, Ampliconic genes on the great ape Y chromosomes: Rapid evolution of copy number but conservation of expression levels. *In Review* (2020).

18. J. Hey, The divergence of chimpanzee species and subspecies as revealed in multipopulation isolation-with-migration analyses. *Mol. Biol. Evol.* **27**, 921–933 (2010).

19. N. Yu, *et al.*, Low nucleotide diversity in chimpanzees and bonobos. *Genetics* **164**, 1511–1518 (2003).

20. D. Bachtrog, Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nat. Rev. Genet.* **14**, 113–124 (2013).

21. S. Warris, *et al.*, Correcting palindromes in long reads after whole-genome amplification. *BMC Genomics* **19**, 798 (2018).

22. N. I. Weisenfeld, *et al.*, Comprehensive variation discovery in single human genomes. *Nat. Genet.* **46**, 1350–1355 (2014).

23. S. Rangavittal, *et al.*, DiscoverY: a classifier for identifying Y chromosome sequences in male assemblies. *BMC Genomics* **20**, 641 (2019).

24. K. Sahlin, F. Vezzi, B. Nystedt, J. Lundeberg, L. Arvestad, BESST--efficient scaffolding of large fragmented assemblies. *BMC Bioinformatics* **15**, 281 (2014).

25. A. H. Wences, M. C. Schatz, Metassembler: merging and optimizing de novo genome assemblies. *Genome Biol.* **16**, 207 (2015).

26. N. I. Weisenfeld, V. Kumar, P. Shah, D. M. Church, D. B. Jaffe, Direct determination of diploid genome sequences. *Genome Res.* **27**, 757–767 (2017).

27. M. Boetzer, W. Pirovano, SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics* **15**, 211 (2014).

28. A. C. English, *et al.*, Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One* **7**, e47768 (2012).

29. B. Gläser, *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).

30. B. Paten, *et al.*, Cactus: Algorithms for genome multiple sequence alignment. *Genome Res.* **21**, 1512–1528 (2011).

31. M. A. Wilson Sayres, K. D. Makova, Genome analyses substantiate male mutation bias in many species. *Bioessays* **33**, 938–945 (2011).

32. P. Moorjani, C. E. G. Amorim, P. F. Arndt, M. Przeworski, Variation in the molecular clock of

primates. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 10607–10612 (2016).

33. W. Iwasaki, T. Takagi, Reconstruction of highly heterogeneous gene-content evolution across the three domains of life. *Bioinformatics* **23**, i230–9 (2007).

34. J. F. Hughes, *et al.*, Strict evolutionary conservation followed rapid gene loss on human and rhesus y chromosomes. *Nature* **483**, 82–87 (2012).

35. C. A. Davis, *et al.*, The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).

36. L. A. Pennacchio, W. Bickmore, A. Dean, M. A. Nobrega, G. Bejerano, Enhancers: five essential questions. *Nat. Rev. Genet.* **14**, 288–295 (2013).

37. B. Mayr, M. Montminy, Transcriptional regulation by the phosphorylation-dependent factor CREB. *Nat. Rev. Mol. Cell Biol.* **2**, 599–609 (2001).

38. M. A. Jobling, Copy number variation on the human Y chromosome. *Cytogenet. Genome Res.* **123**, 253–262 (2008).

39. Y. Zheng, F. Ay, S. Keles, Generative modeling of multi-mapping reads with mHi-C advances analysis of Hi-C studies. *eLife* **8** (2019).

40. I. E. Eres, K. Luo, C. J. Hsiao, L. E. Blake, Y. Gilad, Reorganization of 3D genome structure may contribute to gene regulatory evolution in primates. *PLoS Genet.* **15**, e1008278 (2019).

41. S. S. P. Rao, *et al.*, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).

42. B. R. Lajoie, J. Dekker, N. Kaplan, The Hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods* **72**, 65–75 (2015).

43. K. D. Makova, W.-H. Li, Strong male-driven evolution of DNA sequences in humans and apes. *Nature* **416**, 624–626 (2002).

44. N. Elango, J. W. Thomas, NISC Comparative Sequencing Program, S. V. Yi, Variable molecular clocks in hominoids. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 1370–1375 (2006).

45. K. E. Jones, *et al.*, PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals: Ecological Archives E090-184. *Ecology* **90**, 2648–2648 (2009).

46. J. P. Blumenstiel, Sperm competition can drive a male-biased mutation rate. *J. Theor. Biol.* **249**, 624–632 (2007).

47. M. J. Anderson, A. F. Dixson, Sperm competition: motility and the midpiece in primates. *Nature* **416**, 496 (2002).

48. A. H. Harcourt, P. H. Harvey, S. G. Larson, R. V. Short, Testis weight, body weight and breeding system in primates. *Nature* **293**, 55–57 (1981).

49. E. B. Smithwick, L. G. Young, K. G. Gould, Duration of spermatogenesis and relative frequency of each stage in the seminiferous epithelial cycle of the chimpanzee. *Tissue Cell* **28**, 357–366 (1996).

50. C. G. Heller, Y. Clermont, Spermatogenesis in man: an estimate of its duration. *Science* **140**, 184–186 (1963).

51. G. Amster, G. Sella, Life history effects on the molecular clock of autosomes and sex chromosomes. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 1588–1593 (2016).

52. K. E. Langergraber, *et al.*, Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 15716–15721 (2012).

53. T. Marques-Bonet, O. A. Ryder, E. E. Eichler, Sequencing primate genomes: what have we learned? *Annu. Rev. Genomics Hum. Genet.* **10**, 355–386 (2009).

54. O. Venn, *et al.*, Strong male bias drives germline mutation in chimpanzees. *Science* **344**, 1272–1275 (2014).

55. S. Besenbacher, C. Hvilsom, T. Marques-Bonet, T. Mailund, M. H. Schierup, Direct estimation of mutations in great apes reconciles phylogenetic dating. *Nat Ecol Evol* **3**, 286–292 (2019).

56. B. Trombetta, F. Cruciani, Y chromosome palindromes and gene conversion. *Hum. Genet.* **136**, 605–619 (2017).

57. G. A. B. Marais, P. R. A. Campos, I. Gordo, Can intra-Y gene conversion oppose the degeneration of the human Y chromosome? A simulation study. *Genome Biol. Evol.* **2**, 347–357 (2010).

58. T. Connallon, A. G. Clark, Gene duplication, gene conversion and the evolution of the Y chromosome. *Genetics* **186**, 277–286 (2010).

59. S. Mano, H. Innan, The evolutionary rate of duplicated genes under concerted evolution. *Genetics* **180**, 493–505 (2008).

60. R. Vegesna, M. Tomaszkiewicz, P. Medvedev, K. D. Makova, Dosage regulation, and variation in gene expression and copy number of human Y chromosome ampliconic genes. *PLoS Genet.* **15**, e1008369 (2019).

61. L. Skov, M. H. Schierup, Analysis of 62 hybrid assembled human Y chromosomes exposes rapid structural changes and high rates of gene conversion. *PLoS Genet.* **13**, 1–20 (2017).

62. S. G. Rozen, *et al.*, AZFc Deletions and Spermatogenic Failure: A Population-Based Survey of 20,000 Y Chromosomes. *Am. J. Hum. Genet.* **91**, 890–896 (2012).

63. J. Lange, *et al.*, Isodicentric Y chromosomes and sex disorders as byproducts of homologous recombination that maintains palindromes. *Cell* **138**, 855–869 (2009).

64. B. Lemos, A. T. Branco, D. L. Hartl, 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).

65. S. Kaufmann, *et al.*, Inter-chromosomal contact networks provide insights into Mammalian chromatin organization. *PLoS One* **10**, e0126125 (2015).

66. Y. Q. S. Soh, *et al.*, Sequencing the mouse Y chromosome reveals convergent gene acquisition and amplification on both sex chromosomes. *Cell* **159**, 800–813 (2014).

67. T.-C. Chang, Y. Yang, E. F. Retzel, W.-S. Liu, Male-specific region of the bovine Y chromosome is gene rich with a high transcriptomic activity in testis development. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 12373–12378 (2013).

68. J. Eriksson, *et al.*, Y-chromosome analysis confirms highly sex-biased dispersal and suggests a low male effective population size in bonobos (Pan paniscus). *Mol. Ecol.* **15**, 939–949 (2006).

69. A. Nater, *et al.*, Sex-biased dispersal and volcanic activities shaped phylogeographic patterns of extant Orangutans (genus: Pongo). *Mol. Biol. Evol.* **28**, 2275–2288 (2011).

70. K. Sahlin, M. Tomaszkiewicz, K. D. Makova, P. Medvedev, Deciphering highly similar multigene family transcripts from Iso-Seq data with IsoCon. *Nat. Commun.* **9**, 4601 (2018).

71. J. R. Dixon, *et al.*, Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331–336 (2015).

72. R. S. Harris, Improved pairwise Alignmnet of genomic DNA (2007).

73. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).

74. E. Garrison, G. Marth, Haplotype-based variant detection from short-read sequencing. *arXiv [q-bio.GN]* (2012).

75. M. J. Chaisson, G. Tesler, Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238 (2012).

76. I. Minkin, P. Medvedev, Scalable multiple whole-genome alignment and locally collinear block construction with SibeliaZ. *bioRxiv*, 548123 (2019).

77. M. Blanchette, *et al.*, Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14**, 708–715 (2004).

78. SMIT, F. A. A., Repeat-Masker Open-3.0. *http://www.repeatmasker.org* (2004) (October 23, 2018).

79. G. Hickey, B. Paten, D. Earl, D. Zerbino, D. Haussler, HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29**, 1341–1342 (2013).

80. D. Earl, *et al.*, Alignathon: a competitive assessment of whole-genome alignment methods.

Genome Res. **24**, 2077–2089 (2014).

81. D. Darriba, G. L. Taboada, R. Doallo, D. Posada, jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* **9**, 772 (2012).

82. S. Guindon, O. Gascuel, A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696–704 (2003).

83. A. Siepel, D. Haussler, Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.* **21**, 468–488 (2004).

84. S. Tavaré, Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences* **17**, 57–86 (1986).

85. E. Paradis, J. Claude, K. Strimmer, APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289–290 (2004).

86. M. J. Hubisz, K. S. Pollard, A. Siepel, PHAST and RPHAST: phylogenetic analysis with space/time models. *Briefings in Bioinformatics* **12**, 41–51 (2011).

87. M. Stanke, S. Waack, Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19 Suppl 2**, ii215–25 (2003).

88. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

89. E. Boutet, *et al.*, UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. *Methods Mol. Biol.* **1374**, 23–54 (2016).

90. D. P. Locke, *et al.*, Comparative and demographic analysis of orang-utan genomes. *Nature* **469**, 529–533 (2011).

91. W. J. Kent, *et al.*, The Human Genome Browser at UCSC. *Genome Research* **12**, 996–1006 (2002).

92. J. Krumsiek, R. Arnold, T. Rattei, Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**, 1026–1028 (2007).

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

94. W. Miller, *et al.*, 28-Way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Research* **17**, 1797–1808 (2007).

95. H. Li, *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

96. F. Ramírez, *et al.*, deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–5 (2016).

97. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic

features. *Bioinformatics* **26**, 841–842 (2010).

98. I. Dunham, *et al.*, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

99. J. M. Gaspar, Improved peak-calling with MACS2. *bioRxiv*, 496521 (2018).

100. J. T. Robinson, *et al.*, Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

101. P. A. Knight, D. Ruiz, A fast algorithm for matrix balancing. *IMA J. Numer. Anal.* **33**, 1029–1047 (2013).

Figure 1. Phylogenetic tree of nucleotide sequences. (**A**) Y chromosome, and (**B**) autosomes. Branch lengths (substitutions per 100 sites) were estimated from multi-species alignment blocks including all five species.

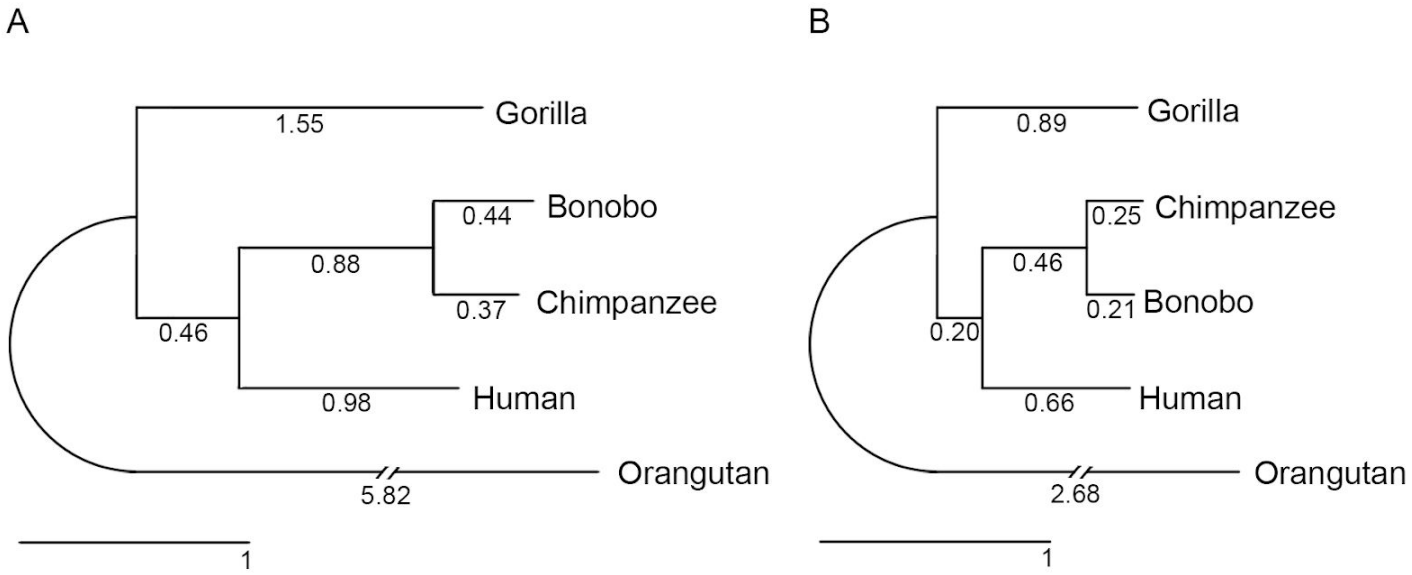

Figure 2. Evolution of Y chromosome gene content in great apes. The reconstructed history of gene birth and death for X-degenerate (blue) and ampliconic (red) genes was overlaid on the great ape phylogenetic tree (not drawn to scale), using macaque as an outgroup. The rates of gene birth and death (in events per million years) are shown in parentheses (for complete data see **Fig. S4**). The list at the root includes the genes that were present in the common ancestor of great apes and macaque. In addition to most of the genes on the human Y, the macaque Y harbors the X-degenerate *MXRA5Y* gene, which we found to be deleted in orangutan and pseudogenized in bonobo, chimpanzee, gorilla, and human. We currently cannot find a full-length copy of the *VCY* gene in bonobo. *TXLNGY* and *DDX3Y* are also known as *CYorf15B* and *DBY*, respectively.

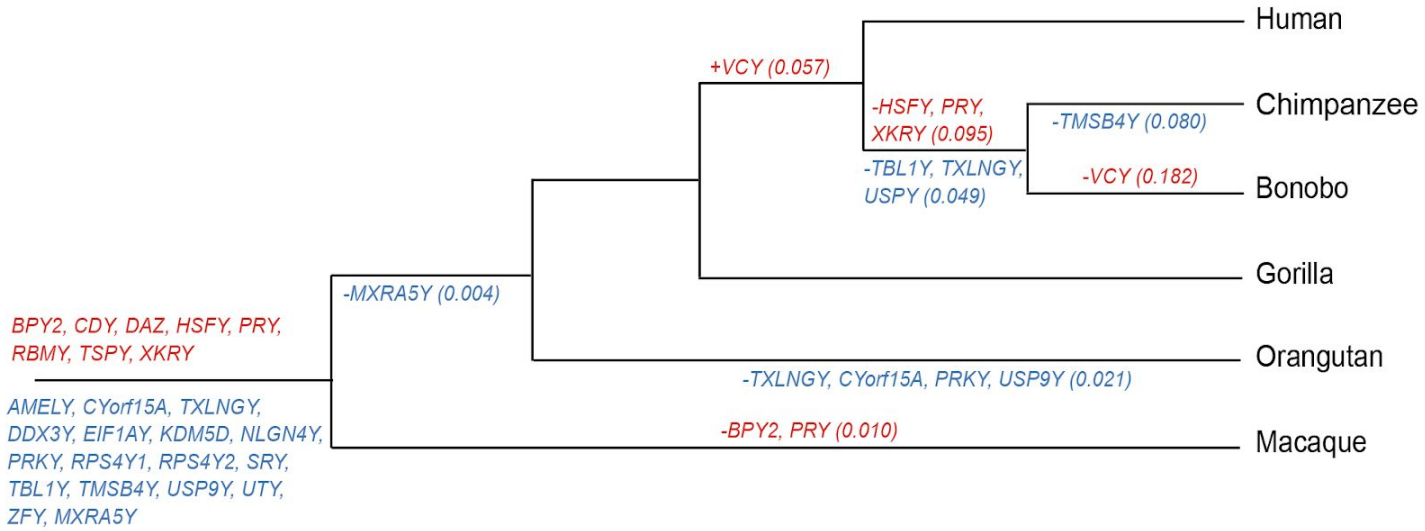

Figure 3. Evolution of sequences homologous to human and chimpanzee palindromes. (A) Heatmaps showing coverage for each palindrome in each species in the multi-species alignment, and box plots representing copy number (natural log) of 1-kb windows which have homology with human or chimpanzee palindromes. **B.** The great ape phylogenetic tree with evolution of human (shown in blue) and chimpanzee (purple) palindromic sequences overlayed on it. Palindrome names in bold indicate that their sequences were present in ≥2 copies. Negative (-) and positive (+) signs indicate gain and loss of palindrome sequence (possibly only partial), respectively. Arrows represent gain (↑) or loss (↓) of palindrome copy number. If several equally parsimonious scenarios were possible, we assumed a later date of acquisition of the multi-copy state for a palindrome (**Note S4**).

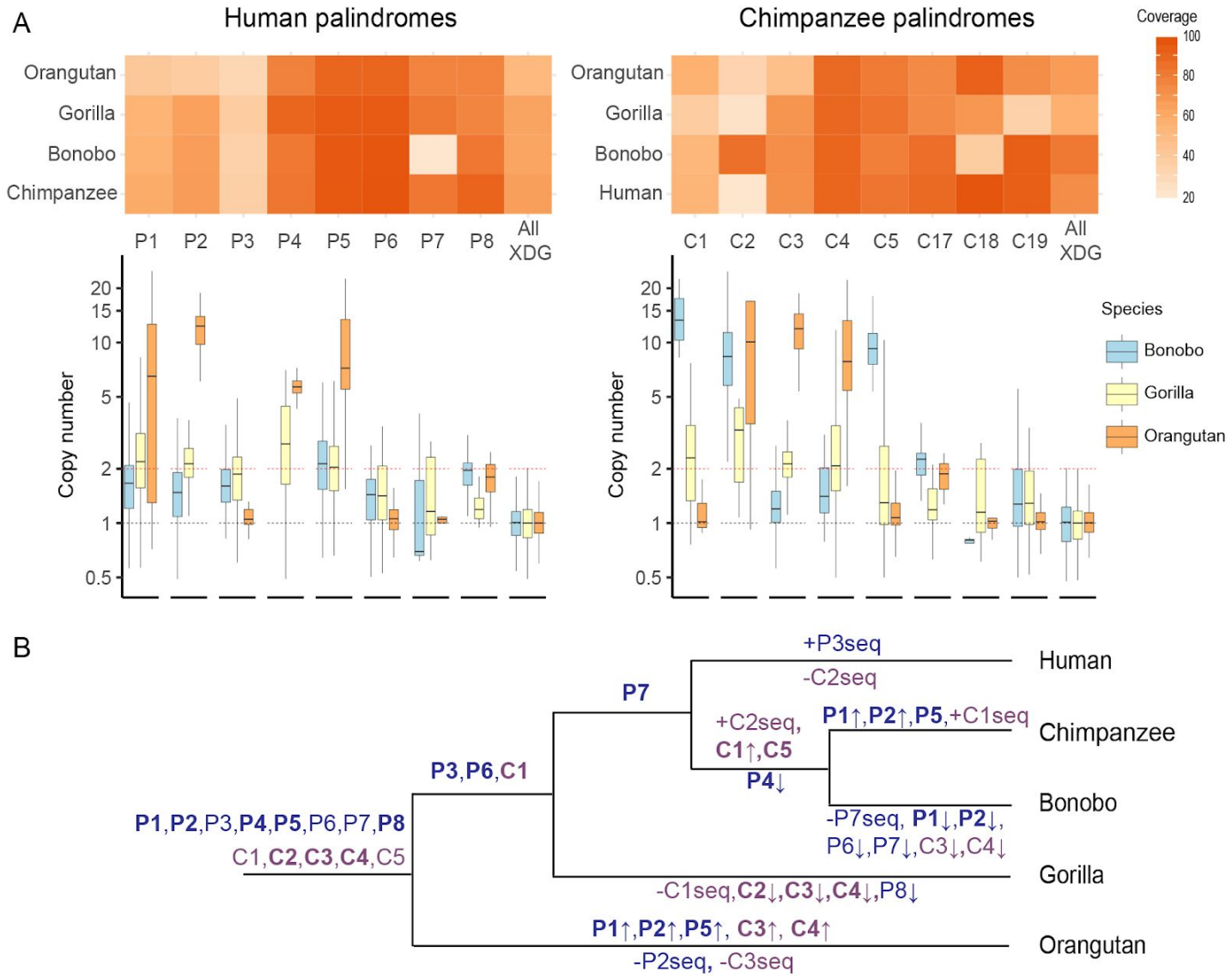
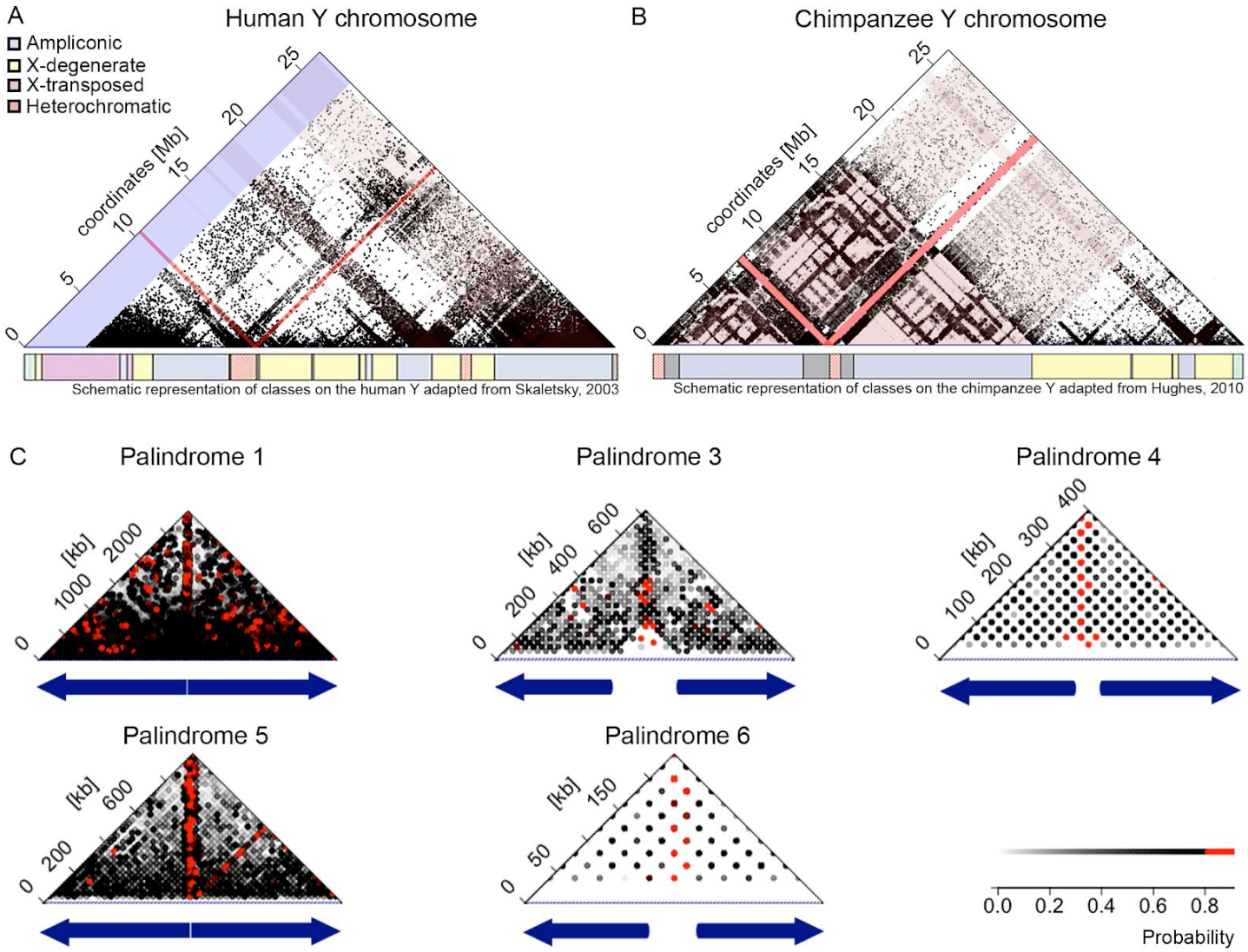
Figure 4. Chromatin contacts on the human and chimpanzee Y chromosomes, as evaluated from iPSC. **A.** Human Y chromosome contacts with palindromes (highlighted in pink), pseudoautosomal regions (blue), and centromere (red). The schematic representation of the sequencing classes on the Y chromosome is adapted from (Skaletsky et al. 2003). **B.** Chimpanzee Y chromosome contacts with palindromes (highlighted in pink). The schematic representation of the sequencing classes on the Y chromosome is adapted from (Hughes et al. 2010). **C**. Chromatin interactions for the five largest palindromes on the human Y. To resolve ambiguity due to multi-mapping reads, each interaction was assigned a probability based on the fraction of reads supporting it. Palindrome arms are shown as blue arrows and the spacer as white space between them.