Large palindromes on the primate X Chromosome are preserved by natural selection

Emily K. Jackson,1,2,3 Daniel W. Bellott,1 Ting-Jan Cho,1 Helen Skaletsky,1,2 Jennifer F. Hughes,1 Tatyana Pyntikova,1 and David C. Page1,2,3

1Whitehead Institute, Cambridge, Massachusetts 02142, USA; 2Howard Hughes Medical Institute, Whitehead Institute, Cambridge, Massachusetts 02142, USA; 3Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Mammalian sex chromosomes carry large palindromes that harbor protein-coding gene families with testis-biased expression. However, there are few known examples of sex-chromosome palindromes conserved between species. We identified 26 palindromes on the human X Chromosome, constituting more than 2% of its sequence, and characterized orthologous palindromes in the chimpanzee and the rhesus macaque using a clone-based sequencing approach that incorporates full-length nanopore reads. Many of these palindromes are missing or misassembled in the current reference assemblies of these species’ genomes. We find that 12 human X palindromes have been conserved for at least 25 million years, with orthologs in both chimpanzee and rhesus macaque. Insertions and deletions between species are significantly depleted within the X palindromes’ protein-coding genes compared to their noncoding sequence, demonstrating that natural selection has preserved these gene families. The spacers that separate the left and right arms of palindromes are a site of localized structural instability, with seven of 12 conserved palindromes showing no spacer orthology between human and rhesus macaque. Analysis of the 1000 Genomes Project data set revealed that human X-palindrome spacers are enriched for deletions relative to arms and flanking sequence, including a common spacer deletion that affects 13% of human X Chromosomes. This work reveals an abundance of conserved palindromes on primate X Chromosomes and suggests that protein-coding gene families in palindromes (most of which remain poorly characterized) promote X-palindrome survival in the face of ongoing structural instability.

[Supplemental material is available for this article.]

The human X Chromosome contains two classes of genomic sequence with distinct evolutionary histories (Mueller et al. 2013). Most human X-Chromosome sequence is ancestral, containing genes retained from the ancestral autosome from which the mammalian sex chromosomes evolved some 200–300 million years ago (Ohno 1967; Lahn and Page 1999); these ancestral genes display diverse patterns of expression. However, around 2% of the human X Chromosome comprises ampliconic sequence, which contains genes that are expressed predominantly in testis and that were not present on the ancestral autosome (Mueller et al. 2013). Amplicons are long, highly identical repeat units, with lengths that can exceed 100 kb in length and sequence identities >99% (Kuroda-Kawaguchi et al. 2001). Although some amplicons form tandem arrays, the majority comprise large palindromes, with mirrored repeats facing head to head (Skaletsky et al. 2003). The testis-biased expression of human X-Chromosome palindromic gene families has led to speculation that X palindromes play roles in spermatogenesis (Warburton et al. 2004; Mueller et al. 2013), consistent with evolutionary theories predicting that the X Chromosome should preferentially accumulate male-beneficial alleles (Rice 1984). In contrast to ancestral sequence, however, few studies have focused on X-Chromosome palindromes, and X palindromes have not been characterized in nonhuman primates. As a result, little is known about the origins of X-Chromosome palindromes or the evolutionary forces that shape them.

Studies of X-Chromosome palindromes have been impeded by the difficulty of obtaining accurate reference sequence. Segmental duplications are commonly collapsed by assembly algorithms into a single repeat unit (Eichler 2001), and they are severely underrepresented in reference genomes assembled using short sequencing reads and whole-genome shotgun (WGS) algorithms (Alkan et al. 2011). Several mammalian Y Chromosomes, which also contain palindromes and other amplicons (Skaletsky et al. 2003; Hughes et al. 2010, 2012, 2020; Soh et al. 2014), were sequenced using labor-intensive but highly accurate clone-based approaches (Kuroda-Kawaguchi et al. 2001; Bellott et al. 2018). Recently, the incorporation of ultralong nanopore reads into a clone-based sequencing approach (Single Haplotype Iterative Mapping and Sequencing 3.0, or SHIMS 3.0) has enabled the time- and cost-effective resolution of amplicons, including the TSPY1 array on the human Y Chromosome, that had been impervious to previous assembly methods (Bellott et al. 2020). Accurate representation of amplicons and other segmental duplications in mammalian genomes is particularly important given the disproportionate roles of segmental duplications in mediating deletions, duplications, inversions, and other complex rearrangements across the human genome (Stankiewicz and Lupski 2010).

Only two mammalian X Chromosomes have been sequenced using a high-resolution, clone-based approach: the mouse X...
Chromosome (Church et al. 2009) and the human X Chromosome (Ross et al. 2005; Mueller et al. 2013). Palindromes containing testis-biased gene families are abundant on both the human and mouse X Chromosomes (Warburton et al. 2004; Mueller et al. 2008); however, X-palindrome gene content largely does not overlap between the two species, suggesting that the palindromes are not orthologous (Mueller et al. 2013). A subset of human X-palindrome gene families have highly divergent orthologs on the chimpanzee X Chromosome, but without a complete reference sequence, their copy number and orientations could not be determined (Stevenson et al. 2007). To determine whether human X palindromes have orthologs in other primates, we used SHIMS 3.0 to generate high-quality reference sequence for portions of the X Chromosome that are orthologous to human X palindromes in two nonhuman primate species, the chimpanzee and the rhesus macaque.

**Results**

**Characterization of 26 palindromes on the human X Chromosome**

To ensure consistency in palindrome annotation between species, we began by re-annotating human X palindromes. We identified palindromic amplicons by searching the reference sequence of the human X Chromosome for inverted repeats >8 kb in length (which excludes repetitive elements such as LINEs and SINEs) that display >99% nucleotide identity (Fig. 1A). We used a k-mer-based method (Teitz et al. 2018) to precisely define the coordinates of each palindrome (see Methods) and triangular dot plots to visualize palindromes and other genomic repeats (Fig. 1B,C). In total, we identified 26 palindromic amplicons on the human X Chromosome (Table 1; Supplemental Table S1); these palindromes, including both arms and spacer sequence, comprise over 3.46 Mb, or 2.2% of the length of the human X Chromosome. Palindrome arm lengths range from 8 kb to 140 kb, with arm-to-arm identities as high as 99.99%, representing one nucleotide difference between arms per 10,000 base pairs.

Palindrome spacer lengths span several orders of magnitude, ranging from 77 bp to 358 kb. The vast majority of palindromes (21 of 26) have at least one protein-coding gene in their arms (Supplemental Table S2). No primate Y palindromes have been observed with protein-coding genes in their spacers (Skaletsky et al. 2003; Hughes et al. 2010, 2012), but we found that 11 of 26 human X palindromes also contain at least one protein-coding gene in their spacer (Supplemental Table S2).

To investigate the functions of palindrome-encoded genes, we examined the expression patterns of all 45 human X-palindrome arm and spacer genes using data from the Genotype-Tissue Expression (GTEx) Consortium (The GTEx Consortium 2017). Palindrome arm genes are present in two nearly identical copies, and RNA sequencing reads map equally well to both, which poses a problem for accurately estimating the expression level of genes in palindrome arms using traditional software that discards multimapping reads, as was recently shown for ampliconic genes on the Y Chromosome (Godfrey et al. 2020). We therefore re-analyzed GTEx data with kallisto, which assigns multimapping reads based on a probability distribution (Bray et al. 2016; Godfrey et al. 2020). We found that 18 of 30 gene families (60%) in palindrome arms were expressed predominantly in testis, consistent with previous reports (Table 1; Supplemental Fig. S1; Warburton et al. 2004). Genes in palindrome spacers showed similar expression patterns: 10 of 17 genes (58.8%) were expressed predominantly in testis (Table 1; Supplemental Fig. S1), suggesting that reproductive specialization of palindrome genes includes both arm genes and spacer genes. We further examined the expression of testis-biased gene families in palindrome arms and spacers across human spermatogenesis using bulk RNA-seq data from Jan et al. (2017) (Supplemental Fig. S2). Across the 20 testis-biased gene families with detectable expression at one or more time points, we observed patterns ranging from highest expression in spermatogonia to highest expression in round spermatids, suggesting that human X-palindrome gene families play roles across multiple stages of spermatogenesis.

Palindromes on the human Y Chromosome are depleted for LINEs and other transposable elements (Skaletsky et al. 2003).

![Figure 1](image-url)
Table 1. Palindromes on the human X Chromosome

| Palindrome | Arm length (kb) | Spacer length (kb) | Arm-to-arm ID (%) | Arm gene families (two copies)* | Spacer genes (one copy)* |
|------------|----------------|-------------------|------------------|-------------------------------|-------------------------|
| P1         | 29.0           | 2.8               | 99.98            | SSX4                          | cGenes annotated as protein-coding in one arm but as a pseudogene in the other. SeeSupplemental Note S1. |
| P2         | 24.9           | 7.4               | 99.92            | CENPVL                        |
| P3         | 36.4           | 99.9              | 99.99            | MAGED4                        |
| P4         | 28.8           | 8.4               | 99.98            | XAGE1                         |
| P5         | 58.7           | 0.5               | 99.97            | SSX2                          |
| P6         | 37.9           | 15.6              | 99.98            | FAM156B                       |
| P7         | 26.6           | 12.9              | 99.97            | CXorf49                       |
| P8         | 57.3           | 0.5               | 99.96            | DMRTC1, FAM236B,             |
| P9         | 119.1          | 0.4               | 99.95            | FAM236D                       |
| P10        | 9.2            | 72.5              | 99.95            | NKF2, TCP11X                  |
| P11        | 140.6          | 10.8              | 99.96            | TMS815B                       |
| P12        | 32.0           | 62.9              | 99.98            | H2BW1, H2BW2                  |
| P13        | 48.6           | 62.7              | 99.90            | RHOXF2                        |
| P14        | 42.0           | 56.3              | 99.93            | RHOXF1                        |
| P15        | 11.0           | 13.9              | 99.89            | ETDB                          |
| P16        | 109.3          | 358.4             | 99.75            | CTSS                          |
| P17        | 10.4           | 0.1               | 99.99            | LDOC1, SPANXc                 |
| P18\*      | 20.0           | 355.1             | 99.66            | EOLA, HSFX3                   |
| P19\*      | 29.1           | 128.9             | 99.91            | MAGEA9, HSFX2                 |
| P20\*      | 26.6           | 41.0              | 99.91            | MAGEA9, TMEM185A, MAGEA9,      |
| P21        | 46.8           | 17.7              | 99.86            | MAGEA9, TMEM185A, MAGEA11,    |
| P22        | 8.4            | 1.2               | 99.89            | MAGEA3, MAGEA2, CSAG2         |
| P23        | 43.5           | 101.0             | 99.88            | CSAG1, MAGEA12                |
| P24        | 10.1           | 37.6              | 99.87            | FLNA, EMD                     |
| P25        | 35.5           | 21.8              | 99.95            | MAGEA1                        |
| P26        | 50.3           | 67.3              | 99.98            | F8A, H2AB                     |

* Genes and gene families with testis-biased expression in GTEx (minimum 2 TPM in testis, and testis accounts for >25% of log2 normalized expression summed across all tissues) are shown in bold. \(^{b}\)PWW4 was not expressed (<2 TPM in all tissues). All other genes are expressed but not testis-biased. Individual genes within each arm gene family are listed in Supplemental Table S2.

We used RepeatMasker to examine the density of transposable elements in human X-palindrome arms and found that transposable elements (LINEs, SINEs, LTRs, and DNA elements) account for 50.2% of palindromic arms, compared to 58.1% of flanking sequence (\(P<0.05\), Mann–Whitney U test). The density of transposable elements is negatively correlated with recombination rates in palindrome arms (Rozen et al. 2003). Structural confidence due to the generation of full-length nanopore reads spanning each clone: For 83 of 107 clones sequenced using SHIMS 3.0, we were able to verify the accuracy of internal repeat structures by comparing the finished sequence to one or more full-length nanopore reads (Fig. 2B; Supplemental Table S3).

Our assemblies revealed 39 palindromes in total on the chimpanzee and rhesus macaque X Chromosomes, collectively comprising 4.90 Mb of palindromic amplicon sequence. Only 10 of these palindromes (25.6%) were represented accurately in existing X-Chromosome assemblies that were generated using primarily short-read whole-genome shotgun approaches (chimpanzee, Pan_tro_3.0; rhesus macaque, Mmul_8.0.1) (Fig. 2B; Supplemental Table S3).

We also compared our SHIMS 3.0 assemblies to two assemblies generated using long-read WGS approaches incorporating Pacific Biosciences (PacBio) reads (chimpanzee, Clint_PTRv2; rhesus macaque, Mmul_10) (Kronenberg et al. 2018; Warren et al. 2020). Here, we found that, although 18 palindromes (46.2%) were represented accurately, 21 palindromes (53.8%) remained missing or incomplete (Fig. 2C; Supplemental Fig. S3; Supplemental Tables S4, S5; Zimin et al. 2014; Kuderna et al. 2017). We also compared our SHIMS 3.0 assemblies to two assemblies generated using long-read WGS approaches incorporating Pacific Biosciences (PacBio) reads (chimpanzee, Clint_PTRv2; rhesus macaque, Mmul_10) (Kronenberg et al. 2018; Warren et al. 2020). Here, we found that, although 18 palindromes (46.2%) were represented accurately, 21 palindromes (53.8%) remained missing or incomplete (Fig. 2C; Supplemental Fig. S3; Supplemental Tables S4, S5; Zimin et al. 2014; Kuderna et al. 2017). We also compared our SHIMS 3.0 assemblies to two assemblies generated using long-read WGS approaches incorporating Pacific Biosciences (PacBio) reads (chimpanzee, Clint_PTRv2; rhesus macaque, Mmul_10) (Kronenberg et al. 2018; Warren et al. 2020). Here, we found that, although 18 palindromes (46.2%) were represented accurately, 21 palindromes (53.8%) remained missing or incomplete (Fig. 2C; Supplemental Fig. S3; Supplemental Tables S4, S5; Zimin et al. 2014; Kuderna et al. 2017).
We therefore carried out all subsequent analyses using our newly generated SHIMS 3.0 assemblies.

Conservation of X palindromes in two nonhuman primates

We annotated palindromes in chimpanzee and rhesus macaque using the same criteria used to annotate human palindromes (minimum 8-kb arm length, and minimum 99% nucleotide identity between arms). We also included one palindrome in rhesus macaque with arms of 6.5 kb, given that the palindrome exhibited 99% arm-to-arm identity and was orthologous to human palindrome P10 (Supplemental Fig. S5). In total, we discovered 21 palindromes in chimpanzee and 18 palindromes in rhesus macaque, demonstrating that X-linked palindromes are not unique to humans but represent a common feature of primate X Chromosomes, at least to Old World monkeys (Fig. 3A).

We applied two criteria to identify orthologs of human X palindromes (Supplemental Fig. S4A,B). First, we required that at least 20% of each palindrome arm in chimpanzee or rhesus macaque align to its putative human ortholog. This excluded three palindromes; we suggest that despite their similar genomic positions, these palindromes arose independently in each lineage (Supplemental Fig. S4C). Second, we required that the portion of palindrome arms that aligned between species have minimal alignment to flanking sequence, defined through high-quality BLAST hits (see Methods). This excluded two palindromes in which the palindrome arm in rhesus macaque mapped equally well to more than two locations in the human sequence, including flanking sequence, consistent with scenarios in which similar palindromes arose independently in each species (Supplemental Fig. S4D). We note that this approach for defining orthologous palindromes is conservative and may exclude palindromes that were
present in the common ancestor but underwent extreme rearrangements in one or more species.

After excluding four regions for which we were not able to generate SHIMS 3.0 assemblies, we found that the vast majority of the remaining human palindromes—20 of 22—have an orthologous palindrome in chimpanzee; the same is true for 14 of 24 palindromes in rhesus macaque (Fig. 3A). For each species, we annotated protein-coding genes in our newly generated sequence and constructed dot plots with accompanying tracks showing the positions and orientations of palindrome arms and genes for each region (Fig. 3B; Supplemental Fig. S5). Comparisons of human sequence with nonhuman primate sequence demonstrated that these palindromes harbor orthologous protein-coding genes (Fig. 3B). Previous literature has reported that nucleotide identity between sex-chromosome palindrome arms is maintained by ongoing gene conversion (Rozen et al. 2003; Skaletsky et al. 2003; Swanepoel et al. 2020); consistent with this, we find that the average arm-to-arm divergence within a species (<0.2% for all species) is lower than the average arm-to-arm divergence between species (0.87% and 5.50% for human vs. chimpanzee and human vs. rhesus macaque, respectively) (Fig. 3C). We conclude that nearly half of human X palindromes—12 of 26—were present in the common ancestor of human, chimpanzee, and rhesus macaque at least 25 million years ago and have been maintained by ongoing gene conversion in all three lineages since their divergence. We subjected these 12 palindromes shared by human, chimpanzee, and rhesus macaque to further analyses to uncover the processes governing their evolution.

Structural changes between species are concentrated around the center of symmetry

Palindromes on the human X and Y Chromosomes are associated with a wide range of pathogenic rearrangements, the majority of which result from nonallelic homologous recombination (NAHR) between near-identical palindrome arms (Lakich et al. 1993; Small et al. 1997; Aradhya et al. 2001; Lange et al. 2009; Scott et al. 2010). Although palindromes are well-known sites of genomic instability in humans, little is known about the stability of palindromic structures on longer evolutionary timescales. We used our set of 12 X palindromes shared by human, chimpanzee, and macaque to investigate structural changes in the palindromes over time.

NAHR between palindrome arms has two common outcomes: NAHR within a single chromatid results in inversions of the spacer, whereas NAHR between misaligned sister chromatids results in acentric and dicentric fragments (Lange et al. 2009). Given that acentric and dicentric fragments are not expected to be stably inherited, we instead looked for evidence of spacer inversions between species. Inversions within X palindromes have been previously reported to exist as neutral polymorphisms in human populations (Small et al. 1997), as well as pathogenic
rearrangements (Lakich et al. 1993; Porubsky et al. 2020). We found abundant evidence for inversions in primate X palindromes: In cases where the orientation of the spacer could be confidently determined, the frequency of inversions between chimpanzee and human, and between human and rhesus macaque, was about 50% (Fig. 4A; Supplemental Fig. S5). This is consistent with the notion that inversions are common events, the majority of which are not harmful, except in rare instances where they disrupt a protein-coding gene. Our SHIMS 3.0 assemblies each derive from a single male individual, but in light of the results from Small and colleagues, who found that human P24 spacers are inverted in 18% of European X Chromosomes, we would anticipate that inversion polymorphisms exist within all three species. Indeed, a recent study found that 10 out of 13 newly identified X-Chromosome inversion polymorphisms in chimpanzee and/or bonobo occurred in X palindromes (Porubsky et al. 2020).

We observed numerous examples in which spacer sequence could not be aligned between species, despite robust alignment of most or all of the palindrome arms (Fig. 4A; Supplemental Fig. S6). This phenomenon was observed in two of 12 spacer comparisons between human and chimpanzee and seven of 12 spacer comparisons between human and rhesus macaque (Fig. 4A). Nonorthologous spacers could be the result of insertions, deletions, or translocations of sequence within a palindrome. Rather than attempting to reconstruct each event, we used the fraction of sequence that is orthologous as a simplified metric for sequence rearrangements and asked whether rearrangements are concentrated within spacers compared to palindrome arms or flanking sequence. We aligned each spacer between species, calculated the fraction of orthologous sequence, and repeated this for palindrome arms and flanking sequence. When comparing human versus chimpanzee, the average fraction of orthologous sequence was lowest in spacers (75.8% for spacers vs. 96.5% and 94.7% for flanking sequence and arms, respectively), though the result did not reach statistical significance (Fig. 4B). When comparing human versus rhesus macaque, the average fraction of orthologous sequence was significantly lower in spacers (35.2%) than in arms (72.9%) or flanking sequences (63.8%) (*P < 0.01 for both, Mann–Whitney U test) (Fig. 4B). Consistent with our observations in Figure 4A, these results were driven by a subset of palindromes in

**Figure 4.** Structural changes between orthologous X-Chromosome palindromes are concentrated around the center of symmetry. (A) Square dot plots comparing the center of the palindrome, including the spacer and 10 kb of inner arm sequence on each side, between the indicated species. (Orthologous spacer) >20% of the spacer from one species aligned to the spacer from the other, in either the same orientation (“Human configuration”) or opposite orientation (“Inversion”); (Non-orthologous) <20% of the spacer from one species aligned to the spacer from the other. Values show the number of orthologous palindromes shared by human, chimpanzee, and macaque in each category. (B) Average fraction of sequence that could be aligned between species. (*) P < 0.05, (**) P < 0.01, Mann–Whitney U test. (C) Sizes of human spacers, binned according to the species between which they are conserved.
which the spacer displayed little or no similarity between species, rather than a small and consistent decrease in similarity affecting all palindromes equally (Fig. 4B). Spacer size was positively correlated with the degree of conservation between species, suggesting that small spacers may be particularly unstable (Fig. 4C). We conclude that, in addition to inversions, palindrome rearrangements are concentrated around the center of symmetry and that palindromes with small spacers are most susceptible to rearrangement.

Natural selection has preserved palindrome gene families
We wondered whether natural selection might provide a countervailing force against structural instability in primate X palindromes. Among the 12 human X palindromes with orthologous palindromes in chimpanzee and rhesus macaque, there are 17 protein-coding gene families in palindrome arms, and five protein-coding genes in palindrome spacers (Supplemental Table S6). The functions of these gene families are poorly characterized in humans: Only one of 17 human palindrome arm gene families (6%) have phenotypes listed in the Online Mendelian Inheritance in Man (OMIM) database (McKusick-Nathans Institute of Genetic Medicine 2020). This represents a fourfold depletion relative to other protein-coding genes on the human X Chromosome, of which 221 of 823 (26.4%) are associated with an OMIM phenotype \( (P<0.05, \text{hypergeometric test}) \). Out of five human palindrome spacer genes, two are associated with OMIM phenotypes: EMD and FLNA, both broadly expressed genes found in the spacer of P24, which are associated with muscular dystrophy and neurological disorders, respectively (Table 1; Bione et al. 1994; Fox et al. 1998; Clapham et al. 2012).

Despite their limited functional characterization, we find that palindrome gene families are well conserved across primates. All 17 gene families in human palindrome arms have at least one intact gene copy in chimpanzee, and 15 of 17 have at least one intact gene copy in rhesus macaque (Supplemental Table S6). Among spacer genes, four of five have at least one intact gene copy in chimpanzee and three of five in rhesus macaque; the two genes with ascribed OMIM phenotypes are conserved in all three species (Supplemental Table S6). Three out of four arm and spacer gene families that are not conserved across all three species have paralogs with at least 85% protein identity elsewhere on the X Chromosome, which may reduce the impact of their loss. Gene families from palindromes shared by human, chimpanzee, and macaque also have conserved expression patterns: 20 of 21 such gene families have the same expression pattern in chimpanzee and human (Supplemental Fig. S7), and the same is true for 16 of 18 gene families conserved in rhesus macaque (Supplemental Fig. S8).

The conservation of palindrome gene families suggests that they are subject to purifying selection. Consistent with this, we found that the ratio of nonsynonymous to synonymous substitution rates \( (d_{ns}/d_s) \) was below 1.0 for 17 of 18 arm and spacer gene families conserved between all three species (Supplemental Table S7), 12 of which were significant using a likelihood ratio test (Supplemental Table S7). Nonetheless, the median \( d_{ns}/d_s \) value for 18 X-palindrome gene families is 0.36, compared to a median \( d_{ns}/d_s \) value of 0.12 for protein-coding genes in the genome (Gayà-Vidal and Albà 2014; Biswas et al. 2016). Elevated \( d_{ns}/d_s \) values could result from either relaxed purifying selection or from positive selection at one or more sites. We therefore also performed a likelihood ratio test for positive selection across all 18 gene families and found evidence of positive selection for two gene families (Supplemental Table S7). We note that, with only three species for comparison, we were likely underpowered to detect positive selection, and our results should not be interpreted as evidence against positive selection in the other 16 gene families (Anisimova et al. 2001).

If palindrome gene families are subject to purifying selection, then we also predicted that they should be depleted for insertions and deletions (indels) between species. To define indels, we used a k-mer-based method to identify stretches of at least 1 kb that lacked orthologous sequence in the other species (Fig. 5A). We then compared the fraction of bases falling within indels for protein-coding gene sequence (including exons, introns, and 1 kb

---

**Figure 5.** Natural selection has preserved X-palindrome gene families. Results based on 12 palindromes conserved between human, chimpanzee, and rhesus macaque. (A) Square dot plot comparing structure of P25 between human and rhesus macaque. Indels highlighted in gray; nearly all fall between protein-coding genes. (B) Fraction of bases within indels for protein-coding gene sequence versus all other sequence. Results are the average of all pairwise species comparisons. Indels were defined as uninterrupted stretches of at least 1 kb in one species without orthologous sequence in the other species. \( * \) \( P<0.01 \), \( ** \) \( P<0.001 \), Mann-Whitney \( U \) test.
upstream) versus other sequence. We performed this analysis individually for palindrome arms, palindrome spacers, and flanking sequence, revealing a significant depletion of indels within protein-coding gene sequence for all three regions (Fig. 5B). We conclude that natural selection has preserved protein-coding gene families in primate X-palindrome arms and spacers, despite their limited functional characterization in humans.

Enrichment of spacer deletions in human X-Chromosome palindromes

Having observed localized structural instability in X palindromes between primate species, we asked whether we could detect signatures of structural instability within the human population. To address this question, we used whole-genome sequencing data from the 1000 Genomes Project (The 1000 Genomes Project Consortium 2015). The 1000 Genomes Project data set consists of short Illumina reads, which are not conducive to finding insertions or rearrangements. Instead, we asked whether we could detect X-palindrome spacer deletions, which we predicted would result in loss of sequence coverage over the palidine spacer.

We searched for X-palindrome spacer deletions among 944 individuals from the 1000 Genomes Project. We limited our analysis to male samples because males have only one X Chromosome; this enabled us to analyze deletions among 944 X Chromosomes and ensured that coverage depth over X-Chromosome deletions should be near zero. To identify palindrome spacer deletions, we screened for X Chromosomes with low normalized spacer coverage depth (Fig. 6A; also see Methods). We found four X Chromosomes with near-zero coverage in the spacer of P2 (Fig. 6A); visual inspection of coverage depth revealed that all four X Chromosomes had a deletion of about 25 kb, spanning not only the palindrome spacer but part of the inner palindrome arm (Fig. 6B). We performed the same analysis for the remaining 25 X palindromes, identifying a total of 149 palindrome spacer deletions across nine different palindromes (Table 2; Supplemental Fig. S9).

Although deletions were identified based on low depth of coverage in spacers, most breakpoints fell within palindrome arms (Fig. 6B; Table 2). In total, 145 of 944 X Chromosomes from the 1000 Genomes data set (15.4%) had a spacer deletion in at least one palindrome. Deletions ranged from 3 kb up to 587 kb in size, and in four cases removed one or more copies of a protein-coding gene family (Table 2). In two cases, breakpoints fell within tandem repeats, suggesting that they arose through NAHR (Supplemental Fig. S10). All other breakpoints fell within unique sequence, suggesting origins independent of NAHR (see Discussion). For the nine human X palindromes in which we identified at least one X Chromosome with a spacer deletion, we examined the structure of orthologous palindromes in chimpanzee. We found that sequence absent in one or more human X Chromosomes is present in the chimpanzee X Chromosome, confirming that the human structural polymorphisms result from deletions rather than insertions (Supplemental Fig. S11).

Out of a total of 149 palindrome spacer deletions, 126 were found in a single palindrome, P17 (Fig. 6C; Table 2). We wondered whether these represented independent deletion events, in which case we would expect different breakpoints in different X Chromosomes. Alternatively, if they represented a common structural polymorphism, all breakpoints should be identical. The deletion breakpoints for all 126 P17 deletions appeared similar by eye (Supplemental Fig. S12); we subsequently identified split reads from the 1000 Genomes Project spanning the same breakpoint for all 126 X Chromosomes and further verified this shared breakpoint by PCR in five individuals selected at random (Supplemental Figs. S13, S14). We conclude that this deletion is a common polymorphism. This P17 deletion spans the palindrome’s spacer and inner arm, removing one copy of CXorf51, a testis-expressed gene family not associated with any phenotypes reported in OMIM (Table 1; Supplemental Table S2). The P17 spacer deletion is found in all five superpopulations in 1000 Genomes, with frequencies ranging from 3% (Africa) to 23% (South Asia) of X Chromosomes (Fig. 6D). The 1000 Genomes data set does not include phenotype information; however, we speculate that the rise of the P17 spacer deletion to high frequency is incompatible with strong reductions in viability or fertility and that phenotypic effects, if any, are likely to be mild.

To determine whether genomic instability is elevated within palindrome spacers, we asked whether deletions were more common in palindrome spacers than in arms and flanking sequence. For each of 26 human X palindromes, we randomly selected regions of the arm and flanking sequence of the same size as the spacer, and we counted deletions by the criteria described above. For spacers, we observed deletions in nine different palindromes; in contrast, we never observed deletions in more than four palindromes using size-matched regions from palindrome arms and flanking sequence (P < 0.01, bootstrapping analysis) (Fig. 6E). The difference was larger in absolute terms: 149 deletions in spacers, with no more than 19 deletions in arms and flanking sequence (P < 0.01, bootstrapping analysis) (Fig. 6F). We conclude that structural instability of X-palindrome spacers has persisted in our own species, and that one manifestation of this instability is deletions. Insertions and rearrangements are possible as well but would not have been detected by our analysis.

Two polymorphic human X-palindrome spacer deletions are not associated with azoospermia

Given that primate X-palindrome gene families are preserved by natural selection, we wondered whether deletions that remove one or more copies of human X-palindrome gene families negatively impact fitness. The two most common human X-palindrome spacer deletions from the 1000 Genomes Project each remove one copy of an uncharacterized testis-expressed gene family: CXorf51 and CXorf49, respectively (Table 2; Supplemental Table S2). Because deletion of testis-expressed Y-palindrome gene families causes azoospermia (Vogt et al. 1996; Kuroda-Kawaguchi et al. 2001), we asked whether deletions of CXorf51 and CXorf49 are also associated with azoospermia, using a publicly available data set containing capture-based targeted sequencing for 301 azoospermia cases and 300 normospermic controls (NCBI database of Genotypes and Phenotypes [dbGaP]; https://www.ncbi.nlm.nih.gov/gap/ accession number phs000102.v3.p1). Selecting samples that met a minimum coverage threshold, we found that CXorf51 deletions were equally prevalent in cases and controls: 47 deletions in 286 cases (16.4%), and 54 deletions in 292 controls (18.5%, ns, Fisher’s exact test) (Supplemental Table S8). We also detected two deletions—both in controls—that appear to remove both copies of the CXorf51 gene family. We found only one case with a CXorf49 deletion and therefore cannot infer an association with azoospermia. To test for milder effects on spermatogenesis, we used PCR screening to identify CXorf51 and CXorf49 deletions in a set of 563 oligozoospermic men (sperm counts 0.1–20 million per cubic cm). We found CXorf51 deletions in 69 men (12.3%), which is not significantly different from the
percentage of men from the 1000 Genomes Project (13.3%, Fisher’s exact test), suggesting that CXorf51 deletions are not enriched in oligozoospermic men (Supplemental Table S8). We identified a single CXorf49 deletion in oligozoospermic men, which is a significantly lower rate than we observed among men from the 1000 Genomes Project (0.18% vs. 1.4%, \( P < 0.05 \), Fisher’s exact test). Although our analyses do not support an association between CXorf51 or CXorf49 deletions and azoospermia or oligozoospermia, we cannot rule out more subtle defects in spermatogenesis, with resultant selection for retention of both gene copies.

**Discussion**

Massive palindromes are hallmarks of mammalian sex chromosomes, yet until now, there were few examples of sex-chromosome palindromes that are conserved between species. We provide evidence that 12 palindromes have been conserved across three primate X Chromosomes.
primate X Chromosomes for at least 25 million years, using a targeted sequencing protocol, SHIMS 3.0, that combines ultralong nanopore reads with a clone-based approach. Comparative genomic analyses of conserved X palindromes shed new light on palindrome evolution, including evidence that natural selection preserves understudied protein-coding genes within X-palindrome arms and spacers. We also report a novel structural instability of X-Chromosome palindromes: Rearrangements between species are concentrated around the center of symmetry, with a high frequency of palindrome spacer deletions observed among individuals from the 1000 Genomes Project.

The deep conservation of primate X palindromes has few parallels in the literature. Out of eight palindromic amplicons on the human Y Chromosome, only two were reported to have orthologous palindromes in rhesus macaque (Hughes et al. 2012). The genes FLNA and EMD, which are found in the human (X Chromosome) P24 spacer, are flanked by inverted repeats in 15 additional mammalian species, leading to the suggestion that these inverted repeats arose over 100 million years ago (Caceres et al. 2007). However, the inverted repeat sequence is not orthologous between all species, leaving open the possibility that they arose through independent duplications (Caceres et al. 2007). Although the chimpanzee Y Chromosome contains 19 massive palindromes, fewer than half of them have homology to human Y palindromes, and abundant rearrangements between the human and chimpanzee Y Chromosomes make it difficult to reconstruct the evolution of putative orthologs (Hughes et al. 2010). In contrast, all 12 palindromes that we found to be shared by human, chimpanzee, and rhesus macaque have at least one protein-coding gene in their arms or spacer. We conclude that palindromes are not inherently structurally stable but rather are preserved through natural selection, most likely acting to preserve the integrity of protein-coding gene families.

The discovery of human X-palindrome spacer deletions in individuals from the 1000 Genomes Project, combined with evidence that purifying selection preserves human X-palindrome gene families, raises the question of whether human X-palindrome spacer deletions are pathogenic. To date, we are aware of one published report describing a pathogenic X spacer deletion. Periventricular nodular heterotopia (PNH), a common neurological disorder, is caused by loss-of-function mutations in FLNA, a broadly expressed spacer gene in P24 that encodes the actin cross-linking protein filamin A (Fox et al. 1998). Although PNH is most frequently caused by missense mutations in FLNA, one affected family was found to have a 39-kb deletion spanning the P24 spacer and inner arm, removing spacer genes FLNA and EMD (Clapham et al. 2012). Although we found spacer deletions in nine different human X palindromes across individuals from the 1000 Genomes Project, none of the deletions occurred in P24. Given that the 1000 Genomes Project does not include phenotype information, we do not know whether any of the spacer deletions we observed are pathogenic. However, four of nine human X spacer deletions removed at least one copy of a protein-coding gene family, and two of these removed all functional copies of a gene family.

### Table 2. Palindrome spacer deletions among 944 X Chromosomes from the 1000 Genomes Project

| Palindrome | # X Chromosomes | Deletion size (kb) | Genes deleted | Breakpoint 1 | Breakpoint 2 |
|------------|-----------------|-------------------|---------------|--------------|--------------|
| P17        | 126             | 3.36              | CXorf51 (1/2 copies) | Arm | Arm |
| P8         | 14              | -47               | CXorf49 (1/2 copies) | Arm | Arm |
| P2         | 4               | 25.01             | Arm | Arm |
| P16        | 1               | -587              | LDOC1, SPANXC, SPANXA1 | Arm | Flanking |
| P11        | 1               | 9.93              | Spacer | Arm |
| P25        | 1               | 21.47             | Spacer | Arm |
| P5         | 1               | 17.89             | Arm | Arm |
| P22        | 1               | 8.46              | PNMA6 (2/2 copies) | Arm | Arm |
| P1         | 1               | 3.97              | Arm | Arm |

For arm genes, the number of gene copies removed by the deletion is indicated in parentheses. All other genes are unique genes from the palindrome spacer or flanking sequence. Deletion sizes marked as approximate (~) were estimated from changes in coverage depth; all others are exact sizes based on split reads spanning the deletion breakpoint. Individual genes within gene families CXorf51, CXorf49, and PNMA6 are listed in Supplemental Table S2.

Although the chimpanzee Y Chromosome contains 19 massive palindromes, fewer than half of them have homology to human Y palindromes, and abundant rearrangements between the human, chimpanzee, and rhesus macaque have at least one protein-coding gene in their arms or spacer. We conclude that palindromes are not inherently structurally stable but rather are preserved through natural selection, most likely acting to preserve the integrity of protein-coding gene families. Although these explanations are not mutually exclusive, our results strongly favor natural selection. First, we demonstrate that palindromes are not inherently stable structures, exhibiting localized structural instability around the center of palindrome symmetry. Rearrangements are enriched around palindrome spacers and inner arms in structural comparisons of palindromes conserved between species; indeed, seven of 12 palindromes shared by human, chimpanzee, and rhesus macaque have no spacer orthology between human and macaque. X-Chromosome palindromes remained structurally unstable during human evolution, as shown by a significant enrichment of deletions in palindrome spacers compared to palindrome arms or flanking sequence. Second, we provide multiple lines of evidence that palindrome gene families are targets of selection. Large (>1 kb) insertions and deletions in palindrome arms and spacers during the last 25 million years of primate evolution were depleted around protein-coding genes, and molecular analyses demonstrate purifying selection on protein-coding genes in X-palindrome arms and spacers. All 12 X palindromes conserved between human, chimpanzee, and rhesus macaque have at least one protein-coding gene in their arms or spacer. We conclude that palindromes are not inherently structurally stable but rather are preserved through natural selection, most likely acting to preserve the integrity of protein-coding gene families.

The discovery of human X-palindrome spacer deletions in individuals from the 1000 Genomes Project, combined with evidence that purifying selection preserves human X-palindrome gene families, raises the question of whether human X-palindrome spacer deletions are pathogenic. To date, we are aware of one published report describing a pathogenic X spacer deletion. Periventricular nodular heterotopia (PNH), a common neurological disorder, is caused by loss-of-function mutations in FLNA, a broadly expressed spacer gene in P24 that encodes the actin cross-linking protein filamin A (Fox et al. 1998). Although PNH is most frequently caused by missense mutations in FLNA, one affected family was found to have a 39-kb deletion spanning the P24 spacer and inner arm, removing spacer genes FLNA and EMD (Clapham et al. 2012). Although we found spacer deletions in nine different human X palindromes across individuals from the 1000 Genomes Project, none of the deletions occurred in P24. Given that the 1000 Genomes Project does not include phenotype information, we do not know whether any of the spacer deletions we observed are pathogenic. However, four of nine human X spacer deletions removed at least one copy of a protein-coding gene family, and two of these removed all functional copies of a gene family.

### Table 2. Palindrome spacer deletions among 944 X Chromosomes from the 1000 Genomes Project

| Palindrome | # X Chromosomes | Deletion size (kb) | Genes deleted | Breakpoint 1 | Breakpoint 2 |
|------------|-----------------|-------------------|---------------|--------------|--------------|
| P17        | 126             | 3.36              | CXorf51 (1/2 copies) | Arm | Arm |
| P8         | 14              | -47               | CXorf49 (1/2 copies) | Arm | Arm |
| P2         | 4               | 25.01             | Arm | Arm |
| P16        | 1               | -587              | LDOC1, SPANXC, SPANXA1 | Arm | Flanking |
| P11        | 1               | 9.93              | Spacer | Arm |
| P25        | 1               | 21.47             | Spacer | Arm |
| P5         | 1               | 17.89             | Arm | Arm |
| P22        | 1               | 8.46              | PNMA6 (2/2 copies) | Arm | Arm |
| P1         | 1               | 3.97              | Arm | Arm |

For arm genes, the number of gene copies removed by the deletion is indicated in parentheses. All other genes are unique genes from the palindrome spacer or flanking sequence. Deletion sizes marked as approximate (~) were estimated from changes in coverage depth; all others are exact sizes based on split reads spanning the deletion breakpoint. Individual genes within gene families CXorf51, CXorf49, and PNMA6 are listed in Supplemental Table S2.
(PNMA6: 2/2 copies; LDOC1: 1/1 copies). We speculate that deletions that remove one or more X-palindrome gene may result in mildly deleterious phenotypes, such as subtle defects in spermatogenesis; this hypothesis is consistent with our observation that insertions and deletions affecting X-palindrome genes are depleted between species.

NAHR is a common cause of palindrome rearrangements on the human X and Y Chromosomes (Lakich et al. 1993; Small et al. 1997; Aradhye et al. 2001; Lange et al. 2009; Scott et al. 2010). However, out of the nine X-palindrome spacer deletions we observed in males from the 1000 Genomes Project, only two displayed breakpoint homology consistent with NAHR. One possible explanation for non-NAHR-based rearrangements is replication errors: Duplication of a single-copy human X gene near P24 has been reported to result from Fork Stalling and Template Switching (FoSTeS), in which replication machinery repeatedly stalls within a low-copy repeat and switches templates, creating a rearrangement of the form duplication-inverted triplication-duplication (Carvalho et al. 2009, 2011). FoSTeS rearrangements lead to increased copy number, yet replication errors in different genomic contexts could lead to sequence loss. Indeed, intra-strand pairing and subsequent replication slippage were proposed to explain deletions within a 15-kb transgenic palindrome in mice, which underwent large asymmetric deletions around the center of palindrome symmetry within a single generation (Akgün et al. 1997). Although replication errors represent one plausible explanation for primate X-palindrome spacer deletions, future studies will be required to rule out other mechanisms.

Our work affirms the importance of high-quality genome assemblies for comparative genomics by revealing a wealth of conserved X palindromes with signatures of natural selection that were largely missing from existing chimpanzee and rhesus macaque X-Chromosome assemblies. In contrast to other genome assembly methods, SHIMS 3.0 enables the verification of palindromes and other genomic structures through the generation of multiple full-length nanopore reads from the same clone. In recent years, long-read technologies have also been incorporated into mammalian genome assemblies generated using a whole-genome shotgun assembly approach (see Gordon et al. 2016; Bickhart et al. 2017; Low et al. 2019; Miga et al. 2020, and others). Although long-read WGS assemblies offer substantial improvements over short-read WGS assemblies, we nevertheless found that the fraction of primate X palindromes represented accurately in two long-read WGS assemblies hovered around 50%, demonstrating that clone-based methods remain necessary to confidently resolve complex genomic structures. Indeed, a recent nanopore WGS assembly of the human genome achieved greater continuity of multiple full-length nanopore reads from the same clone. Nonetheless, long-read whole-genome shotgun assemblies and SHIMS 3.0 may be used in tandem to improve the representation of palindromes in mammalian genomes.

Our study used synteny between primate X Chromosomes to identify candidate regions in chimpanzee and rhesus macaque that were likely to contain palindromes; candidate regions were then targeted with SHIMS 3.0 to generate finished sequence. We propose that long-read WGS assemblies may serve a similar role: In our own comparison of primate X-Chromosome assemblies, we found that, whereas some palindromes were missing entirely from long-read WGS assemblies, the majority were present but incomplete. Long-read WGS assemblies may thus serve as a guide for identifying the positions of putative palindromes, which can then be finished using SHIMS 3.0 or other clone-based approaches. Future comparative analyses using high-resolution sequence will reveal whether conserved palindromes are a feature of other mammalian X Chromosomes and, if so, shed further light on the balance of structural instability and natural selection that govern their evolution. High-resolution X-Chromosome sequence for other great apes (gorilla and orangutan) may be particularly useful for probing the dynamics of X-palindrome spacer inversions and deletions. Finally, although our study was limited to X-Chromosome palindromes, a similar methodology could be applied to study palindromes on other chromosomes. To date, no human autosomes have been sequenced using a single-haplotype clone-based approach comparable to that used for the human and mouse X Chromosomes (Church et al. 2009; Mueller et al. 2013); SHIMS 3.0 could therefore also be used to search for new human autosomal palindromes (Warburton et al. 2004) and their orthologs in other species.

The conserved X-Chromosome palindromes that we describe here represent a substantially understudied class of genomic sequence. The technical challenges presented by palindromes go beyond the generation of accurate reference sequence: Many common bioinformatic tools, such as those used for quantifying gene expression or detecting mutations, automatically discard multimapping reads, rendering palindromes invisible in downstream analyses (Godfrey et al. 2020). The issue can be overcome by selection of tools like kallisto that probabilistically assign multimapping reads (Bray et al. 2016), yet this is not routinely done for large-scale genomic analyses. Given these challenges, it is not surprising that human X-palindrome gene families remain poorly characterized compared to other human X-linked genes (see also Mueller et al. 2013). Many human X-palindrome genes are classified as cancer-testis antigens (CTAs), genes defined by expression in the testis as well as in cancerous tumors, yet the mechanisms and significance of this phenomenon are not well understood (Simpson et al. 2005). Deletion or inversion of a single arm of a palindrome containing a testis-specific gene family in mouse yielded no observable phenotypes, leading to the suggestion that palindrome structures may primarily have benefits over longer evolutionary timescales, perhaps through purging deleterious mutations or fixing beneficial mutations through rapid gene conversion (Kruger et al. 2018). We propose that palindromes have a fundamentally different biology than unique sequence—a biology that does not readily align with our expectations or our standard methods of imputing function, including murine mouse models or association with Mendelian disease, both of which depend on observing a strong phenotype within a single generation. Technical advances that facilitate the study of X palindromes and other amplicons will be essential to illuminate their biology, with implications for X-Chromosome evolution as well as human health and disease.

Methods

Palindrome annotation

Candidate regions likely to contain human X-Chromosome palindrome were identified from the genomicSuperDups track in the UCSC Genome Browser (Kent et al. 2002) using the following criteria: Inverted repeats >8 kb in length and displaying >95% sequence identity, with <500 kb between arms. For each candidate
region, we divided the sequence into overlapping 100-bp windows, then aligned these windows back to the candidate region using Bowtie 2 (Langmead and Salzberg 2012), with settings to return up to 10 alignments with alignment scores $\geq -11$ (Teitz et al. 2018). We then created a bedGraph file for each candidate region in which the value for each position represents the number of times the window starting at that position aligns to that region and visualized the bedGraph file using the Integrative Genomics Viewer (IGV) (Robinson et al. 2011). Putative palindromes boundaries were annotated manually based on the start and end of long stretches of multimapping windows, and filtered for arms $>8$ kb and arm-to-arm identity $>99\%$. The same method was used to annotate palindromes within chimpanzee and rhesus macaque SHIMS 3.0 assemblies for regions orthologous to human X palindromes.

### Sequence alignments and dot plots

Square and triangular dot plots were generated using custom Perl code (http://pagelabsupplement.wi.mit.edu/fast_dot_plot.pl). Unless otherwise noted, sequence alignments were performed using ClustalW with default parameters (Thompson et al. 1994). To identify and exclude regions of poor alignment, ClustalW sequence alignments were scanned using a sliding 100-bp window and filtered to exclude windows with fewer than 60 matches between species, using custom Python code.

### Human gene annotation

GENCODE 34 gene annotations for the human X Chromosome were downloaded from Ensembl using the BioMart package in R (R Core Team 2020). Annotations were filtered for protein-coding genes only, and the APPRIS principal transcript was selected for each gene. If there were multiple principal transcripts, the longest principal transcript was selected, and if there were multiple principal transcripts of equal length, the longest principal transcript with the highest transcript support level (TSL) was selected. There were three exceptions as follows:

1. There were two protein-coding genes with the same Ensembl gene name: ENSG00000158427 and ENSG00000269226, both named TMSB15B. We refer to them here as TMSB15BA and TMSB15BB, respectively.

2. In two cases, the gene name in GENCODE 34 differed from the current approved nomenclature from HGNC; in both cases, we use the gene name approved by HGNC. The gene name for ENSG00000224931 in GENCODE 34 is AC520101.1, but there is no current approved gene name from HGNC; we therefore refer to it as novel gene ENSG00000224931.

3. The palindrome arm genes PNMA6B and novel gene ENSG00000224931 were included despite being annotated as pseudogenes, because each gene was annotated as having a protein-coding paralog in the other arm (see Supplemental Note S1).

4. The principal transcript for palindrome arm gene TCP11X2 encoded a different protein than the principal transcript for its paralog TCP11X1. For consistency, the isoform encoding the longer protein was selected as the principal transcript for both; transcript ENST00000642911 (marked “alternative2”) was therefore used for TCP11X2.

### Human gene expression

Gene expression was calculated for a subset of samples from the GTEx project (v8) as follows. For each tissue subtype in GTEx, 5–10 of the highest quality samples were selected based on a combination of RNA integrity (RIN), mapped-read library size, and intronic read mapping rate. BAM files containing all reads (mapped and unmapped) from these samples were accessed through Terra (https://app.terra.bio) and used to generate FASTQ files. Transcript expression levels in TPM were estimated using kallisto with sequence-bias correction (−bias) using GENCODE 34 gene annotations, then summed to obtain gene expression levels. Results were filtered to include protein-coding genes only and TPM values were renormalized to 1 million for each sample. For human X-palindrome arm gene families, expression levels for both arm genes were averaged to return gene family expression levels. To analyze expression of human X-palindrome gene families in spermatogenesis, we downloaded publicly available NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) files from Jan et al. (2017) (accession number SRP069329) and analyzed them with kallisto as described above.

### Transposable elements

We analyzed transposable element density using RepeatMasker (https://www.repeatmasker.org) with default settings.

### Clone selection and sequencing

All chimpanzee clones selected for sequencing were from BAC library CH251 (https://bac pacresources.org), which derives from a single male individual (“Clint”) used in initial sequencing of the chimpanzee genome (The Chimpanzee Sequencing and Analysis Consortium 2005). All rhesus macaque clones selected for sequencing were from BAC library CH250, which derives from a single male individual of Indian origin (https://bacpacresources.org). Sequencing was performed using the SHIMS 3.0 protocol (Bellett et al. 2020). Regions covered by one or more nanopore reads, but no Illumina reads, were marked as “problem regions” and excluded from downstream analysis; these regions represented $<1\%$ of all sequence generated for this project. Our assemblies also include sequence from seven chimpanzee clones previously sequenced and deposited in the NCBI GenBank database (https://www.ncbi.nlm.nih.gov/genbank/); these clones each contained part or all of a palindrome arm but did not contain internal repeats, making them suitable for assembly without long reads (Supplemental Table S9).

### Comparison to existing X-Chromosome assemblies

X-Chromosome assemblies for Pan_tro_3.0 (CM000336.3), Mmul_8.0.1 (CM002997.3), and Mmul_10 (CM014356.1) were downloaded from Ensembl (www.ensembl.org). The X-Chromosome assembly for Clint_PTRv2 (CM0009261.2) was downloaded from NCBI (https://www.ncbi.nlm.nih.gov/). Regions orthologous to each chimpanzee and rhesus macaque palindrome, as identified in SHIMS 3.0 assemblies, were extracted from each X-Chromosome reference assembly using custom Python code. We generated triangular dot plots for each extracted region and square dot plots comparing each extracted region to the orthologous SHIMS 3.0 assembly. Categorizations were made as follows: (1) “Missing” = no palindrome; (2) “Incomplete” = a palindrome was partially present but misassembled; and (3) “Accurate” = palindrome arms and spacer aligned fully to the orthologous SHIMS 3.0 assembly.

### Primate gene annotation

Pitrate gene annotations were performed manually using alignment of human exons and alignment of testis RNA-seq reads for
Palindromes are conserved on primate X Chromosomes

Primate gene expression
The latest transcriptomes for chimpanzee and rhesus macaque were downloaded from Ensembl (Pan_troglodytes.Pan_tro_3.0.cdna.all.fa and Macaca_mulatta.Mmul_10.cdna.all.fa, respectively) and merged with newly annotated palindrome arm and spacer genes from SHIMS 3.0 assemblies. To prevent redundancy between our transcripts and transcripts representing the same genes that were already present in existing transcriptomes, we used BLAST to identify and remove existing transcripts that aligned to newly annotated genes over >50% of their length and with >95% sequence ID. Gene expression was calculated using RNA-seq reads from the following publicly available data sets containing at least five different tissues, including testis: Chimpanzee (Brawand et al. 2011); Rhesus macaque (Merkin et al. 2012). Transcript expression levels were calculated using kallisto with sequence-bias correction (~bias), and summed to gene expression levels. To enable comparison of expression between conserved human and primate gene families, all primate X-palindrome genes were grouped based on their closest human X-palindrome gene family, and gene family expression levels were calculated accordingly.

Definition of orthologous palindromes
For each palindrome identified in chimpanzee or rhesus macaque SHIMS 3.0 assemblies, we generated alignments between Arm 1 of the nonhuman primate and Arm 1 of the putative human ortholog. Orthologous palindromes were required to meet two criteria, designed to establish an unambiguous common origin. First, at least 20% of the nonhuman primate palindrome arm was required to align to the putative human ortholog. Second, the alignable portion of the human palindrome arm was analyzed with BLAST against the complete nonhuman primate region (including palindrome arms, spacer, and flanking sequence) using default parameters. More than 90% of positions in high-quality BLAST hits (>1 kb, 95% for chimpanzee vs. human; >1 kb, 90% for rhesus macaque vs. human) were required to map to the palindrome arms.

Calculation of divergence
Divergence was calculated by generating pairwise alignments using ClustalW, then calculating p-distance with MEGA X (Kumar et al. 2018). For alignment of arms between species, we generated pairwise alignments using Arm 1 from each species.

Calculation of fraction of orthologous sequence
Pairwise alignments between species were generated as described above. The fraction of orthologous sequence was calculated as (total bases in unfiltered alignment windows)/(total bases in starting sequence), after excluding bases from problem areas.

Analysis of Online Mendelian Inheritance in Man phenotypes
We downloaded the genemap2.txt file from the OMIM database (https://www.omim.org/) (McKusick-Nathans Institute of Genetic Medicine 2020). We filtered for phenotypes linked to a single X-linked gene using custom Python code, and calculated the fraction of all protein-coding X-linked genes with an OMIM phenotype, relative to the fraction of X-palindrome arm genes and X-palindrome spacer genes with an OMIM phenotype.

Calculation of \( \frac{d_s}{d_S} \)
Alignments of coding sequence from X-palindrome arm and spacer genes conserved between human, chimpanzee, and rhesus macaque were performed using default parameters from ClustalW. \( d_s/d_S \) values were calculated using the basic model in PAML (model = 0, NSsites = 0) (Yang 2007). To test the significance of calculated \( d_s/d_S \) values, we compared the likelihood of calculated values against a model where \( d_s/d_S \) was fixed at one. To test for positive selection, we compared the likelihood of model M1a (neutral evolution) versus model M2a (positive selection at one or more sites). We defined significance for both comparisons using the chi-squared distribution and appropriate degrees of freedom.

Depletion of indels within protein-coding genes
Sequence from one species was broken into overlapping k-mers with step size = 1 and aligned to orthologous sequence from the other species using Bowtie 2 (Langmead and Salzberg 2012), with settings to return up to 10 alignments with alignment scores ≥11. k-mer size was either 100 (human-chimpanzee comparisons) or 40 (human-rhesus macaque comparisons). Indels were defined as stretches of at least 1 kb from one species that had no aligned k-mers from the other species.

1000 Genomes data analysis
We analyzed whole-genome sequencing data from 1225 males from the 1000 Genomes Project (The 1000 Genomes Project Consortium 2015). Data selection, sequence alignment, GC bias correction, and repeat masking were performed as described in Teitz et al. (2018). We calculated average read depth for palindrome arms, spacers, and flanking sequence, normalized to a 1-Mb region of the X Chromosome without palindromes, using custom Python scripts. We filtered for males whose X Chromosome normalization region had an average read depth ≥2, restricting our downstream analysis to 944 males. For small spacers (<3 kb), we expanded the area over which we calculated average read depth symmetrically into the inner palindrome arm until reaching 3 kb.

To identify candidate spacer deletions, we initially filtered for palindrome spacers with a normalized read depth below 0.25. After visualizing histograms of spacer depth across 944 males for each palindrome, we noticed a second peak centered around 0.25 for P17. To include all candidate P17 spacer deletions, we therefore raised our initial filtering threshold for P17 to 0.5. Read depths for all candidate spacer deletions were viewed using IGV. Candidate deletions that did not have a clear reduction in read depth in the spacer were excluded; all others were included in Table 2.
Identification of split reads spanning deletion breakpoints
Forward and reverse reads from individuals with deletions were aligned separately to the region of the suspected deletion with settings to return up to 10 matches of minimum alignment score ≥11. We identified read pairs in which one read aligned and the other did not and returned the sequence of the read that did not align. In the case of P17 deletions, we inspected unaligned reads by eye from four males to identify reads spanning the breakpoint. Finding that all of these males had the same breakpoint, we then used this breakpoint (± 10 base pairs on each side) to screen unmapped reads from all other males with suspected P17 deletions. For individuals where a breakpoint read could not be found using the primary 1000 Genomes data set reads (1000genomes.sequence.index), we used a deeper 1000 Genomes data set (1000G_2504_high_coverage.sequence.index).

PCR verification of human palindrome spacer deletions
Patient genomic DNAs were purchased from Corell Cell Repositories (HGO1872, HGO2070, HGO2398, HGO2687, HG03295, HG04015, HG04219, NA11919, NA18645, NA19086, NA19562, NA20351, NA20897, NA20905, NA21116, NA21117, NA21133). DNAs were tested for the presence or absence of palindrome spacers using primer pairs described in Supplemental Table S10. PCR was performed using 50 ng of DNA as template in a total volume of 20 μL (10 mM Tris–HCl [pH 9], 1.5 mM MgCl2, 50 mM KCl, 0.1% Triton X-100, 0.2 mM dNTPs, 0.5 μM primers, 0.5 U Taq polymerase). PCR cycling conditions for all primers were as follows: 94°C (30 sec), 61°C (30 sec), 72°C (1 min) for 35 cycles. Long range PCR was performed using Advantage 2 Polymerase following the manufacturer’s protocol (Clontech Laboratories).

Association of CXorf51 and CXorf49 gene family deletions with azoospermia
Palindrome spacer deletions removing one copy of CXorf51 or CXorf49 in dbGaP data set phs0001023 were detected based on reduced coverage depth, as described above for 1000 Genomes, with modifications as follows. Data set phs0001023 was generated using target-based capture sequencing rather than whole-genome shotgun sequencing, making it inappropriate for de novo deletion discovery. We therefore selected coordinates for detection of CXorf51 and CXorf49 spacer deletions based on three criteria: (1) coordinates overlap part or all of the deletion identified from the 1000 Genomes analysis; (2) coordinates contain at least two target-ed probes; and (3) coverage depth within coordinates is predicted to decrease by >50% when the deletion identified from 1000 Genomes is present.

Association of CXorf51 and CXorf49 gene family deletions with oligoazoospermia
We analyzed 562 DNA samples from oligoazoospermic men previously collected by our lab. We excluded samples from men with Y-Chromosome deletions, varicocele, undescended testicles, or other known risk factors for oligoazoospermia. Palindrome spacer deletions removing one copy of CXorf51 or CXorf49 were detected using the same primers and PCR conditions used for verification of deletions from 1000 Genomes (see above). Each DNA sample was tested using one set of primers expected to yield no product in samples with the deletion (P17 inner arm, P8 inner arm) and one set of primers expected to yield a specific product in samples with the deletion (P17 breakpoint, P8 breakpoint) (Supplemental Table S10).

Human data
These studies were approved by the Massachusetts Institute of Technology's Committee on the Use of Humans as Experimental Subjects. Informed consent was obtained from all subjects.

Data access
BAC sequences generated in this study have been submitted to the NCBI Genbank database (https://www.ncbi.nlm.nih.gov/genbank/) under the accession numbers listed in Supplemental Table S11. Codes for replicating these analyses are included in Supplemental Code as well as on Github (https://github.com/ejackson054/primate-X-palindromes).

Competing interest statement
The authors declare no competing interests.

Acknowledgments
We thank A.K. San Roman and J.L. Mueller for comments on the manuscript, and A.K. Godfrey and other current and former Page lab members for valuable advice throughout the project. We thank S. Silber, R. Oates, M.C. Summers, V. Cardone, P. Patrizio, B. Gilbert, W.A. Hogge, C. Brunnning, M. Jannethor, B. Shapiro, K. Monaghan, J.C. Petrozza, L. Omar-Ganes, and J. Roberson for patient samples. This work was supported by the Howard Hughes Medical Institute, the Whitehead Institute, and generous gifts from Brit and Alexander d’Arbeloff, Arthur W. and Carol Tobin Brill, and Matthew Brill.

Author contributions: E.K.J., D.W.B, H.S., J.F.H., and D.C.P. designed the study. T.J.C. generated nanopore and Illumina sequencing reads. E.K.J. and D.W.B. assembled and finished clones. T.P. performed PCR experiments. E.K.J. performed computational analyses with assistance from D.W.B. and H.S. E.K.J. and D.C.P. wrote the manuscript.

References
The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. Nature 526: 68–74. doi:10.1038/nature15393
Alkan C, Zahn J, Baumés S, Brown G, Liang F, Romanienko PJ, Lewis S, Jasin M. 1997. Palindrome resolution and recombination in the mammalian germ line. Mol Cell Biol 17: 5559–5570. doi:10.1128/MCB.17.9.5559
Alkan C, Sahjadian S, Eichler EE. 2011. Limitations of next-generation genome sequence assembly. Nat Methods 8: 61–65. doi:10.1038/nmeth.1527
Anisimova M, Bielawski JP, Yang Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol 18: 1585–1592. doi:10.1093/oxfordjournals.molbev.a003945
Aradhya S, Bardaro T, Galgczy P, Yamagata T, Esposito T, Patlan H, Ciccodicola A, Munnich A, Kenwrick S, Platzer M, et al. 2001. Multiple pathogenic and benign genomic rearrangements occur at a 35 kb duplication involving the NEMO and LAGE2 genes. Hum Mol Genet 10: 2557–2567. doi:10.1038/hmg.10.22.2557
Bellott DW, Cho TJ, Hughes JF, Skalesky H, Page DC. 2018. Cost-effective high-throughput single-haplotype iterative mapping and sequencing for complex genomic structures. Nat Protoc 13: 787–809. doi:10.1038/nprot.2018.019
Bellott DW, Cho T-J, Jackson E, Skalesky H, Hughes JF, Page DC. 2020. Highly efficient single-haplotype iterative mapping and sequencing using ultra-long nanopore reads. bioRxiv doi:10.1101/2020.09.18.303735
Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, Lee J, Lam ET, Liachko I, Sullivan ST, et al. 2017. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nat Genet 49: 645–650. doi:10.1038/ng.3802
Bione S, Maestrini E, Rivella S, Mancini M, Regis S, Romeo G, Toniolo D. 1994. Identification of a novel X-linked gene responsible for Emery-Dreifuss muscular dystrophy. Nat Genet 8: 323–327. doi:10.1038/ng1294-323
human spermatogenesis. Development 144: 3659–3673. doi:10.1242/dev.152413

Jensen-Seaman MI, Furey TS, Page A, Yu L, Roskin KM, Chen C-F, Thomas MA, Haussler D, Jacob HJ. 2004. Comparative recombination rates in the rat, mouse, and human genomes. Genome Res 14: 528–531. doi:10.1101/gr.197030

Kent WJ. 2002. BLAT—the BLAST-like alignment tool. Genome Res 12: 656–664. doi:10.1101/gr.22920

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. Genome Res 12: 996–1006. doi:10.1101/gr.121902

Kronenbergn ZN, Veldes IT, Gordon D, Murali S, Cantisbiers L, Meyerson OS, Unschuld GJ, Nelson EC, Chaisson MJ, Dougherty ML, et al. 2018. High-resolution comparative analysis of great ape genomes. Science 360: eaar6343. doi:10.1126/science.aar6343

Kruger AN, Ellison Q, Brogly MA, Gerlinger ER, Mueller JL. 2018. Male mice with large inversions or deletions of X-chromosome patellar anomalies are fertile and express their associated genes during meiosis. Sci Rep 8: 8985. doi:10.1038/s41598-018-27360-x

Kuderna LFK, Tomlinson C, Hillier LW, Tran A, Fiddes IT, Armstrong J, Laasooymi H, Gordon D, Huddleston J, Perez RG, et al. 2017. A 3-way hybrid approach to generate a new high-quality chimpanzee reference genome (Pan tro tro). Gigascience 6:1–6. doi:10.1093/gigascience/gix098

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TreeIME: a resource for timelines, timetrees, and divergence times. Mol Biol Evol 34: 1812–1819. doi:10.1093/molbev/msx116

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. Mol Biol Evol 35: 1547–1549. doi:10.1093/molbev/msy021

Kuroda-Kawaguchi T, Skaletsky H, Brown LG, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Silber S, Oates R, Rozen S, et al. 2007. The AYZ region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. Nat Genet 39: 279–286. doi:10.1038/ng10757

Lahn B, Page DC. 1999. Four evolutionary strata on the human X chromosome. Science 286: 964–967. doi:10.1126/science.286.5441.964

Lachke D, Kaczazian HH, Antonarakis SE, Gitschier J. 1993. Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. Nat Genet 5: 236–241. doi:10.1038/ng1193-236

Lange J, Skaletsky H, van Daalen SKM, Embry SL, Korer CM, Brown LG, Oates RD, Silber S, Repping S, Page DC. 2009. Isolecinic Y chromosomes and sex disorders as byproducts of homologous recombination that maintains palindromes. Cell 138: 885–889. doi:10.1016/j.cell.2009.07.042

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357–359. doi:10.1038/nmeth.1923

Low WY, Tauerle R, Rickhart DM, Rosen BD, Kingan SB, Swale T, Thibaud-Nissen F, Murphy TD, Young R, Lefere L, et al. 2019. Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity. Nat Commun 10: 2680. doi:10.1038/s41467-018-08260-0

Mckusick-Nathans Institute of Genetic Medicine. 2010. Online Mendelian Inheritance in Man, OMIM®. 2010. Johns Hopkins University. Baltimore, MD.

Merkin JR, Russell C, Pen Burge CB. 2012. Evolutionary dynamics of genes and isoform regulation in mammalian tissues. Science 338: 1593–1599. doi:10.1126/science.1228186

Miga KH, Koren S, Rihe A, Tyson CE, Gishrnan A, Rizkazdke A, Brooks S, kose H, porubsky D, Logsdon GA, et al. 2020. Telomere-to-telomere assembly of a complete human X chromosome. Nat Genet 52: 79–84. doi:10.1038/s41588-020-02547

Mueller JL, Mahadeviah SK, Park P, Warrburn PE, Page DC, Turner JMA. 2008. The mouse X chromosome is enriched for multicopy testis genes showing post-meiotic expression. Nat Genet 40: 794–799. doi:10.1038/ng.126

Mueller JL, Skaletsky H, Brown LG, Zalhuhl S, Rock S, Graves T, Auger K, Warren WC, Wilson RK, Page DC. 2013. Independent specialization of the human and mouse X chromosomes for male germ cell functions. Genome Res 23: 1098–1107. doi:10.1101/gr.135124.112

Ohto S. 1967. Sex chromosomes and sex-limited genes. Springer, Berlin.

Porubsky D, Sanders AD, Hops W, Siheh P, Solomon A, Li R, Mercuri L, Sorensen M, Murali SC, Gordon D, et al. 2020. Recurrent inversion tugging and great ape genome evolution. Nat Genet 52: 849–883. doi:10.1038/s41588-020-00646-x

R Core Team. 2020. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/

Rice WR. 1984. Sexual chromosomes and the evolution of sexual dimorphism. Evolution (N Y) 38: 735–742. doi:10.1111/j.1558-5646.1984.tb00346.x

Robinson JT, Thrallsdottir H, Wincker W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. Nat Biotechnol 29: 24–26. doi:10.1038/nbt.1754
Ross MT, Graffham DV, Coffey AJ, Scherer S, McLay K, Muzny D, Platzer M, Howell GR, Burrows C, Bird CP, et al. 2005. The DNA sequence of the human X chromosome. *Nature* **434**: 325–337. doi:10.1038/nature03440

Rozen S, Skalskya H, Marszałek JD, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Page DC. 2003. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* **423**: 873–876. doi:10.1038/nature01723

Scott SA, Cohen N, Brandt T, Warburton PE, Edelmann L. 2010. Large inverted repeats within Xp11.2 are present at the breakpoints of isodicentric X chromosomes in Turner syndrome. *Hum Mol Genet* **19**: 3383–3393. doi:10.1093/hmg/ddq250

Simpson AJ, Caballero OL, Jungbluth A, Chen Y-T, Old LJ. 2005. Cancer/testis antigens, gametogenesis, and cancer. *Nat Rev Cancer* **5**: 615–625. doi:10.1038/nrc1669

Skalskya H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier LG, Repping S, Pyntikova T, Ali J, Bieri T, et al. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**: 825–837. doi:10.1038/nature01722

Small K, Iber J, Warren ST. 1997. Emerin deletion reveals a common X-chromosome inversion mediated by inverted repeats. *Nat Genet* **16**: 96–99. doi:10.1038/ng0597-96

Soh YQS, Alfoldi J, Pyntikova T, Brown LG, Minx PJ, Fulton RS, Kremitski C, Koutseva N, Mueller JL, Rozen S, et al. 2014. Sequencing the mouse Y chromosome reveals convergent gene acquisition and amplification on both sex chromosomes. *Cell* **159**: 800–813. doi:10.1016/j.cell.2014.09.052

Stankiewicz P, Lupski JR. 2010. Structural variation in the human genome and its role in disease. *Annu Rev Med* **61**: 437–455. doi:10.1146/annurev-med-100708-204735

Stevenson BJ, Iseli C, Panji S, Zahn-Zabal M, Hide W, Old LJ, Simpson AJ, Jongeneel CV. 2007. Rapid evolution of cancer/testis genes on the X chromosome. *BMC Genomics* **8**: 129. doi:10.1186/1471-2164-8-129

Swanepoel CM, Gerlinger ER, Mueller JL. 2020. Large X-linked palindromes undergo arm-to-arm gene conversion across Mus lineages. *Mol Biol Evol* **37**: 1979–1985. doi:10.1093/molbev/msaa059

Tett LS, Pyntikova T, Skalskya H, Page DC. 2018. Selection has countered high mutability to preserve the ancestral copy number of Y chromosome amplexins in diverse human lineages. *Am J Hum Genet* **103**: 261–275. doi:10.1016/j.ajhg.2018.07.007

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**: 4673–4680. doi:10.1093/nat/22.22.4673

Vogt PH, Edelmann A, Kirsch S, Henegarina O, Hirschmann P, Kiesewetter F, Köhn FM, Schill WB, Farah S, Ramos C, et al. 1996. Human Y chromosome azoospermia factors (AZF) mapped to different subregions in Yq11. *Hum Mol Genet* **5**: 933–943. doi:10.1093/hmg/5.7.933

Warburton PE, Giordano J, Cheung F, Gelfand Y, Benson G. 2004. Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. *Genome Res* **14**: 1861–1869. doi:10.1101/gr.2542904

Warren WC, Harris RA, Haukness M, Fidites IT, Murali SC, Fernandes J, Dishuck PC, Storer JM, Raveendran M, Hillier LW, et al. 2020. Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. *Science* **370**: eabc6617. doi:10.1126/science.eabc6617

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586–1591. doi:10.1093/molbev/msm088

Zimin AV, Cornish AS, Maudhoo MD, Gibbs RM, Zhang X, Pandey S, Meehan DT, Wipfler K, Bosinger SE, Johnson ZP, et al. 2014. A new rhesus macaque assembly and annotation for next-generation sequencing analyses. *Biol Direct* **9**: 20. doi:10.1186/s13064-015-0020-2

Received December 29, 2020; accepted in revised form May 17, 2021.