Evolution of large palindromes on the primate X chromosome

by

Emily Katherine Jackson

B.A., Biology
Amherst College, 2013

Submitted to the Department of Biology
in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2021

Signature of Author ....Emily K. Jackson
Department of Biology
May 21, 2021

Certified by ....David C. Page
Professor of Biology
Member, Whitehead Institute
Investigator, Howard Hughes Medical Institute
Thesis Supervisor

Accepted by .... Amy Keating
Professor of Biology and Biological Engineering
Co-Director, Biology Graduate Committee

Evolution of large palindromes on the primate X chromosome

by

Emily Katherine Jackson

Submitted to the Department of Biology on May 21, 2021
in partial fulfillment of the requirements
for the Degree of Doctor of Philosophy in Biology

ABSTRACT

Human sex chromosomes are enriched for complex genomic architecture, including massive palindromes with arms that can exceed 1 Mb in length and arm-to-arm sequence identity higher than 99%. Palindrome arms harbor protein-coding genes with testis-biased expression, suggesting roles in male fertility. However, palindromes are under-represented in non-human reference genomes due to technical challenges associated with genomic repeats, limiting our understanding of palindrome origins and evolution.

In this thesis, we used specialized methods to investigate the evolution of X-chromosome palindromes in primates. We used a clone-based sequencing approach that incorporates ultralong nanopore reads to generate accurate reference sequence for regions orthologous to human X palindromes in two non-human primates, the chimpanzee and the rhesus macaque. Twelve human X palindromes have conserved orthologs in both species, demonstrating a common origin at least 25 million years ago. The majority of these palindromes were missing or misassembled in existing reference genomes for these species. Comparative analyses demonstrate that natural selection preserves X-palindrome gene families, despite limited functional characterization of these genes in humans. Unexpectedly, structural comparisons of conserved palindromes between species revealed frequent rearrangements around the center of palindrome symmetry; this instability persists among human X chromosomes, which are enriched for deletions within the spacer that separates palindrome arms.

Sequence identity between palindrome arms is maintained by high rates of intra-chromosomal gene conversion, which led us to hypothesize that palindromes may be subject to amplified effects from GC-biased gene conversion. Among twelve conserved primate X palindromes, we find that palindrome arms are significantly more GC-rich than flanking sequence, and that GC content in primate X-palindrome arms is increasing over time. Evolutionary simulations reveal that nucleotide replacement patterns between species are consistent with a magnitude of GC bias in gene conversion of around 70%, consistent with previous estimates derived from analyses of human meiosis. Altogether, the work presented in this thesis demonstrates an unexpectedly deep evolutionary history of primate X palindromes that is shaped by a complex mixture of natural selection, localized structural instability, and GC-biased gene conversion.

Thesis Supervisor: David C. Page
Title: Professor of Biology

ACKNOWLEDGEMENTS

The research described in this dissertation would not have been possible without the help and support of my community at MIT and beyond. I am deeply grateful to my advisor, David Page, for giving me the opportunity to learn in his lab—and in particular, for allowing me the chance to tackle computational projects that expanded my horizons as a scientist. In addition to his carefully thought-out approach to research, David cultivates a lab environment that is supportive and collaborative, with an emphasis on clear thinking and communication. The Page lab has been a wonderful place to work, and I have learned a great deal from David's example.

My experience in the lab would not have been the same without all of my colleagues over the years. I want to thank Winston and Helen in particular for their mentorship, and for each guiding me out of countless holes that I have dug myself—whether through giving advice for a new analysis, providing feedback on manuscript drafts, directing me to the right scientific paper, or simply being willing to talk through an issue that I have been staring at for too long. I will always be grateful for the time they invested in me, and I hope to be able to pay it forward to others in the future. Thank you to Gabe, for your jokes, your patience, and for making my day every time you generated another amazing run of nanopore reads. Thank you to Tatyana for sharing your time, expertise, and energy to help with my PCR experiments. Thank you to Susan and Jorge for all the work you do to keep the lab running smoothly, and to Jenn for patiently answering every question I've ever asked you. Thank you to everyone for the encouragement and laughter shared over lunch breaks, cookie hours, retreats, ski trips, and more.

I would like to thank my thesis committee members, Chris Burge and Jackie Lees, for their thoughtful feedback and support as I developed these projects. I would also like to thank Karen Miga for sharing her time and expertise as my external thesis committee member.

Thank you to all the friends and family who have supported me over the past few years, providing balance and distractions when I've needed them most.

Most of all, thank you to my three "house people." To my mom, for providing endless love, support, encouragement, and muffins, and for reminding me through her own example what's important in life. To my big sister, Molly, my littlest and fiercest defender, for always knowing what I need. And to my husband, Jeremy, for being there every step of the way, and making every one of them better. I love you all more than I can say.

TABLE OF CONTENTS

CHAPTER 1:

Introduction

I thank Helen Skaletsky, Jennifer Hughes, and David Page for comments on this chapter.

Mammalian sex is determined by karyotype: Females have two X chromosomes, and males have one X chromosome and one Y chromosome. The X and Y chromosomes can be easily distinguished under a microscope based on their relative sizes (the X chromosome is much larger than the Y chromosome). Visible differentiation between the X and Y chromosomes results from their distinct evolutionary trajectories over the past 300 million years, since they first arose from a pair of ordinary autosomes (Ohno 1967, Lahn and Page 1999). The Y chromosome lacks a partner for meiotic recombination and therefore has suffered dramatic genetic decay, resulting in its diminutive current size, while the X chromosome maintains the majority of its ancestral gene content across a wide range of mammalian species.

Over the past twenty years, a second theme in the evolution of mammalian sex chromosomes has emerged. In addition to the loss or maintenance of ancestral gene content, the X and Y chromosomes have each acquired and amplified gene families that are expressed solely or predominantly in the testis, suggesting specialized roles in male reproduction (Skaletsky et al. 2003, Mueller et al. 2013). Amplified gene families are often found within massive palindromes, with lengths that can exceed 1 Mb and arm-to-arm identities up to 99.99%. While the coherent expression patterns of palindrome gene families suggest a unique relationship between genomic structure and function, the study of sex-chromosome palindromes has been inhibited by technical challenges. Fewer than ten mammalian X and Y chromosomes have been sequenced using methods capable of resolving palindromes, including only two X chromosomes: those of human (Ross et al. 2005, Mueller et al. 2013) and mouse (Church et al. 2009). Deletion of a subset of human Y-palindrome genes has been directly linked to spermatogenic defects (Vogt et al. 1996), yet the functions of nearly all X-palindrome genes remain unknown (Mueller et al. 2013). In this thesis, we used specialized methods to determine whether massive palindromes are present on other primate X chromosomes, and if so, to investigate how X palindromes and their associated gene families evolve over time.

In this introduction, I will first provide context on the unique evolutionary history of the X chromosome, and how it can illuminate the present-day gene content of mammalian X chromosomes. I will review previous literature describing palindromes on the human and mouse X chromosomes, while also discussing evolutionary insights and lingering questions from studies of mammalian Y-chromosome palindromes. Importantly, I will also introduce several challenges to the study of sex-chromosome palindromes, including the generation of accurate reference sequence and the design of downstream bioinformatic analyses that deal appropriately with short next-generation sequencing reads that originate from near-identical palindrome arms. The evolutionary insights that result from addressing these challenges—including new evidence that natural selection preserves X-palindrome gene families, the discovery that human X palindromes are susceptible to deletions around the center of palindrome symmetry, and an evolutionary history of GC-biased gene conversion in primate X-palindrome arms—will be the topics of Chapters 2 and 3.

EVOLUTION OF MAMMALIAN SEX CHROMOSOMES FROM ANCESTRAL AUTOSOMES

Early differentiation of the mammalian X and Y chromosomes

The mammalian X and Y chromosomes began as an ordinary pair of autosomes. Evolution of a pair of sex chromosomes from ancestral autosomes is not unique to mammals, but has occurred independently many times in diverse taxa including birds (Fridolfsson et al. 1998, Nanda et al. 1999, Bellott et al. 2010), snakes (Matsubara et al. 2006, Bellott and Page 2021), fish (Nanda et al. 1990), flies (Kaiser et al. 2011, Zhou and Bachtrog 2012), and plants (Atanassov 2001, Nicholas et al. 2004). Sex chromosome differentiation begins with the acquisition of a sex-determining gene by one homolog (Ohno 1967) (Figure 1.1); for mammals, the sex-determining gene is the Y-chromosome gene *SRY*, which is both necessary and sufficient to promote male gonadal development (Koopman et al. 1991). For further differentiation between the proto-X and proto-Y chromosomes to occur, recombination between the homologs must be suppressed, typically through one or more inversions of sequence on the proto-Y chromosome. The mammalian Y chromosome has undergone at least four sequential inversions that suppress recombination with the X chromosome (Lahn and Page 1999), so that today recombination only occurs in small regions at the tip of each chromosome known as the pseudo-autosomal regions (i.e. PAR1

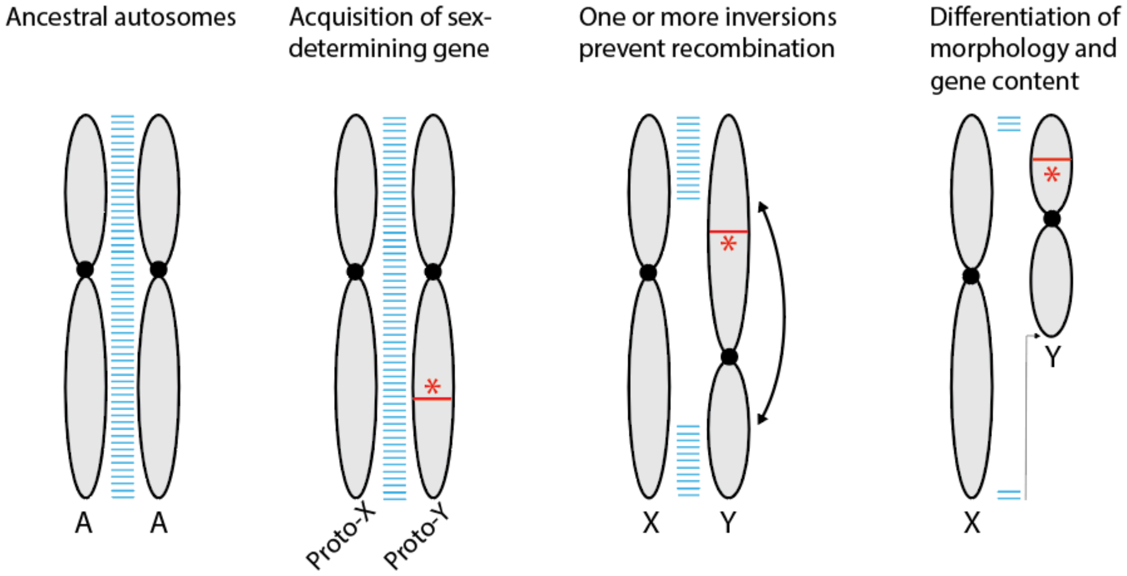

Ancestral autosomes
Acquisition of sex-determining gene
One or more inversions prevent recombination
Differentiation of morphology and gene content
A
A
Proto-X
Proto-Y
*
X
Y
*
X
Y
*

Figure 1.1 Evolution of sex chromosomes from ancestral autosomes. Black: Centromere. Red line with asterisk: Sex-determining gene. Blue lines: Regions of recombination. Sex chromosomes arise from an identical pair of autosomes when one homolog acquires a sex-determining gene. One or more inversions prevent recombination between portions of the proto-X chromosome and proto-Y chromosome, allowing differentiation of sequence and gene content. The X chromosome continues to undergo recombination along its full length during female meiosis (not shown), allowing it to retain the majority of its ancestral sequence and gene content. The X and Y chromosomes continue to recombine during male meiosis within specialized regions at the tips of each chromosome known as the pseudo-autosomal regions (PARs). The remainder of the Y chromosome undergoes progressive genetic decay due to its inability to purge deleterious mutations through recombination.

and PAR2). Ongoing recombination in the pseudo-autosomal regions is essential for the correct pairing of the X and Y chromosomes during male meiosis (Burgoyne et al. 1992, Mohandas et al. 1992), and ensures that the PARs remain identical even while the remainder of the X and Y chromosomes diverge. Future mentions of the Y chromosome in this thesis can be assumed to refer to the male-specific region of the Y chromosome (MSY), meaning the regions of the Y chromosome outside of the PARs, unless otherwise stated. Over time, the sequence and gene content of the Y chromosome has decayed, while ancestral gene content is largely preserved on the X chromosome (Ohno 1967, Bellott et al. 2010). The reasons for the divergent evolutionary trajectories taken by the non-PAR regions of the mammalian X and Y chromosomes are described in detail in the following two sections.

Degeneration of the Y chromosome to a handful of dosage-sensitive ancestral genes

The identification of the mammalian Y chromosome in 1921 was enabled in part by its distinctive small size relative to the X chromosome (Painter 1921). Although the Y chromosome is male-specific, small size is also a feature of the female-specific snake and avian W chromosomes (by convention, sex chromosome systems with a female-specific sex chromosome are described using the nomenclature ZW, where W represents the female-specific sex chromosome and Z represents the sex-shared chromosome). It can thus be inferred that Y chromosome decay does not result from male specificity, but rather from the general consequences of lacking a partner during meiotic recombination.

Several non-mutually exclusive models have been proposed to explain the decay of the Y chromosome following the loss of meiotic recombination (reviewed in Charlesworth and Charlesworth 2000). Hermann Muller noted in 1964 that a Y chromosome can never become “less mutated”—that is, Y chromosomes cannot reshuffle deleterious mutations during meiosis, which would otherwise generate some Y chromosomes with an elevated mutation load and others with a reduced load. In the absence of a mechanism to generate Y chromosomes with fewer mutations, a population of Y chromosomes can only accumulate more deleterious mutations over time, as the least mutated Y chromosome is repeatedly lost through genetic drift (Muller 1964). This phenomenon, known as Muller’s Ratchet, is compounded by the reduced effective population size of Y chromosomes: Y chromosomes are only one-quarter as common as autosomes, further increasing their susceptibility to genetic drift (Nei 1970). Finally, the complete linkage of Y-chromosome genes ensures that natural selection must act on the Y chromosome as a whole, which reduces the efficiency of natural selection on individual alleles. Complete linkage can lead to genetic hitchhiking, in which slightly deleterious alleles are preserved because of linkage to beneficial alleles (Smith and Haigh 1974, Rice 1987) or, conversely, background selection, in which slightly beneficial alleles are lost because of linkage to deleterious alleles (Charlesworth et al. 1993). The well-documented decay of Y chromosomes is believed to result from the intersecting effects of each of these processes.

In recent years, it has become clear that not all ancestral genes were lost from the mammalian Y chromosome. Complete sequencing of the Y chromosomes of human (Skaletsky et al. 2003) and rhesus macaque (Hughes et al. 2012) revealed that the loss of ancestral genes ceased at least 25 million years ago, suggesting that remaining ancestral genes on the mammalian Y chromosome are not lingering relics but instead are preserved by natural selection (Hughes et al. 2012). Indeed, a survey of ancestral Y-chromosome genes across eight mammalian species revealed that surviving Y-chromosome genes are enriched for characteristics including dosage-sensitivity, broad expression across the human body, and regulatory functions, suggesting essential functions that made them resistant to decay (Bellott et al. 2014). Similar observations have been made regarding surviving W-chromosome genes in snakes and birds, demonstrating that the same principles drive ancestral gene survival across independent vertebrate sex chromosome systems (Bellott et al. 2017, Bellott and Page 2021). Contrary to predictions that the Y chromosome is heading for extinction (Graves 2006), there is growing understanding that remaining Y chromosome genes are permanent residents that are likely essential for male viability (Bellott et al. 2014). This supposition is supported by the fact that most human 45,X embryos are inviable (Hook and Warburton 1983); such embryos are distinguished from typical 46,XY males only by the absence of a single Y chromosome.

Conservation of ancestral gene content on the X chromosome

Unlike the Y chromosome, the X chromosome still undergoes recombination in females, shielding it from the degeneration described above. Evolutionary reconstructions of the ancestral autosome that gave rise to mammalian sex chromosomes have shown that the human X chromosome retains a remarkable ~98% of its ancestral gene content, compared to only ~3% for the male-specific region of the Y chromosome (Bellott et al. 2010, Bellott et al. 2014). To emphasize this point: Genes from the ancestral autosome have not only survived within the human genome, but moreover remain on the X chromosome, with few translocations or karyotype rearrangements over the past 300 million years (Figure 1.2). This phenomenon extends across diverse placental mammals, as X-chromosome gene content is highly conserved between species including mice (Mueller et al. 2013), pigs (Quilter et al. 2002), horses (Raudsepp et al. 2004), cows (Zimin et al. 2009), cats (Murphy et al. 2007), dogs (Murphy et al. 2007), and elephants (Delgado et al. 2009). In contrast, autosomal gene content is frequently reshuffled between chromosomes in different species, such that a block of genes found on Chromosome 1 in one species might be found on Chromosome 8 or Chromosome 12 in another (Mouse Sequencing and Analysis Consortium 2002, Carbone et al. 2006, Murphy et al. 2007) (Figure 1.2). What explains this striking synteny of X-chromosome gene content?

The answer was first hypothesized by Susumo Ohno in his comprehensive 1967 monograph on sex-chromosome evolution (Ohno 1967). He and others inferred in 1960 that a deeply staining chromosome present in a single copy in both diploid female mouse cells and tetraploid male mouse cells must correspond to a condensed X chromosome (Ohno and Hauschka 1960). Mary Lyon further proposed that the condensation of a single X chromosome in females signaled genetic inactivation of that chromosome (Lyon 1961) and could represent a form of dosage compensation that would equalize the expression of X-linked genes between males and females (Lyon 1962). Based on these studies, Ohno proposed a two-part model for the evolution of dosage compensation: 1) The expression of X-chromosome genes was up-regulated to maintain ancestral levels of gene expression in males, 2) One X chromosome was later inactivated in females to prevent over-expression of X-linked genes (Ohno 1967).

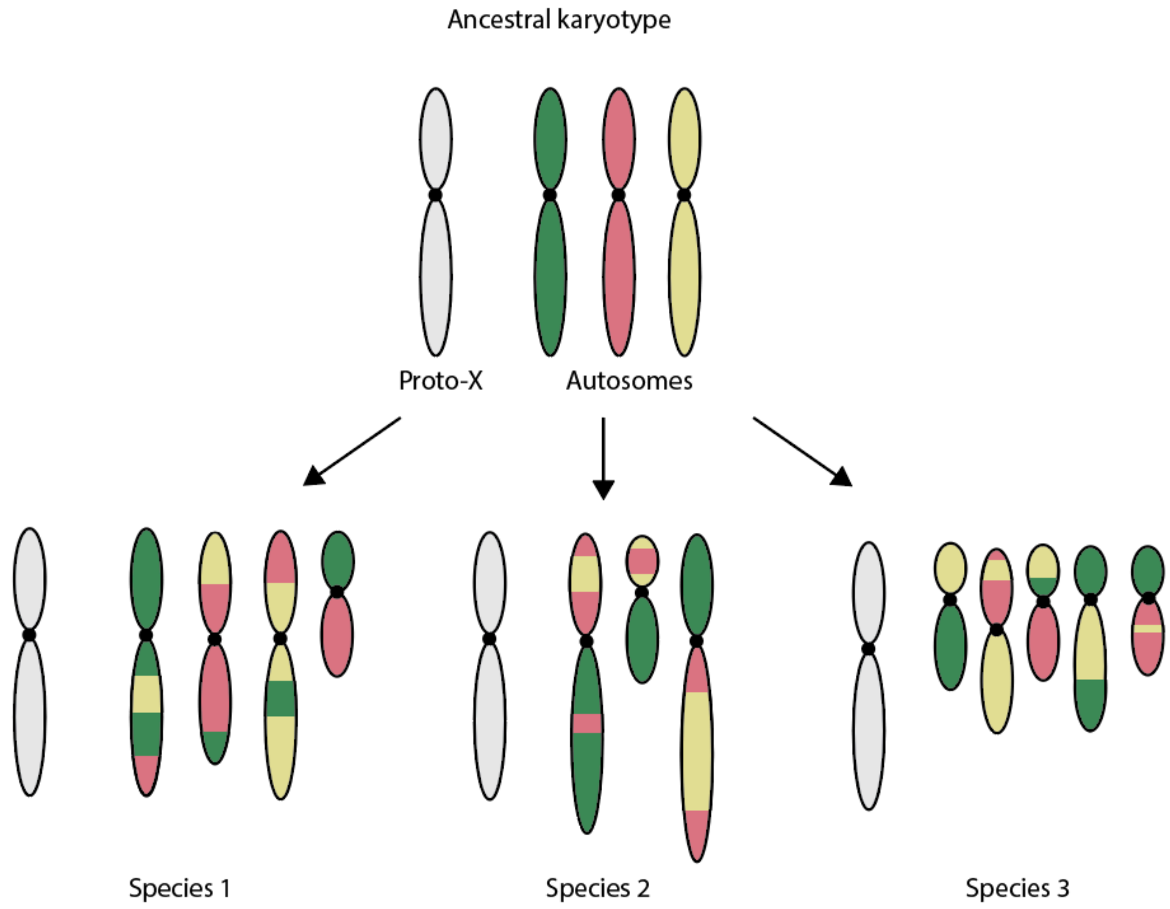


Figure 1.2 Conservation of ancestral gene content across mammalian X chromosomes. Sequence present on the proto-X chromosome (gray) is conserved in a single orthologous block among modern X chromosomes, while autosomal sequence (green, yellow and pink) is highly rearranged between species.

Key aspects of this model have been validated in mammals (reviewed in Disteche 2012), although it is now known that around 15% of human X-chromosome genes escape from X inactivation (Carrel and Willard 2005). From this model, Ohno inferred the explanation for conservation of X-chromosome gene content across mammals: Following the evolution of dosage compensation, translocations of sequence between the X chromosome and autosomes would disrupt gene expression levels; such translocations would therefore be strongly disfavored by natural selection (Ohno 1967). Dosage compensation acts as a barrier to movement of genes to and from the X chromosome, preserving the gene content of mammalian X chromosomes largely as it appeared on the ancestral autosome hundreds of millions of years ago.

DISCOVERY OF LARGE PALINDROMES ON MAMMALIAN SEX CHROMOSOMES

Ampliconic sequence on mammalian Y chromosomes

Fundamental tenets of sex-chromosome evolution, including Y-chromosome degeneration and X-chromosome preservation, were inferred with remarkable accuracy decades before the sequencing of a single mammalian chromosome, based on observations of sex-chromosome size and gene linkage across different species (Muller 1914, Ohno 1967). Complete sequencing of the first mammalian Y chromosome—the human Y chromosome in 2003—validated theories of Y-chromosome degeneration by identifying 14 homologous gene pairs present on the human X and Y chromosomes, along with additional Y-chromosome pseudogenes with functional homologs on the X chromosome (Skaletsky et al. 2003). The most striking finding, however, had little explanation in existing theories of sex-chromosome evolution. Around 30% of the male-specific region of the human Y chromosome was composed of large, nearly identical repeat units, or amplicons, harboring gene families that in most cases were not present on the ancestral autosome (Skaletsky et al. 2003). Ampliconic gene families had been acquired from other chromosomes, and nearly all were expressed predominantly in the testis, specifically within male germ cells. Even while ancestral genes had decayed, the Y chromosome had actively acquired and amplified new genes, specializing large portions of its sequence for male reproduction (Skaletsky et al. 2003).

The human Y chromosome contained one long ampliconic tandem array, but most ampliconic repeats were found in inverted orientations, forming eight massive palindromes (Skaletsky et al. 2003). Ampliconic sequence is generally defined as repeat units >10 kb in length, with minimum 99% sequence identity between repeats, to distinguish them from other segmental duplications (Mueller et al. 2013). However, some Y-chromosome palindromes dwarfed these minimum requirements: One palindrome had arms spanning 1.45 Mb each, and five other palindromes had arms spanning more than 100 kb, all with arm-to-arm identities higher than 99.9% (Skaletsky et al. 2003). Y-chromosome palindrome arms were separated by small stretches of unique sequence referred to as spacers, with lengths ranging from 2 kb to 170 kb (Skaletsky et al. 2003). Although no protein-coding genes were found in palindrome spacers, six out of eight palindromes contained one or more testis-biased protein-coding gene families in their arms, suggesting an association between genomic structure and gene content. A strong hint as to the nature of this association came from the observation that all 25 intact genes from testis-biased gene families on the human Y chromosome were found within amplicons, while more than 50 pseudogenes from the same gene families were scattered equally among amplicons and single-copy sequence (Skaletsky et al. 2003). Palindromes and other amplicons are thus strongly associated with the survival of testis-biased genes, perhaps due to the evolutionary benefits of recombination within palindromes, discussed below.

The high sequence identity between palindrome arms would be expected if they were young duplications which had not yet accumulated differences between arms. However, at least three human Y palindromes had orthologous palindromes in chimpanzee (Rozen et al. 2003), which diverged from humans around 7 million years ago (Kumar et al. 2017). The high arm-to-arm identity between palindrome arms therefore could not result from recent duplications, and was inferred to result instead from an ongoing exchange of information between palindrome arms through intra-chromosomal gene conversion (Figure 1.3). This inference was confirmed by the finding that single-nucleotide differences between palindrome arms in the human reference Y chromosome were polymorphic in human

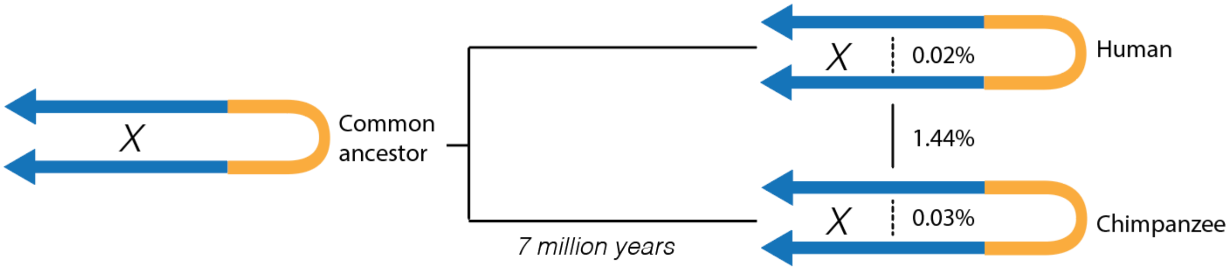


Figure 1.3 Sex-chromosome palindromes are maintained by ongoing arm-to-arm gene conversion. Blue: Palindrome arms. Yellow: Palindrome spacer. *X* indicates gene conversion between arms. Percentages show percent divergence between palindrome arms within species (dashed lines) versus between species (solid line). Values are the averages from three Y-chromosome palindromes conserved between human and chimpanzee (Rozen et al. 2003). New mutations over the past 7 million years have been rapidly homogenized within each species, leading to divergence of palindrome arms between human and chimpanzee, while maintaining near-perfect sequence identity within each species.

populations—e.g. a site that was C/T in the reference genome could be C/C, C/T, or T/T in different men, demonstrating the erasure of nucleotide differences through gene conversion (Rozen et al. 2003). Given the observed divergence between arms and the known mutation rate of the Y chromosome, it was calculated that ~600 base pairs of Y-palindrome arms must undergo gene conversion during every male meiosis (Rozen et al. 2003). The non-PAR portion of the Y chromosome had previously been known as the "non-recombining region of the Y chromosome" (NRY) due to its known inability to recombine with the X chromosome; in fact, palindrome formation enabled regions of the Y chromosome to recombine with itself, leading the NRY to be re-christened the MSY, or "male-specific region of the Y chromosome" (Skaletsky et al. 2003).

The sequencing of additional mammalian Y chromosomes, including those of chimpanzee (Hughes et al. 2010), rhesus macaque (Hughes et al. 2012), mouse (Soh et al. 2014), and bull (Hughes et al. 2020), revealed the existence of Y amplicons in every species studied. Yet while amplicons are a common feature of mammalian Y chromosomes, they differ dramatically between species: Ampliconic sequence represents less than 5% of the rhesus macaque Y chromosome (Hughes et al 2012) yet comprises more than 80% of the Y chromosome in bull (Hughes et al. 2020), and a remarkable 98% of the Y chromosome in mouse (Soh et al. 2014). Amplicons on the mouse Y chromosome contain testis-biased gene families that have no homologs in the human MSY, demonstrating that mouse and human amplicons were acquired independently (Soh et al. 2014). Out of two highly amplified testis-biased gene families on the bull Y chromosome, one was acquired independently by the bull Y, while the other has a homologous gene family on the human Y chromosome (Hughes et al. 2020). Y-chromosome palindromes are poorly conserved even among primates: The chimpanzee Y chromosome has 19 massive palindromes, only seven of which have sequence orthology to human Y palindromes, despite humans and chimpanzees sharing a common ancestor only seven million years ago (Hughes et al. 2010, Kumar et al. 2017).

Understanding the evolution of Y-chromosome amplicons is complicated by two factors: Complex amplicon architecture within each species, and dramatic structural rearrangements of the Y chromosome between species. Several human Y palindromes have partial orthology to other human Y-palindromes, creating stretches of sequence that have both direct and inverted repeats elsewhere on the chromosome (Skaletsky et al. 2003). The same is true of Y palindromes in chimpanzee, which are frequently present in two or more copies (Hughes et al. 2010). Among bull and mouse Y amplicons, the situation is even more challenging, as repeat units are nearly always found in a complex mixture of palindromes and tandem arrays, with repeat units found in dozens or even hundreds of copies (Soh et al. 2014, Hughes et al. 2020). These factors, combined with additional large-scale rearrangements of the Y chromosome between species, make it difficult to identify 1:1 structural orthologs between species. As I will explore below, the situation is very different on the human X chromosome, where 26 massive palindromes are each found in a single distinct copy—and where, as discussed above, ancestral X-chromosome sequence is highly conserved among placental mammals, providing a clear roadmap to identifying orthologous stretches of sequence between species.

The human X chromosome contains large palindromes with testis-biased gene families

Historically, only two mammalian X chromosomes were sequenced using methods capable of resolving palindromes and other amplicons: The human X chromosome (International Human Genome Sequencing Consortium 2004, Ross et al. 2005, Mueller et al. 2013) and the mouse X chromosome (Church et al. 2009) (Figure 1.4). I will discuss reasons for this dearth, including the technical difficulties of resolving complex genomic repeats using traditional genome assembly methods, in a later section of this introduction. First, I will discuss insights from comparing the structure and gene content of mouse and human X chromosomes, as well as a limited but suggestive pool of literature about the function of human and mouse X-palindrome genes.

A comparative study in 2013 used the complete sequences of the human and mouse X chromosomes to rigorously test Susumo Ohno's prediction that X-chromosome gene content should be conserved across placental mammals (Ohno 1967, Mueller et al. 2013). Among single-copy genes, 94-95% were indeed conserved between the human and mouse X chromosomes, consistent with Ohno's predictions. However, the X chromosomes from both species were also found to contain extensive

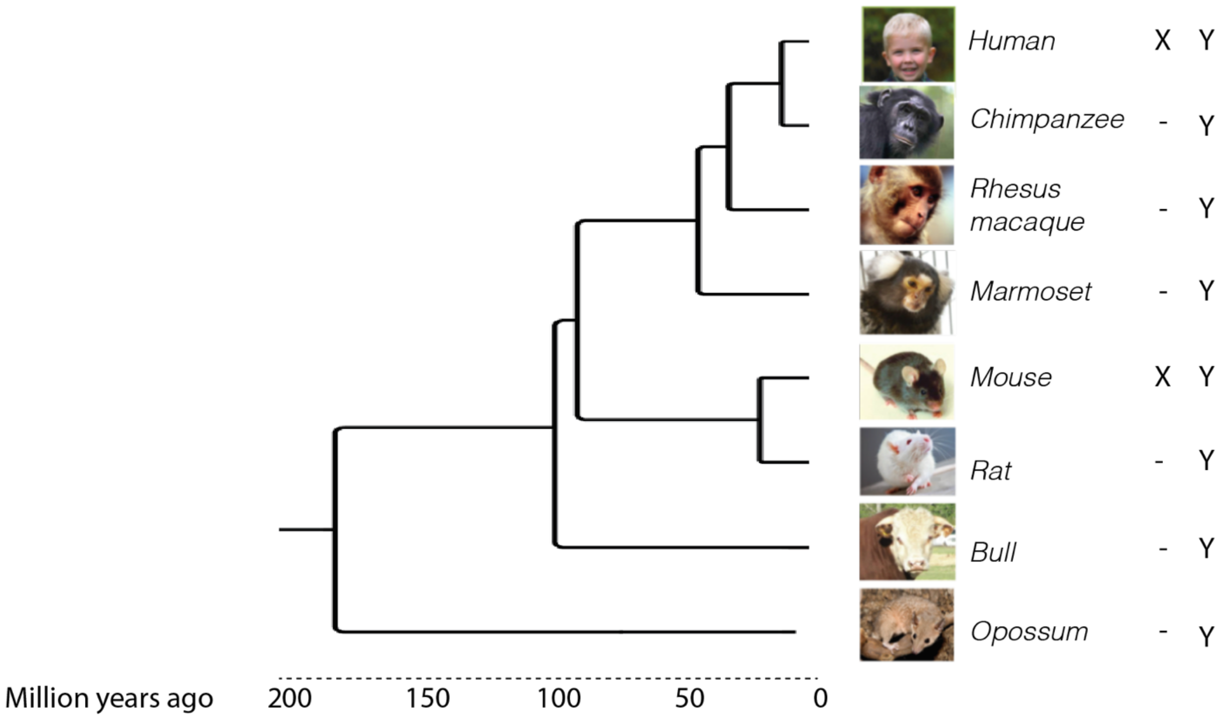


Figure 1.4 Mammalian sex chromosomes with complete sequence available. To be included, each chromosome must have been sequenced using a clone-based method from a single haplotype (see a discussion of the sequencing challenges associated with sex-chromosome amplicons in Introduction section "Technical barriers to the study of X-chromosome palindromes"). Three of these Y-chromosome assemblies—those of marmoset, rat, and opossum—are not yet published (H. Skaletsky, personal communication).

ampliconic sequence: The human X chromosome has 3.15 Mb of ampliconic sequence with 107 ampliconic protein-coding genes, while the mouse X chromosome has 19.42 Mb of ampliconic sequence with 149 ampliconic protein-coding genes (Mueller et al. 2008, Mueller et al. 2013). Altogether, it was found that around 2% of the human X chromosome is comprised of ampliconic sequence, including at least two dozen large palindromes with arm lengths up to 140 kb (Warburton et al. 2004, Mueller et al. 2013). In contrast to single-copy genes, fewer than one-third of ampliconic genes found on the X chromosome in one species were present on the other (Mueller et al. 2013) (Figure 1.5). Ampliconic genes on the X chromosome had strong parallels to ampliconic genes on the Y chromosome: They were expressed predominantly in testis, and inferred to have been independently acquired since the divergence of mouse and human around 90 million years ago based on their absence from outgroup species (Mueller et al. 2013, Kumar et al. 2017). These results suggested that evolution of the mammalian X chromosome has been bimodal, with ancestral portions of the chromosome strongly conserved between species, and newer ampliconic regions containing species-specific genes involved in male reproduction. Male reproductive genes are not expected to be regulated by X-inactivation, which occurs only in females. However, a subset of X-amplicon genes are expressed in both male and female tissues (Warburton et al. 2004, Mueller et al. 2008, Mueller et al. 2013); to our knowledge, the potential impact of X-inactivation on such X-amplicon genes in females has not been characterized. Notably, while the mouse X chromosome contains many amplicons that are present in multiple copies and include both tandem and inverted repeats, all palindromes on the human X chromosome are present only once, with two arms that are unique to that palindrome.

While the finding that the mammalian Y chromosome is specialized for male reproduction is somewhat intuitive, given that the Y chromosome is only found in males, it is less obvious why the mammalian X chromosome should show the same tendency. However, theoretical work predicted this result as early as 1931. In brief, we can consider the case of sexually antagonistic traits, i.e. traits that benefit one sex but are harmful to the other. Ronald Fisher noted that sexually antagonistic male-benefit alleles should flourish on the Y chromosome, because their harmful effects in females would never be

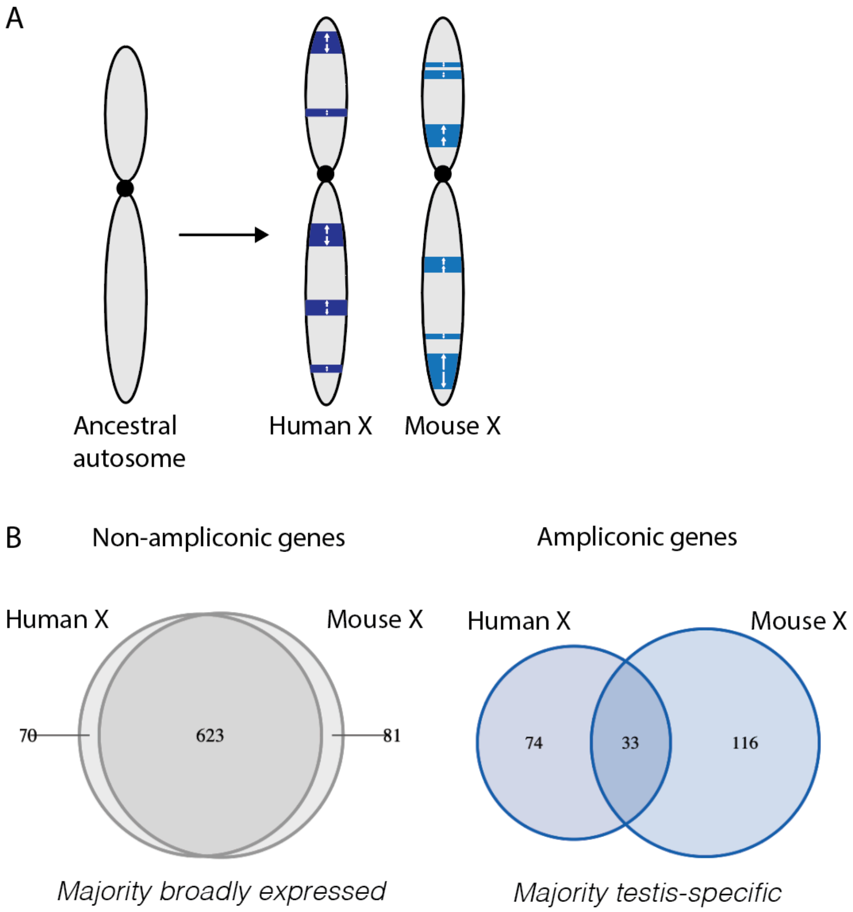


Figure 1.5 Amplicons on the human and mouse X chromosomes. A) Schematic of bimodal evolution of the X chromosome. The human and mouse X chromosomes both retain the majority of their ancestral sequence (gray), but have independently acquired ampliconic sequence, including palindromes and tandem arrays (dark blue: human, light blue: mouse). B) Conservation and expression patterns of ampliconic genes versus non-ampliconic genes between human and mouse. Data from Mueller et al. 2013.

expressed and therefore would never be subject to negative selection (Fisher 1931). When sexually antagonistic male-benefit alleles are recessive, however, they are also favored to accumulate on the X chromosome: Such alleles would be subject to positive selection immediately in males, yet would not be selected against in females until reaching sufficiently high frequency in the population for homozygosity to become common (Fisher 1931, Rice 1984). While this hypothesis could explain the enrichment of testis-biased genes on the human (Ross et al. 2005, Mueller et al. 2013) and mouse (Wang et al. 2001, Mueller et al. 2008, Mueller et al. 2013) X chromosomes, the specific functions of most human and mouse X-amplicon genes remain unknown. I briefly summarize what is known about these genes in the following section.

Limited functional characterization of X-palindrome genes

We start with the case of human X-palindrome genes. As a whole, human X-palindrome genes are significantly depleted for association with Mendelian phenotypes (Mueller et al. 2013, McKusick-Nathans Institute of Genetic Medicine 2020). While their testis expression might predict roles related to male fertility, the majority of literature searches for human X-palindrome genes instead reflect their involvement in an unexpected process: Many are designated as "cancer-testis antigens," or CTAs, based on their expression in the testis as well as a range of solid tumors (reviewed in Simpson et al. 2005). This unique expression pattern has generated interest in CTAs as potential targets for cancer immunotherapy, given that the testis is an immune-privileged tissue, and therapies that target these genes could in theory target cancer cells with few effects on healthy tissues.

Results from clinical trials targeting CTAs have been mixed. Early attempts at targeting one X-palindrome CTA, an antigen produced by the gene *CTAG1*, with engineered autologous T-cells were successful in a small number of patients with metastatic melanoma (Robbins et al. 2011). However, clinical trials for CTAs were set back when engineered autologous T-cells against a second X-palindrome CTA, an antigen produced by the gene *MAGEA3*, resulted in unexpected neurotoxicity due to low-level expression of *MAGEA3* in the brain (Morgan et al. 2013). More recently, an RNA vaccine against four cancer antigens, including those produced by *CTAG1* and *MAGEA3*, induced clinical responses in a subset of patients with advanced melanoma (Sahin et al. 2020). While studies of X-palindrome gene expression patterns and potential clinical utility are widespread, we note that little of this literature has described functional roles for X-palindrome genes in either cancerous tumors or the testis. Given the interest that X-palindrome genes have generated as clinical targets, the lack of information about their origins and functions in a healthy context is even more striking.

More information is available about the functions of X-amplicon genes in mouse. Knockouts of six *Magea* family genes, which are present in highly divergent copies on both the human and mouse X chromosomes, result in lower testis weights and increased apoptosis of male germ cells, consistent with hypothesized roles in spermatogenesis (Hou et al. 2016). However, litter size was unaffected, suggesting that such defects were below the threshold to negatively impact male fertility (Hou et al. 2016). Consistent with these results, knockout of a different subset of eight *Magea* family genes resulted in decreased male fertility, but only under stress conditions including treatment with DNA damaging agents or starvation (Fon Tacer et al. 2019). Both of these studies were hindered by potential redundancy between the targeted genes and other more divergent genes from the MAGE superfamily, yet taken together, these studies suggested that *Magea* genes may play modulatory roles in spermatogenesis, particularly under stress conditions. The finding that mouse X-amplicon genes have quantitative rather than absolute effects on male fertility may explain why so few functions for X-palindrome genes have been identified in humans, as quantitative associations would not have been revealed by studies looking at infertile men, and would not appear in databases of Mendelian phenotypes.

One particularly interesting case is that of the mouse gene family *Slx/Sly*, which is co-amplified in amplicons on the mouse X and Y chromosomes: The X chromosome contains more than 30 copies of two closely related sub-families, *Slx* and *Slxl1*, while the Y chromosome contains more than 120 copies of *Sly*, which encodes a protein with around 43% identity to the X-linked copies (Soh et al. 2014). Male mice deficient for *Slx/Slxl1* or *Sly* are sub-fertile, and male mice deficient for *Slx/Slxl1* have a significant bias towards male offspring (Cocquet et al. 2009, Cocquet et al. 2012, Kruger et al. 2019). Remarkably, male mice deficient for both *Slx/Slxl1* and *Sly* show normal fertility and a normal offspring sex ratio, demonstrating that the primary function of these co-amplified gene families is not to promote male fertility, but rather to compete with each other for transmission during spermatogenesis (Cocquet et al. 2012). Similar co-amplified X/Y gene families have been observed in Drosophila (Ellison and Bachtrog 2019), mouse (Soh et al. 2014), and bull (Hughes et al. 2020), and are also hypothesized be involved in meiotic drive, although the effects of removing these gene families have not been directly tested. There are two examples of co-amplified gene families on the human X and Y chromosomes, albeit with much lower copy numbers: *VCX/VCY* (4 copies and 2 copies, respectively) (Lahn and Page 2000) and *HSFX/HSFY* (4 copies and 2 copies, respectively). It is not known whether these gene families have antagonistic roles in humans, or whether similar conflicts occur between other unrelated X and Y amplicon gene families.

EVOLUTIONARY DYNAMICS OF SEX-CHROMOSOME PALINDROMES

Consequences of high rates of gene conversion in sex-chromosome palindromes

As discussed above, several previous studies have compared the structure and gene content of sex-chromosome amplicons between species (e.g. Hughes et al. 2010, Hughes et al. 2012, Mueller et al. 2013). However, given the small number of palindromes that are conserved between species, there have been few opportunities to examine the molecular evolution of palindromes. It has been recognized for more than a decade that the maintenance of arm-to-arm identity by high rates of intra-chromosomal gene conversion over millions of years creates opportunities for unusual patterns of molecular evolution (Rozen et al. 2003). Among three Y palindromes conserved between human and chimpanzee, sequence divergence was lower between species in the palindrome arm than in the spacer or the non-ampliconic MSY (Rozen et al. 2003). Reduced divergence was observed even in transposable elements and other non-coding sequence, suggesting that it did not result from selective constraint in palindrome arms, but rather from neutral evolutionary forces. Rozen et al. suggested that this could result from a slight preference for gene conversion to restore the ancestral base rather than a new substitution, also known as a derived base (Rozen et al. 2003). Two later studies showed that gene conversion within human Y palindromes was indeed more likely to restore the ancestral state than the derived base (Hallast et al. 2013, Skov et al. 2017). However, studies of palindromes in other contexts have found the opposite result: A comparative study of X palindromes conserved between several species of mouse revealed that palindrome arms evolve more rapidly than surrounding sequence (Swanepoel et al. 2020), and the same result was observed for a Y-chromosome palindrome conserved between several species of rabbits (Geraldes et al. 2010). These conflicting results suggest context-dependent effects of high rates of gene conversion on sequence divergence. Importantly, divergence rates are also influenced by natural selection: High rates of recombination are associated with more efficient natural selection in cases of both positive and negative selection (Felsenstein 1974), which could enable rapid evolution in regions under positive selection, and slower evolution in regions under purifying selection. Interpretation of altered divergence rates in sex-chromosome palindromes is therefore made challenging by the poor functional characterization of palindromes discussed above.

While evidence for the ability of gene conversion to preferentially restore the ancestral base remains mixed, a second preference is much more well established: GC-biased gene conversion (reviewed in Galtier et al. 2001, Marais 2003, Duret and Galtier 2009). Mammalian genomes are characterized by large regions with similar GC content, gene density, and density of interspersed repeats, sometimes referred to as isochores (Bernardi et al. 1985). Adam Eyre-Walker first suggested in 1993 that the existence of isochores could be explained by a bias of gene conversion to preferentially fix GC bases over AT bases. His two key arguments were as follows: 1) GC content in mammalian genomes is positively correlated with the density of chiasmata, suggesting an association with recombination; 2) GC content is particularly low on the Y chromosome, which does not recombine (except for within the PAR and palindromes, as discussed above) (Eyre-Walker 1993). This hypothesis was bolstered by the discovery that genes that translocate from a GC-poor region into a GC-rich region undergo a subsequent increase in GC content (Montoya-Burgos et al. 2003, Li et al. 2016). Most compellingly, GC-biased gene conversion has now been detected directly through identifying gene conversion events from human pedigrees: In a choice between an AT base and GC base, gene conversion was found to fix the GC base around 70% of the time (Williams et al. 2015, Halldorsson et al. 2016) (Figure 1.6). Evidence for GC-biased gene conversion is now widespread across taxa including plants (Muyle et al. 2011), yeast (Mancera et al. 2008), birds (Smeds et al. 2016), and mammals (Galtier et al. 2009, Clément and Arndt 2011, Odenthal-Hesse et al. 2014). Although the mechanism of GC-biased gene conversion is still debated, it likely arises from GC bias in the mismatch repair machinery that fixes heteroduplex DNA formed during recombination (Lesecque et al. 2013).

It was previously proposed that the high rate of gene conversion in sex-chromosome palindrome arms might lead to elevated GC content over time. A human X palindrome with putative orthologs in 16 mammalian species was found to have GC content significantly higher than that of flanking sequence in all species studied (Caceres et al. 2007). Another study examined nucleotide replacements patterns in a Y-chromosome palindrome conserved between human, chimpanzee, and gorilla, and found an elevated frequency of AT→GC nucleotide replacements in palindrome arms compared to the spacer, consistent with an evolutionary history of GC-biased gene conversion in palindrome arms (Hallast et al. 2013). However, among six human Y-chromosome palindromes examined, only two showed elevated GC content compared to nearby single-copy sequence (Hallast et al. 2013). We note that there is little evidence that human Y-chromosome palindromes are conserved outside of apes, meaning that they may be too young to show long-term effects from GC-biased gene conversion. Additionally, ampliconic sequence on the human Y chromosome is more GC-rich than single-copy sequence when considered as a whole—40.2% versus 38.4%, respectively (Skaletsky et al. 2003)—suggesting that real signal is present, but may vary between different amplicons. It has been previously proposed that palindromes may have unique epigenetic states (Skaletsky et al. 2003); DNA methylation status could also influence the evolution of GC-rich regions, given that highly methylated DNA is prone to CpG hypermutability (Bird 1980). For example, if palindromes are highly methylated, they may experience a high frequency of GC→AT mutations, which could partially offset the expected increase in GC content from GC-biased gene conversion. Altogether, previous results suggest the potential for unusual patterns of molecular

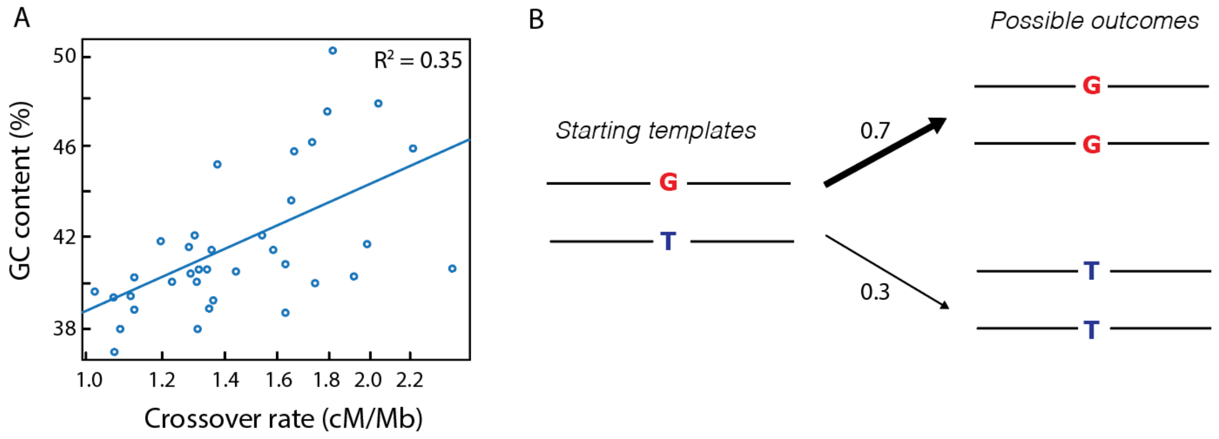


Figure 1.6 GC-biased gene conversion. A) GC content is positively correlated with recombination rates across the human genome. Crossover rates are plotted in log scale. Adapted from Duret and Galtier 2009. B) Schematic of GC-biased gene conversion. When gene conversion occurs between one template with a "strong" base (G or C) and another with a "weak" base (A or T), the strong base is transmitted around 70% of the time (Williams et al. 2015, Halldorsson et al. 2016).

evolution in palindrome arms driven by high rates of arm-to-arm gene conversion, but insights have been limited by the small number of palindromes conserved between species. In Chapter 3, I will describe the effects of GC-biased gene conversion in a set of twelve primate X palindromes conserved between human, chimpanzee, and rhesus macaque.

Structural rearrangements associated with sex-chromosome palindromes

The previous section discussed the consequences of high rates of gene conversion between sex-chromosome palindrome arms. However, recombination between palindrome arms can also lead to crossover events in which non-homologous sequence is exchanged between arms. Crossovers in X- and Y- chromosome palindromes are associated with a wide range of benign and pathogenic rearrangements, including inversion of the palindrome spacer (when recombination occurs within a sister chromatid) (Lakich et al. 1993, Small et al. 1997) as well as the formation of isodicentric chromosomes (when recombination occurs between sister chromatids) (Lange et al. 2009, Scott et al. 2010) (Figure 1.7). Isodicentric chromosomes are mitotically unstable, and the transmission of an isodicentric sex chromosome to offspring typically leads to loss of that chromosome early in embryonic development, which has been linked clinical outcomes including sex reversal, spermatogenic failure, and Turner's syndrome (i.e. a 45,X karyotype) (Lange et al. 2009). As mentioned previously, duplication of part or all of some human Y-chromosome palindromes creates tandem repeats that are also substrates for recombination (Figure 1.7). Recombination between tandem repeats in the human *AZFc* region causes deletion of intervening sequence, including several Y-amplicon gene families, resulting in male infertility (Vogt et al. 1996, Kuroda-Kawaguchi et al. 2001). On the Y chromosome, the combination of tandem and inverted repeats creates abundant opportunities for non-allelic recombination, contributing to the highly divergent Y-chromosome structures observed between primates (Hughes et al. 2010, Hughes et al. 2012). On the X chromosome, where palindromes are present in a single copy, the potential for rearrangements over long evolutionary timescales is more limited: Without additional opportunities for non-allelic recombination from duplicated palindromes, the only predicted viable rearrangements are inversions of

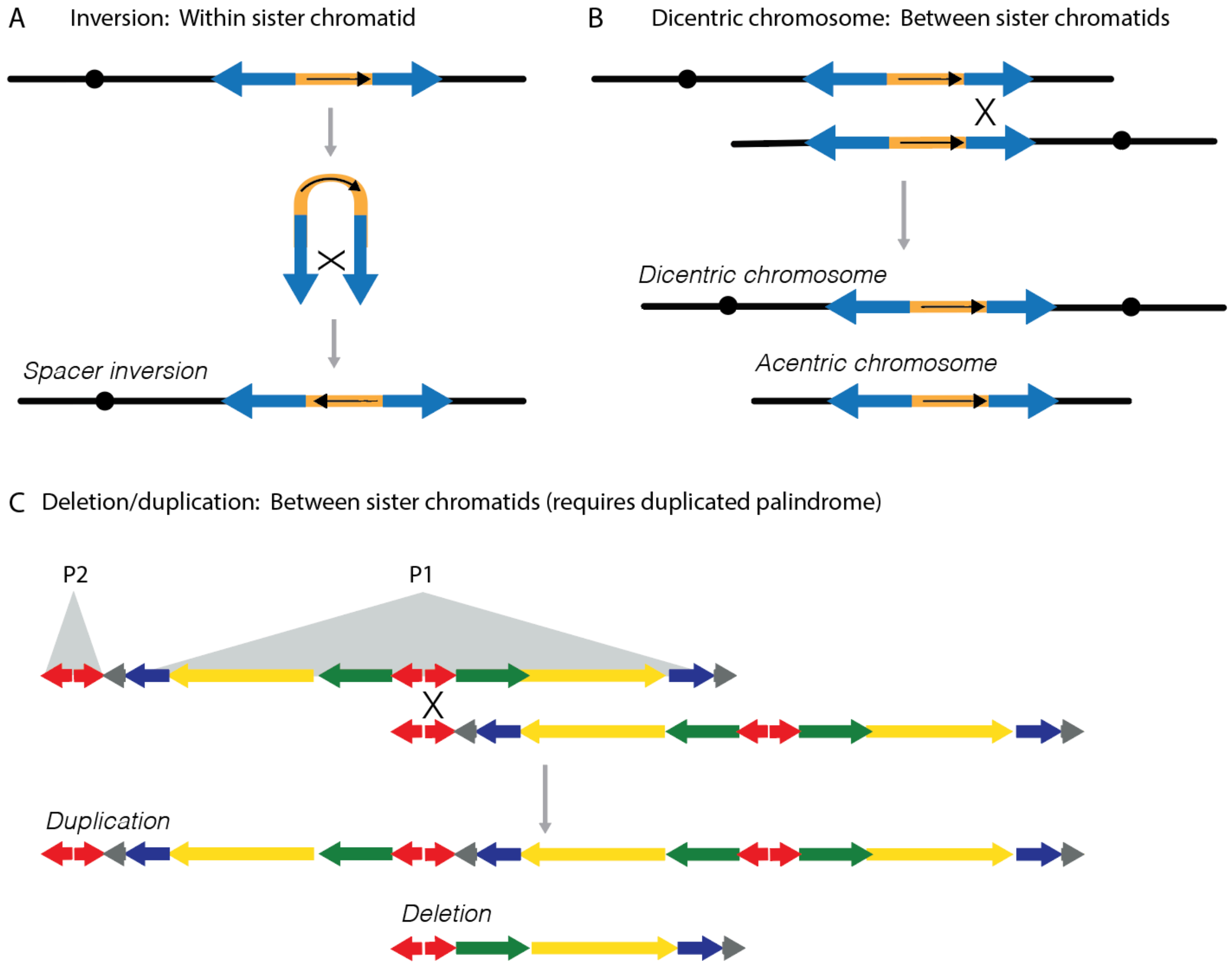

A
Inversion: Within sister chromatid
Spacer inversion
B
Dicentric chromosome: Between sister chromatids
X
Dicentric chromosome
Acentric chromosome
C
Deletion/duplication: Between sister chromatids (requires duplicated palindrome)
P2
P1
X
Duplication
Deletion

Figure 1.7 Rearrangements mediated by recombination between sex-chromosome palindrome arms. A) Generation of a spacer inversion. B) Generation of dicentric and acentric chromosomes. C) Generation of duplications and deletions. Note that the mechanism in C) requires that the palindrome be at least partially duplicated elsewhere on the chromosome, generating a tandem repeat. The example shown represents human Y-chromosome palindromes P1 and P2; P2 is nearly identical to the innermost portion of P1. Adapted from Teitz et al. 2018.

the palindrome spacer. X-chromosome palindromes have been implicated in the formation of isodicentric X chromosomes (Scott et al. 2010), but such mitotically unstable rearrangements are not conducive to long-term evolutionary transmission and fixation.

TECHNICAL BARRIERS TO THE STUDY OF X PALINDROMES

Generation of accurate X-palindrome reference sequence

Palindromes on the human X chromosome were not completely sequenced until specialized efforts in 2013 (Mueller et al. 2013), despite publication of more than 99% of the sequence of the human X chromosome in 2005 (Ross et al. 2005). Similarly, the complete Y-chromosome sequences discussed previously—those of bull, chimpanzee, and rhesus macaque—were each published years after the initial genome sequence for those respective species became available (Chimpanzee Sequencing and Analysis Consortium 2005, Rhesus Macaque Sequencing and Analysis Consortium 2007, Bovine Genome Sequencing and Analysis Consortium 2009). These delays resulted from the unique challenges of accurately sequencing long and highly identical segmental duplications, including sex-chromosome palindromes.

An analysis of segmental duplications in the working draft of the human genome (International Human Genome Sequencing Consortium 2001) identified two primary issues (Bailey et al. 2001). First, fewer than half of inter-chromosomal segmental duplications that had been previously identified using fluorescence in-situ hybridization (FISH) were present in the working draft sequence, suggesting that most segmental duplications were missing or assigned to the wrong chromosome (Bailey et al. 2001). When read lengths are shorter than the segmental duplication repeat unit (as is true for Sanger reads, with lengths up to 1 kb, and especially for Illumina reads, with lengths up to 300 bp), automated assembly programs generally fail to recognize that multiple repeat units are present, resulting in collapsed assemblies that contain a single repeat unit (Figure 1.8A). Second, the working draft assembly showed—conversely—an unexpectedly high number of segmental duplications with identities >98%. The human draft sequence had been generated using DNA from multiple individuals, and many of these highly

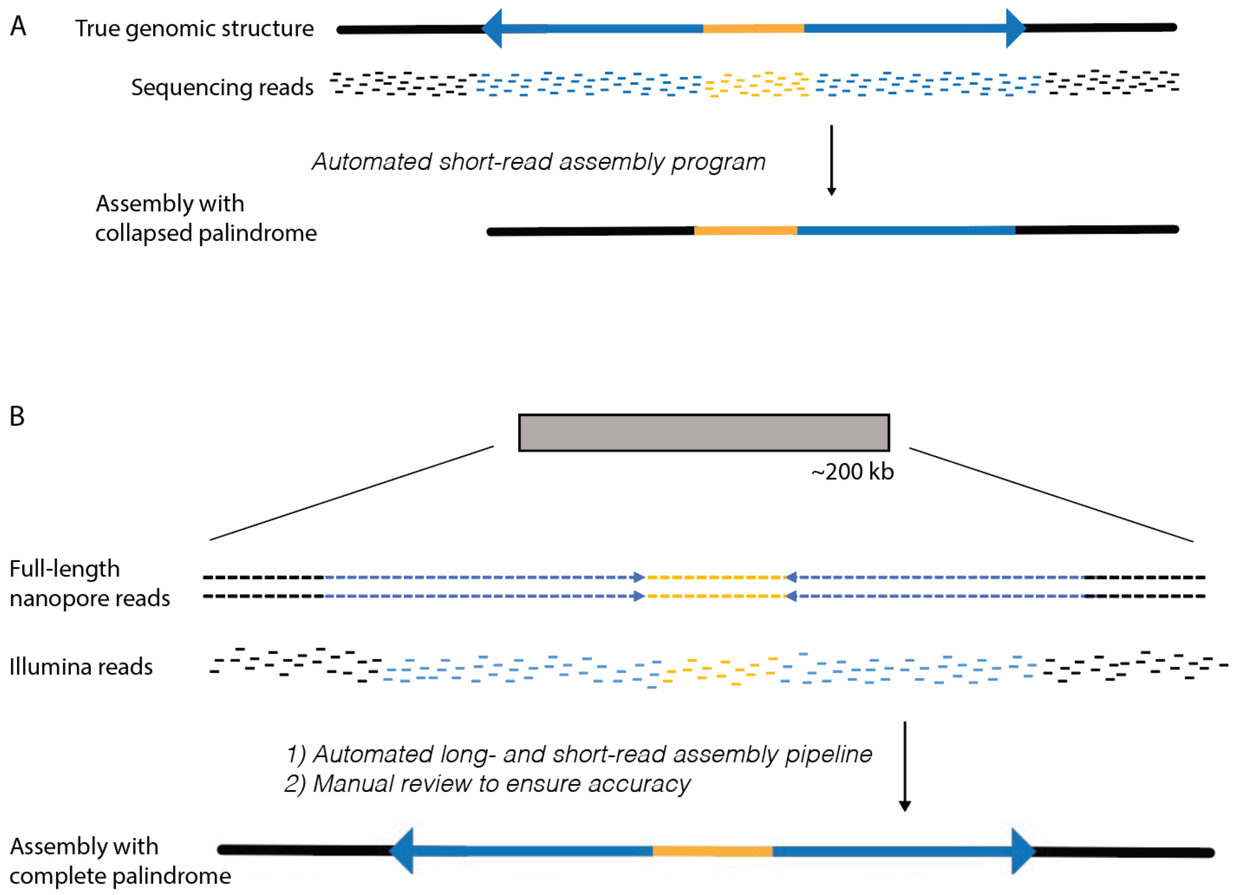


Figure 1.8 Generation of accurate reference sequence for palindromes. A) Collapsed sequence assembly caused by read lengths shorter than the length of the palindrome arm. B) Generation of accurate reference sequence using a clone-based approach that combines long (>100 kb) nanopore reads plus short Illumina reads. Full method is described in the Appendix.

identical ‘segmental duplications’ were found to represent different haplotypes, rather than true genomic duplications (Bailey et al. 2001). The task of fixing these sequencing errors was grimly described as “akin to ‘mopping up the dance floor after the band has gone home”—an “arduous task with little reward, done by a few people willing to don the overalls, put the trash where it belongs, and pick up the pieces” (Eichler 2001). Segmental duplications are well-known hotspots of human genetic variation, including both pathogenic rearrangements (Stankiewicz and Lupski 2010) and sites of genomic adaptation (Dennis and Eichler 2016), making their accurate representation in reference genomes essential.

Both of these technical challenges were addressed by the creation of a new sequencing method, now known as Single-Haplotype Iterative Mapping and Sequencing (SHIMS), in 2001. SHIMS was developed to sequence the azoospermia factor c (*AZFc*) region of the human Y chromosome. At the time, the *AZFc* region was known to contain testis-specific gene families whose deletion was a common cause of male infertility (Vogt et al. 1996), but the copy number and orientations of these genes were unknown (Kuroda-Kawaguchi et al. 2001). SHIMS had two key features: 1) the use of DNA from a single haplotype (which for sex chromosomes can be achieved by choosing one XY male), which prevented different haplotypes from being mistaken for genomic repeats, and 2) a clone-based approach, which allowed repeat units to be sequenced independently in different clones, preventing them from being collapsed by automated assembly programs. SHIMS was used successfully to sequence not only the *AZFc* region of the human Y chromosome (Kuroda-Kawaguchi et al. 2001), but also the entire human Y chromosome (Skaletsky et al. 2003), the chimpanzee (Hughes et al. 2010), the rhesus macaque (Hughes et al. 2012), and mouse (Soh et al. 2014) Y chromosomes, and amplicons on the human X chromosome (Mueller et al. 2013). However, the task of generating and assembling Sanger reads for each clone made it expensive and laborious, particularly for projects requiring dozens to hundreds of clones.

An update to the SHIMS protocol in 2017 took advantage of the relatively low cost of Illumina reads and barcoding reads from different clones to create a more efficient and cost-effective protocol (Bellott et al. 2018). This revised protocol allowed the sequencing of up to 192 clones simultaneously, and was more than two orders of magnitude cheaper than original method (Bellott et al. 2018). However, assembly was still difficult or impossible in cases where a repeat unit is shorter than the length of the clone (typically 150 to 200 kb), as short Illumina reads could still be collapsed into a single repeat unit by automated assembly algorithms. The invention of long-read technologies such as PacBio (Eid et al. 2009) and Oxford Nanopore (Jain et al. 2016), with read lengths that can exceed 40 kb and 800 kb, respectively, now provide the means to address this issue. In the Appendix, we describe an update to the SHIMS protocol that involves incorporating one or more nanopore reads than spans the entire length of a clone, enabling instant resolution of the genomic structure and minimizing the need for manual finishing by a bioinformatics scientist (Figure 1.8B). Nanopore and PacBio reads have also been used within a whole-genome shotgun approach to generate new mammalian genome assemblies with extremely high contiguity (see Kronenberg et al. 2018, Miga et al. 2020, Warren et al. 2020, and others). Although the whole-genome shotgun approach with long reads still fails to accurately assemble a subset of regions with complex genomic architecture (Miga et al. 2020, see also Figure 2 in Chapter 2), improvements in both long-read technologies and assembly algorithms will continue to increase the representation of palindromes and other segmental duplications in mammalian reference genomes.

Appropriate treatment of X-chromosome palindromes in bioinformatic analyses

Challenges to studying sex-chromosome palindromes do not end with the generation of high-quality reference sequence. Off-the-shelf bioinformatic tools designed for downstream genomic and transcriptomic analysis frequently fail to deal appropriately with sequencing reads that map equally well to two or more genomic locations, also known as multi-mapping reads (reviewed in Treangen and Salzberg 2012). This issue has become more widespread as Sanger reads have been gradually replaced by shorter Illumina reads, which typically have read lengths of 50 to 150 bp, and where the true origin of the read cannot always be determined. The most common solution to this dilemma is to include only uniquely mapping reads in downstream sequence analysis. While this strategy prevents the assignment of reads to incorrect locations, it also makes highly identical repeat units such as palindromes and tandem repeats functionally invisible in these analyses.

Specialized bioinformatic tools, and sometimes specialized settings within off-the-shelf bioinformatic tools, can be used to address this problem. A recent study used an RNA-Seq quantification tool called kallisto, which probabilistically distributes multi-mapping reads using an optimization algorithm, to quantify the expression of Y-chromosome genes across the human body (Bray et al. 2016, Godfrey et al. 2020). This project re-analyzed raw data from the GTEx Consortium, which had previously published their own gene expression estimates using software that discarded multi-mapping reads (The GTEx Consortium 2020). A comparison between the results reported from GTEx and those found using kallisto revealed stark differences. Among genes from the human MSY, the majority of which are ampliconic, only around 40% were reported to be expressed in the published GTEx data, while 80% were found to be expressed using kallisto—sometimes at levels two orders of magnitude higher than the levels reported by GTEx (Godfrey et al. 2020). Tools that retain information from multi-mapping reads have also been successfully developed for other methods that require mapping short sequencing reads back to a reference genome, including ChIP-Seq (Chung et al. 2011) and the detection of genomic copy level variation (Teitz et al. 2018). However, as in the case of the GTEx dataset, not all large-scale genomic and transcriptomic analyses take advantage of these solutions, which frequently require more effort to implement than default software.

A separate issue affecting the study of X palindromes is the inclusion or exclusion of the X chromosome itself from large-scale genomic and transcriptomic analyses. As noted above, human X-amplicon genes are depleted for Mendelian disorders (Mueller et al. 2013). However, not all disorders have a Mendelian inheritance pattern; as discussed above, disruption of a subset of mouse X-amplicon genes appears to have smaller quantitative effects on spermatogenesis and male fertility. The contribution of genetic variation to complex disorders is often determined through genome-wide association studies, or GWAS. A review study found that the X chromosome was analyzed in only around one-third of published GWAS studies; as a result, the X chromosome is significantly depleted for GWAS hits compared to autosomes (Wise et al. 2013). A commonly cited reason for the omission of the X chromosome is the need for specialized methods to account for different X-chromosome copy number in men and women, despite the fact that methods have previously been developed to address this issue (Zheng et al. 2007, Clayton 2008). As with the technical issues presented by multi-mapping reads, solutions to including the X chromosome in GWAS studies exist, but are not always utilized by large-scale studies for the sake of convenience. Taken together, all three issues described in this section—the under-representation of palindromes in mammalian reference genomes, the under-utilization of methods that deal appropriately with palindromes in downstream bioinformatics analyses, and the exclusion of the X chromosome as a whole from some large-scale genomic analyses—likely account for our limited understanding of the functions and evolution of X palindromes and their associated gene families.

SUMMARY

The evolution of sex and sex chromosomes has captured scientists' imagination for more than a century (Muller 1914). In recent years, understanding of the autosomal origins of mammalian sex chromosomes has informed new research into sex differences in health and disease, including the roles of surviving Y-chromosome genes across the human body (Bellott et al. 2014, Godfrey et al. 2020) and conserved sex differences in autosomal gene expression (e.g. Naqvi et al. 2019). However, understanding of amplicons and other non-ancestral sequence on mammalian sex chromosomes has lagged behind. Amplicons were first sequenced and described only twenty years ago (Kuroda-Kawaguchi et al. 2001), and their study remains limited by the technical challenges described above. Although Y-chromosome amplicons have been characterized in three primate species—human (Kuroda-Kawaguchi et al. 2001, Skaletsky et al. 2003), chimpanzee (Hughes et al. 2010), and rhesus macaque (Hughes et al. 2012)—the limited conservation of Y palindromes between species has limited insights into Y-palindrome evolution.

In this introduction, I described several features of the X chromosome that make it a promising place to search for conserved palindromes, including highly conserved ancestral gene content across mammals and the fact that each human X palindrome is present in a single distinct copy. In Chapter 2, I describe how we searched for orthologous palindromes in two non-human primates, the chimpanzee and the rhesus macaque, by sequencing orthologous portions of the X chromosome using a clone-based approach incorporating long nanopore reads (see also the Appendix). I present evidence that twelve palindromes have been conserved on the primate X chromosome for 25 million years, and that poorly characterized human X-palindrome gene families are preserved by natural selection. In Chapter 3, I investigate a second dynamic of palindrome evolution: The high rates of intra-chromosomal recombination that maintained X palindromes over millions of years have also contributed to unusual nucleotide replacement patterns in X-palindrome arms, through GC-biased gene conversion that favors the fixation of GC bases over AT bases. Altogether, our results demonstrate that primate X palindromes have ancient origins, and have been shaped by a complex mixture of selective and non-selective forces. In Chapter 4, I consider our results in the context of revisiting the relationship between X-palindrome structures and their associated gene families. Finally, I will suggest that the increasing use of long-read technologies to generate mammalian reference genomes will continue to reveal new examples of conserved palindromes in other genomic contexts, opening up new fields of inquiry into the functions and evolution of mirrored genomic repeats.

REFERENCES

Atanassov I, Delichère C, Filatov DA, Charlesworth D, Negrutiu I, Monéger F. 2001. Analysis and evolution of two functional Y-linked loci in a plant sex chromosome system. *Mol Biol Evol* **18**: 2162–2168.

Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. 2001. Segmental duplications: Organization and impact within the current human genome project assembly. *Genome Res* **11:** 1005–1017.

Bellott DW, Cho TJ, Hughes JF, Skaletsky H, Page DC. 2018. Cost-effective high-throughput single-haplotype iterative mapping and sequencing for complex genomic structures. *Nat Protoc* **13**: 787-809.

Bellott DW, Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Cho T-J, Koutseva N, Zaghlul S, Graves T, Rock S, et al. 2014. Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* **508**: 494–499.

Bellott DW, Page DC. 2021. Dosage-sensitive functions in embryonic development drove the survival of genes on sex-specific chromosomes in snakes, birds, and mammals. *Genome Res* **31**: 198–210.

Bellott DW, Skaletsky H, Cho T-J, Brown L, Locke D, Chen N, Galkina S, Pyntikova T, Koutseva N, Graves T, et al. 2017. Avian W and mammalian Y chromosomes convergently retained dosage sensitive regulators. *Nat Genet* **49**: 387–394.

Bellott DW, Skaletsky H, Pyntikova T, Mardis ER, Graves T, Kremitzki C, Brown LG, Rozen S, Warren WC, Wilson RK, et al. 2010. Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition. *Nature* **466**: 612–616.

Bernardi G, Olofsson B, Filipski F, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F. 1985. The mosaic genome of warm-blooded vertebrates. *Science* **228:** 953–958.

Bird AP. 1980. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res* **8**: 1485–1497.

Bovine Genome Sequencing and Analysis Consortium. 2009. The genome sequence of taurine cattle: A window to ruminant biology and evolution. *Science* **324:** 522-528.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**: 525–527.

Burgoyne PS, Mahadevaiah SK, Sutcliffe MJ, Palmer SJ. 1992. Fertility in mice requires X-Y pairing and a Y-chromosomal “spermiogenesis” gene mapping to the long arm. *Cell* **3**: 391–398.

Caceres M, McDowell JC, Gupta J, Brooks S, Bouffard GG, Blakesley RW, Green ED, Sullivan RT, Thomas JW. 2007. A recurrent inversion on the eutherian X chromosome. *PNAS* **104**: 18571–18576.

Carbone L, Vessere GM, ten Hallers BFH, Zhu B, Osoegawa K, Mootnick A, Kofler A, Wienberg J, Rogers J, Humphray S, et al. 2006. A high-resolution map of synteny disruptions in gibbon and human genomes. *PLoS Genet* **2**: e223.

Carrel L, Willard HF. 2005. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434**: 400–404.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**: 1289–1303.

Charlesworth B, Charlesworth D. 2000. The degeneration of Y chromosomes. *Philos Trans R Soc Lond B Biol Sci* **355**: 1563–1572.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**: 69–87.

Chung D, Kuan PF, Li B, Sanalkumar R, Liang K, Bresnick EH, Dewey C, Keles. 2011. Discovering transcription factor binding sites in highly repetitive regions of genomes with multi-read analysis of ChIP-seq data. *PLoS Comput Biol* **7**: e1002111.

Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, Bult CJ, Agarwala R, Cherry JL, Dicuccio M et al. 2009. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol* **7**:e1000112.

Clayton D. 2008. Testing for association on the X chromosome. *Biostatistics* **9:** 593–600.

Clément Y, Arndt PF. 2011. Substitution patterns are under different influences in primates and rodents. *Genome Biol Evol* **3**: 236–245.

Cocquet J, Ellis PJI, Mahadevaiah SK, Affara NA, Vaiman D, Burgoyne PS. 2012. A genetic basis for a postmeiotic X versus Y chromosome intragenomic conflict in the mouse. *PLoS Genet* **8**: e1002900.

Cocquet J, Ellis PJI, Yamauchi Y, Mahadevaiah SK, Affara NA, Ward MA, Burgoyne PS. 2009. The multicopy gene *Sly* represses the sex chromosomes in the male mouse germline after meiosis. *PLoS Biol* **7**: e1000244.

Delgado CL, Waters PD, Gilbert C, Robinson TJ, Graves JA. 2009. Physical mapping of the elephant X chromosome: Conservation of gene order over 105 million years. *Chromosome Res* **17**: 917–926.

Dennis MY, Eichler EE. 2016. Human adaptation and evolution by segmental duplication. *Curr Opin Genet Dev* **41:** 44–52.

Disteche CM. 2012. Dosage compensation of the sex chromosomes. *Annu Rev Genet* **46**: 537–560.

Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet* **10**: 285–311.

Eichler EE. 2001. Segmental duplications: what's missing, misassigned, and misassembled—and should we care? *Genome Res* **11**: 653–656.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* **323**: 133–138.

Ellison C, Bachtrog D. 2019. Recurrent gene co-amplification on Drosophila X and Y chromosomes. *PLoS Genet* **15:** e1008251.

Eyre-Walker A. 1993. Recombination and mammalian genome evolution. *Proc R Soc Lond B* **252**: 237–243.

Felsenstein J. 1974. The evolutionary advantage of recombination. *Genetics* **78**: 737–756.

Fisher RA. 1931. The evolution of dominance. *Biol Rev* **6**: 345–368.

Fon Tacer K, Montoya MC, Oatley MJ, Lord T, Oatley JM, Klein J, Ravichandran R, Tillman H, Kim MS, Connelly JP, et al. 2019. MAGE cancer-testis antigens protect the mammalian germline under environmental stress. *Sci Adv* **5:** eaav4832.

Fridolfsson AK, Cheng H, Copeland NG, Jenkins NA, Liu HC, Raudsepp T, Woodage T, Chowdhary B, Halverson J, Ellegren H. 1998. Evolution of the avian sex chromosomes from an ancestral pair of autosomes. *Proc Natl Acad Sci* **95:** 8147–8152.

Galtier N, Piganeau G, Mouchiroud D, Duret L. 2001. GC-content evolution in mammalian genomes: The biased gene conversion hypothesis. *Genetics* **159**: 907–911.

Galtier N, Duret L, Glémin S, Ranwez V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet* **25**: 1–5.

Geraldes A, Rambo T, Wing RA, Ferrand N, Nachman MW. 2010. Extensive gene conversion drives the concerted evolution of paralogous copies of the SRY gene in European rabbits. *Mol Biol Evol* **27**: 2437–2440.

Godfrey AK, Naqvi S, Chmátal L, Chick JM, Mitchell RN, Gygi SP, Skaletsky H, Page DC. 2020. Quantitative analysis of Y-chromosome gene expression across 36 human tissues. *Genome Res* **30**: 1–14.

Graves JAM. 2006. Sex chromosome specialization and degeneration in mammals. *Cell* **124**: 901–914.

The GTEx Consortium. 2020. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**: 1318–1330.

Hallast P, Balaresqu P, Bowden GR, Ballereau S, Jobling MA. 2013. Recombination dynamics of a human Y-chromosomal palindrome: Rapid GC-biased gene conversion, multikilobase conversion tracts, and rare inversions. *PLoS Genet* **9**: e1003666.

Halldorsson BV, Hardarson MT, Kehr B, Styrkarsdottir U, Gylfason A, Thorleifsson G, Zink F, Jonasdottir A, Jonasdottir A, Sulem P, et al. 2016. The rate of meiotic gene conversion varies by sex and age. *Nat Genet* **48**: 1377–1384.

Hook EB, Warburton D. 1983. The distribution of chromosomal genotypes associated with Turner's syndrome: livebirth prevalence rates and evidence for diminished fetal mortality and severity in genotypes associated with structural X abnormalities or mosaicism. *Hum Genet* **64**: 24–27.

Hou S, Xian L, Shi P, Li C, Lin Z, Gao X. 2016. The *Magea* gene cluster regulates male germ cell apoptosis without affecting the fertility in mice. *Sci Adv* **6:** 26735.

Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Graves T, Fulton RS, Dugan S, Ding Y, Buhay CJ, Kremitzki C, et al. 2012. Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* **483**: 82–86.

Hughes JF, Skaletsky H, Pyntikova T, Graves TA, van Daalen SKM, Minx PJ, Fulton RS, McGrath SD, Locke DP, Friedman C, et al. 2010. Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* **463**: 536–539.

Hughes JF, Skaletsky H, Pyntikova T, Koutseva N, Raudsepp T, Brown LG, Bellott DW, Cho T-J, Dugan-Rocha S, et al. 2020. Sequence analysis in *Bos taurus* reveals pervasiveness of X-Y arms races in mammalian lineages. *Genome Res* **30**: 1716–1726.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431**: 931–945.

Jain M, Olsen HE, Paten B, Akeson M. 2016. The Oxford Nanopore MinION: Delivery of nanopore sequencing to the genomics community. *Genome Biol* **17**: 239.

Koopman P, Gubbay J, Vivian, N, Goodfellow, P, Lovell-Badge, R. 1991. Male development of chromosomally female mice transgenic for *Sry*. *Nature* **351:** 117–121.

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: A resource for timelines, timetrees, and divergence times. *Mol Biol Evol* **34:** 1812–1819.

Kronenberg ZN, Fiddes IT, Gordon D, Murali S, Cantsilieris S, Meyerson OS, Underwood JG, Nelson BJ, Chaisson MJP, Dougherty ML, et al. 2018. High-resolution comparative analysis of great ape genomes. *Science* **360:** eaar6343.

Kuroda-Kawaguchi T, Skaletsky H, Brown LG, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Silber S, Oates R, Rozen S, et al. 2001. The *AZFc* region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. *Nature Genet* **29**: 279–186.

Lahn B, Page DC. 1999. Four evolutionary strata on the human X chromosome. *Science* **286**: 964–967.

Lahn B, Page DC. 2000. A human sex-chromosomal gene family expressed in male germ cells and encoding variably charged proteins. *Hum Mol Genet* **9**: 311-319.

Lakich D, Kazazian HH, Antonarakis SE, Gitschier J. 1993. Inversions disrupting the factor VIII gene are a common cause of several haemophilia A. *Nature Genet* **5**: 236–241.

Lange J, Skaletsky H, van Daalen SKM, Embry SL, Cindy M, Brown LG, Oates RD, Silber S, Repping S, Page DC. 2009. Isodicentric Y chromosomes and sex disorders as byproducts of homologous recombination that maintains palindromes. *Cell* **138**: 855–869.

Lesecque Y, Mouchiroud D, Duret L. 2013. GC-biased gene conversion in yeast is specifically associated with crossovers: Molecular mechanisms and evolutionary significance. *Mol Biol Evol* **30**: 1409–1419.

Li F-W, Kuo L-Y, Pryer KM, Rothfels CJ. 2016. Genes translocated into the plastid inverted repeat show decelerated substitution rates and elevated GC content. *Genome Biol Evol* **8**: 2452–2458.

Lyon MF. 1961. Gene action in the X-chromosome of the mouse (*Mus musculus L.*). *Nature* **190**: 372–373.

Lyon MF. 1962. Sex chromatin and gene action in the mammalian X-chromosome. *Am J Hum Genet* **14**: 135–48.

Kaiser VB, Zhou Q, Bachtrog D. 2011. Nonrandom gene loss from the Drosophila miranda neo-Y chromosome. *Genome Biol Evol* **3:** 1329–1337.

Kruger AN, Brogley MA, Huizinga JL, Kidd JM, de Rooij DG, Hu Y-C, Mueller JL. 2019. A neofunctionalized X-linked ampliconic gene family is essential for male fertility and equal sex ratio in mice. *Curr Biol* **29**: 3699–3706.

Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz M. 2008. High-resolution mapping of meiotic crossovers and noncrossovers in yeast. *Nature* **454**: 479–485.

Marais G. 2003. Biased gene conversion: implications for genome and sex evolution. *Trends Genet* **19**: 330–338.

Matsubara K, Tarui H, Toriba M, Yamada K, Nishida-Umehara C, Agata K, Matsuda Y. 2006. Evidence for different origin of sex chromosomes in snakes, birds, and mammals and step-wise differentiation of snake sex chromosomes. *Proc Natl Acad Sci* **103:** 18190–18195.

McKusick-Nathans Institute of Genetic Medicine. 2020. *Online Mendelian Inheritance in Man, OMIM.* Johns Hopkins University, Baltimore, MD.

Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA, et al. 2020. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585:** 79–84.

Mohandas TK, Speed RM, Passage MB, Chen PH, Chandley AC, Shapiro LJ. 1992. Role of the pseudoautosomal region in sex-chromosome pairing during male meiosis: meiotic studies in a man with a deletion of distal Xp. *Am J Hum Genet* **51**: 526-533.

Montoya-Burgos JI, Boursot P, Galtier N. 2003. Recombination explains isochores in mammalian genomes. *Trends Genet* **19:** 128–130.

Morgan RA, Chinnasamy N, Abate-Daga D, Gros A, Robbins PF, Zheng Z, Dudley ME, Feldman SA, Yang JC, Sherry RM, et al. 2013. Cancer regression and neurological toxicity following anti- MAGE-A3 TCR gene therapy. *J Immunother* **36**: 133–151.

Mueller JL, Mahadevaiah SK, Park PJ, Warburton PE, Page DC, Turner JMA. 2008. The mouse X chromosome is enriched for multicopy testis genes showing post-meiotic expression. *Nature Genet* **40**: 794–799.

Mueller JL, Skaletsky H, Brown LG, Zaghlul S, Rock S, Graves T, Auger K, Warren WC, Wilson RK, Page DC. 2013. Independent specialization of the human and mouse X chromosomes for the male germline. *Nature Genet* **45**: 1083–1087.

Muller HJ. 1914. A gene for the fourth chromosome of Drosophila. *J Exp Zool* **17**: 325–336.

Muller HJ. 1964. The relation of recombination to mutational advance. *Mutat Res* **1**: 2–9.

Murphy WJ, Davis B, David VA, Agarwala R, Schäffer AA, Wilkerson AJP, Neelam B, O’Brian SJ, Menotti-Raymond M. 2007. A 1.5 megabase resolution radiation hybrid map of the cat genome and comparative analysis with the canine and human. *Genomics* **89**: 189–196.

Muyle A, Serres-Giardi L, Ressayre A, Escobar J, Glémin S. 2011. GC-biased gene conversion and selection affect GC content in the *Oryza* genus (rice). *Mol Bio Evol* **28:** 2695–2706.

Nanda I, Feichtinger W, Schmid M, Schröder JH, Zischler H, Epplen JT. 1990. Simple repetitive sequences are associated with differentiation of the sex chromosomes in the guppy fish. *J Mol Evol* **30**: 456–462.

Nanda I, Shan Z, Schartl M, Burt DW, Koehler M, Nothwang H, Grützner F, Paton IR, Windsor D, Dunn I, et al. 1999. 300 million years of conserved synteny between chicken Z and human chromosome 9. *Nat Genet* **21**: 258–259.

Naqvi S, Godfrey AK, Hughes JF, Goodheart ML, Mitchell RN, Page DC. 2019. Conservation, acquisition, and functional impact of sex-biased gene expression in mammals. *Science* **365:** eaaw7317.

Nei M. 1970. Accumulation of nonfunctional genes on sheltered chromosomes. *Am Nat* **104**: 311–322.

Nicholas M, Marais G, Hykelova V, Janousek B, Laporte V, Vyskot B, Mouchiroud D, Negrutiu I, Charlesworth D, Monéger F. 2004. A gradual process of recombination restriction in the evolutionary history of the sex chromosomes in dioecious plants. *PLoS Biol* **3**: e4.

Ohno S. 1967. *Sex chromosomes and sex-linked genes*. Springer, Berlin.

Ohno S, Hauschka TS. 1960. Allocycly of the X-chromosome in tumors and normal tissues. *Cancer Res* **20**: 541–545.

Odenthal-Hesse L, Berg IL, Veselis A, Jeffreys AJ, May CA. 2014. Transmission distortion affecting human noncrossover but not crossover recombination: A hidden source of meiotic drive. *PLoS Genet* **10**: e1004106.

Painter TS. 1921. The Y-chromosome in mammals. *Science* **53:** 503–504.

Quilter CR, Blott SC, Mileham AJ, Affara NA, Sargent CA, Griffin DK. 2002. A mapping and evolutionary study of porcine sex chromosome genes. *Mamm Genome* **13**: 588–594.

Raudsepp T, Lee E-J, Kata SR, Brinkmeyer C, Mickelson JR, Skow LC, Womack JE, Chowdhary BP. 2004. Exceptional conservation of horse–human gene order on X chromosome revealed by high-resolution radiation hybrid mapping. *PNAS* **101**: 2386–2391.

Rhesus Macaque Sequencing and Analysis Consortium. 2007. Evolutionary and biomedical insights from the Rhesus macaque genome. *Science* **316:** 222–234.

Rice WR. 1984. Sex chromosomes and the evolution of sexual dimorphism. *Evolution* **38:** 735–742.

Rice WR. 1987. Genetic hitchhiking and the evolution of reduced genetic activity of the Y sex chromosome. *Genetics* **116***:* 161–167.

Robbins PF, Morgan RA, Feldman SA, Yang JC, Sherry RM, Dudley ME, Wunderlich JR, Nahvi AV, Helman LJ, Mackall CL, et al. 2011. Tumour regression in patients with metastatic synovial cell sarcoma and melanoma using genetically engineered lymphocytes reactive with NY-ESO-1. *J Clin Oncol* **29**: 917–924.

Ross MT, Grafham 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.

Rozen S, Skaletsky H, Marszalek 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.

Sahin U, Oehm P, Derhovanessian E, Jabulowsky RA, Vormehr M, Gold M, Maurus D, Schwarck-Kokarakis D, Kuhn AN, Omokoko T, et al. 2020. An RNA vaccine drives immunity in checkpoint-inhibitor-treated melanoma. *Nature* **585:** 107–112.

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.

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

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown 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.

Skov L, The Danish Pan Genome Consortium, Schierup MH. 2017. Analysis of 62 hybrid assembled human Y chromosomes exposes rapid structural changes and high rates of gene conversion. *PLoS Genet*. **13**: e1006834.

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

Smeds L, Mugal CF, Qvarnström A, Ellegren H. 2016. High-resolution mapping of crossover and non-crossover recombination events by whole-genome re-sequencing of an avian pedigree. *PLoS Genet* **12**: e1006044.

Smith JM, Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genet Res* **23:** 23–35.

Soh YQS, Alfoldi J, Pyntikova T, Brown LG, Minx PJ, Fulton RS, Kremitzki 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.

Stankiewicz P, Lupski JR. 2010. Structural variation in the human genome and its role in disease. *Annu Rev Med* **61**: 437–455.

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.

Teitz LS, Pyntikova T, Skaletsky H, Page DC. 2018. Selection has countered high mutability to preserve ancestral copy number of Y chromosome amplicons in diverse human lineages. *Am J Hum Genet* **103**: 261–275.

Treangen TJ, Salzberg SL. 2012. Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nat Rev Genet* **13**: 36–46.

Vogt PH, Edelmann A, Kirsch S, Henegariu 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.

Wang PJ, McCarrey JR, Yang F, Page DC. 2001. An abundance of X-linked genes expressed in spermatogonia. *Nature Genet* **27**: 422–426.

Warburton PE, Giordano J, Cheung F, Gelfand Y, Benson G. 2004. Inverted repeat structures 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.

Warren WC, Harris RA, Haukness M, Fiddes 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.

Williams AL, Genovese G, Dyer T, Altemose N, Truax K, Jun G, Patterson N, Myers SR, Curran JE, Duggirala R, et al. 2015 Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *eLife* **4**: e04637.

Wise AL, Gyi L, Manolio TA. 2013. eXclusion: Toward integrating the X chromosome in genome-wideassociation analyses. *Am J Hum Genet* **92:** 643–647.

Zheng G, Joo J, Zhang C, Geller NL. 2007. Testing association for markers on the X chromosome. *Genet Epidemiol* **31**: 834–843.

Zhou Q, Bachtrog D. 2012. Sex-specific adaptation drives early sex chromosome evolution in Drosophila. *Science* **337:** 341–345.

Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, Hanrahan F, Pertea G, Van Tassell CP, Sonstegard TS, et al. 2009. A whole-genome assembly of the domestic cow, *Bos Taurus. Genome Biol* **10**: R42.

CHAPTER 2:

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

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

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.

Acknowledgements
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. Brunning, M. Jamehdor, B. Shapiro, K. Monaghan, J.C. Petrozza, L. Oman-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 and Arthur W. and Carol Tobin Brill.

Adapted from: Jackson E, Bellott DW, Cho T-J, Skaletsky H, Hughes JF, Pyntikova T, and Page DC (in revision at *Genome Research*). 2021. Large palindromes on the primate X Chromosome are preserved by natural selection. *bioRxiv* doi: https://doi.org/10.1101/2020.12.29.424738.

ABSTRACT

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 non-coding sequence, demonstrating that natural selection has preserved these gene families. Unexpectedly, the spacers that separate the left and right arms of palindromes are a site of localized structural instability, with 7 of 12 conserved palindromes showing no spacer orthology between human and rhesus macaque. Analysis of the 1000 Genomes Project dataset 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.

INTRODUCTION

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 gene families 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 palindrome 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 non-human 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, Hughes et al. 2012, Soh et al. 2014, Hughes et al. 2020), were sequenced using labor-intensive but highly accurate clone-based approaches (Kawaguchi-Kuroda 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 *TSPY* 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 non-human 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 kmer-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

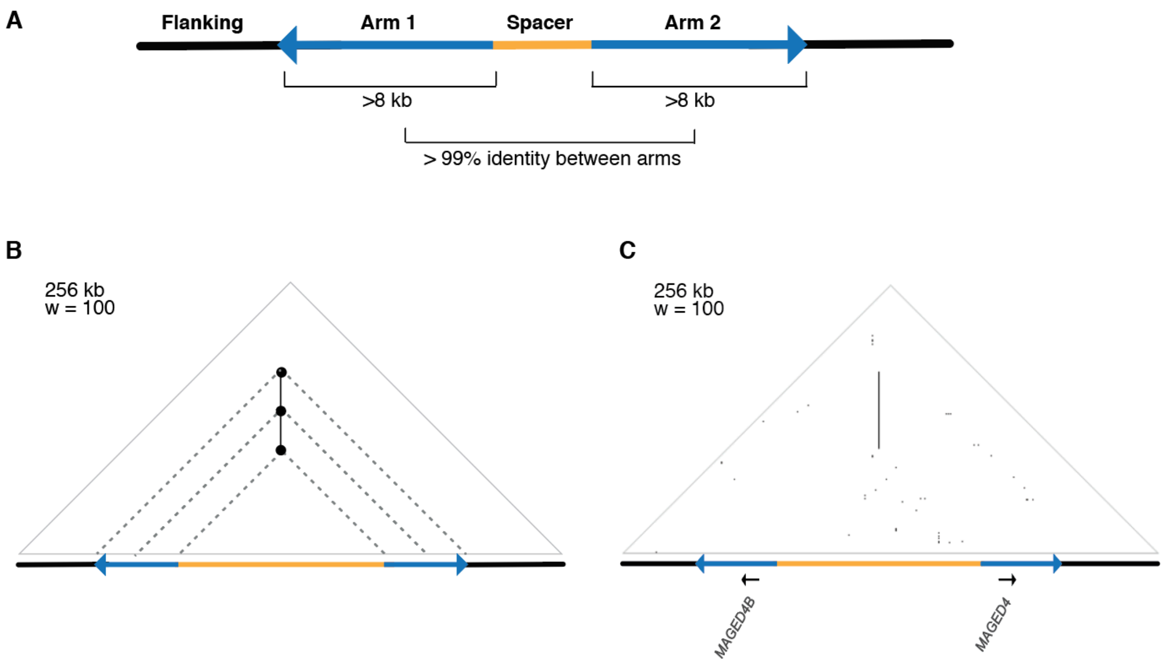
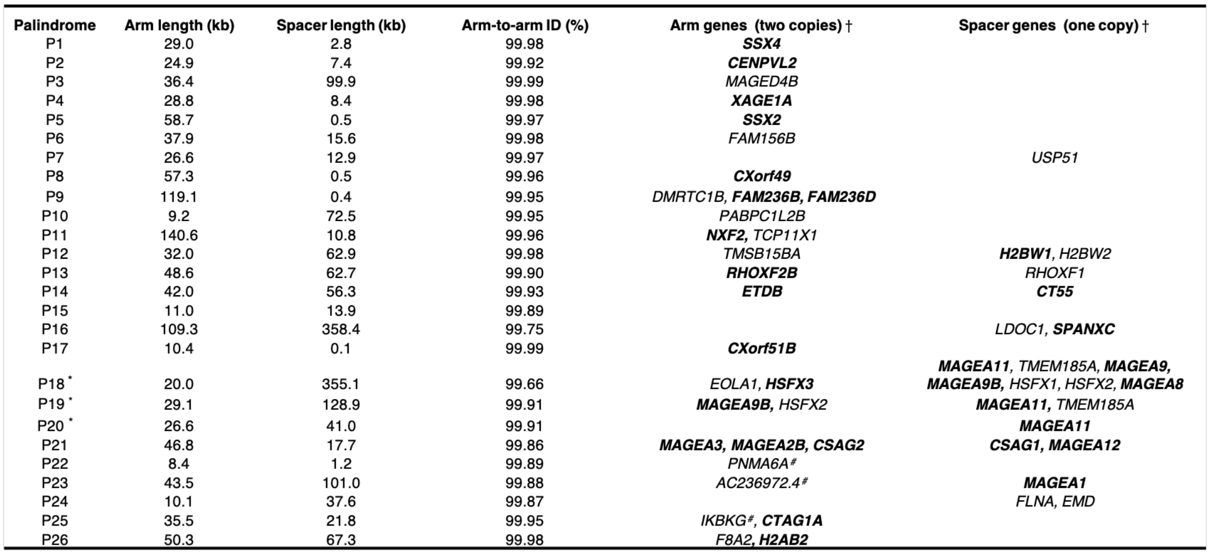


Figure 1. Overview of human X-Chromosome palindromes. A) Schematic of a palindrome. B) Schematic of a triangular dot plot. Dots are placed at a 90-degree angle between identical kmers, or "words," within a DNA sequence. Palindromes appear as vertical lines. "w": word size used to construct the dot plot. C) Triangular dot plot for human X palindrome P3, including annotated protein-coding genes.

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. No primate Y palindromes have been observed with protein-coding genes in their spacers (Skaletsky et al. 2003, Hughes et al. 2010, Hughes et al. 2012), but we found that 11 of 26 human X palindromes also contain at least one protein-coding gene in their spacer.

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

Table 1. Palindromes on the human X Chromosome



Palindrome	Arm length (kb)	Spacer length (kb)	Arm-to-arm ID (%)	Arm genes (two copies) †	Spacer genes (one copy) †
P1	29.0	2.8	99.98	***SSX4***	
P2	24.9	7.4	99.92	***CENPVL2***	
P3	36.4	99.9	99.99	*MAGED4B*	
P4	28.8	8.4	99.98	***XAGE1A***	
P5	58.7	0.5	99.97	***SSX2***	
P6	37.9	15.6	99.98	*FAM156B*	
P7	26.6	12.9	99.97		*USP51*
P8	57.3	0.5	99.96	***CXorf49***	
P9	119.1	0.4	99.95	*DMRTC1B*, ***FAM236B, FAM236D***	
P10	9.2	72.5	99.95	*PABPC1L2B*	
P11	140.6	10.8	99.96	***NXF2,*** *TCP11X1*	
P12	32.0	62.9	99.98	*TMSB15BA*	***H2BW1***, *H2BW2*
P13	48.6	62.7	99.90	***RHOXF2B***	*RHOXF1*
P14	42.0	56.3	99.93	***ETDB***	***CT55***
P15	11.0	13.9	99.89		
P16	109.3	358.4	99.75		*LDOC1*, ***SPANXC***
P17	10.4	0.1	99.99	***CXorf51B***	
P18 *	20.0	355.1	99.66	*EOLA1*, ***HSFX3***	***MAGEA11***, *TMEM185A*, ***MAGEA9, MAGEA9B,*** *HSFX1*, *HSFX2*, ***MAGEA8***
P19 *	29.1	128.9	99.91	***MAGEA9B,*** *HSFX2*	***MAGEA11,*** *TMEM185A*
P20 *	26.6	41.0	99.91		***MAGEA11***
P21	46.8	17.7	99.86	***MAGEA3, MAGEA2B, CSAG2***	***CSAG1, MAGEA12***
P22	8.4	1.2	99.89	*PNMA6A#*	
P23	43.5	101.0	99.88	*AC236972.4#*	***MAGEA1***
P24	10.1	37.6	99.87		*FLNA*, *EMD*
P25	35.5	21.8	99.95	*IKBKG#*, ***CTAG1A***	
P26	50.3	67.3	99.98	*F8A2*, ***H2AB2***	

* P20 is found within the spacer of P19, which is found within the spacer of P18. See Supplemental Fig. S4.

# Genes annotated as protein-coding in one arm, but as a pseudogene in the other. See Supplemental Note S1.

† Genes 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. AC236972.4 was not expressed (<2 TPM in all tissues). All other genes are expressed but not testis-biased.

(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 multi-mapping 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 multi-mapping 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 timepoints, 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). 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 palindrome arms, compared to 58.1% of flanking sequence ($p<0.05$, Mann-Whitney U). The density of transposable elements is negatively correlated with recombination rates across mammalian genomes, which may reflect the increased efficiency of natural selection in removing mildly deleterious insertions (Jensen-Seaman et al. 2004). We conclude that transposable elements are depleted both in X and Y palindromes in humans, and that this likely results from elevated recombination rates in palindrome arms (Rozen et al. 2003).

Generation of high-resolution X-palindrome reference sequence for chimpanzee and rhesus macaque

To understand the origins of human X palindromes, we searched for orthologous palindromes in two non-human primates, the chimpanzee and the rhesus macaque. Existing reference genomes for chimpanzee and rhesus were not generated using a clone-based approach comparable to that for the human genome, so to address this limitation, we generated 14.43 Mb of non-redundant high-quality reference sequence for portions of the chimpanzee and rhesus X Chromosomes that are orthologous to human palindromes (Fig. 2A). We used the recently developed SHIMS 3.0 method, a clone-based approach that incorporates both Illumina and nanopore reads (Bellott et al. 2020). This method provides high structural confidence due to the generation of full-length nanopore reads spanning each clone: For 83 of 107 clones sequenced using

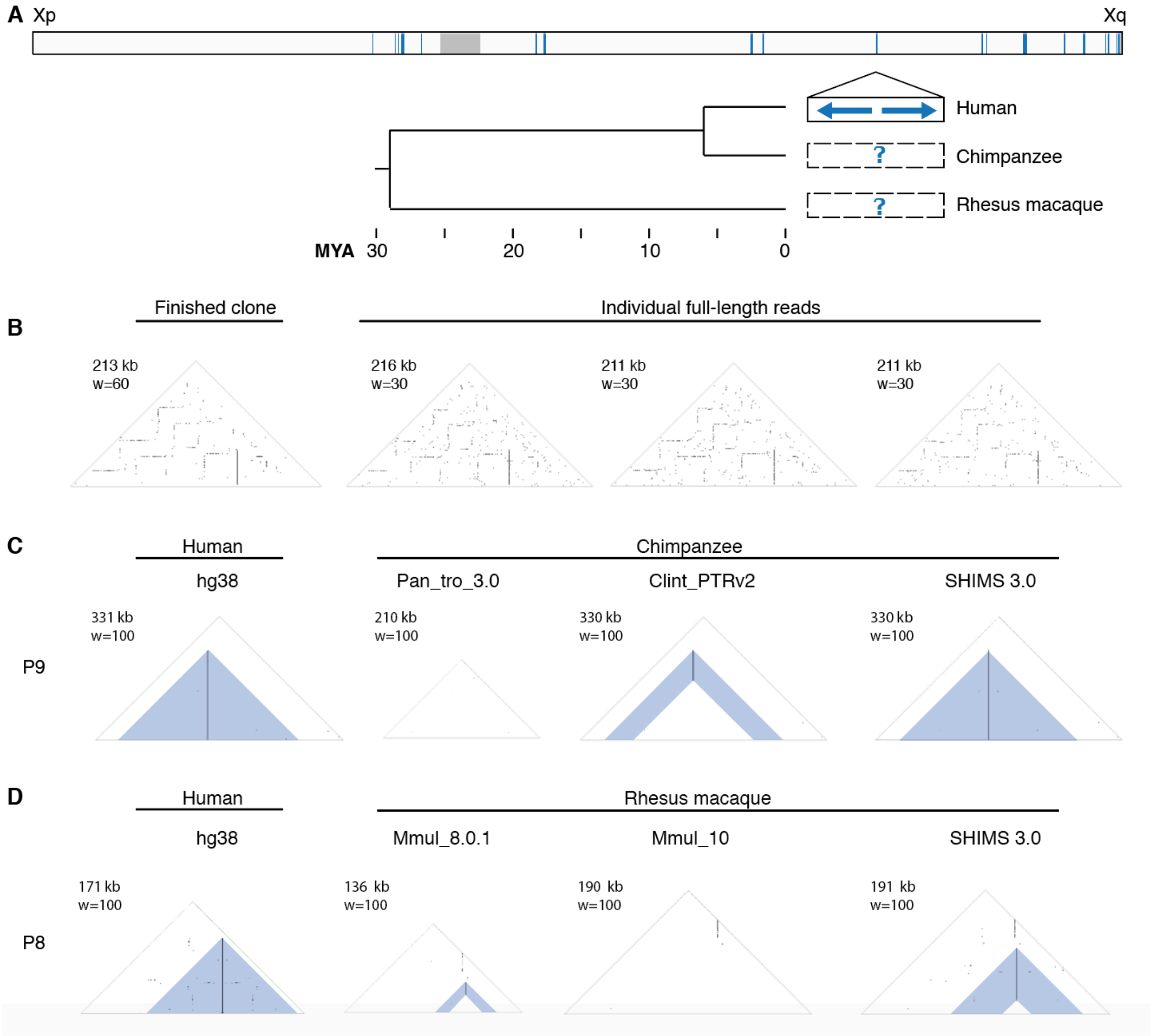


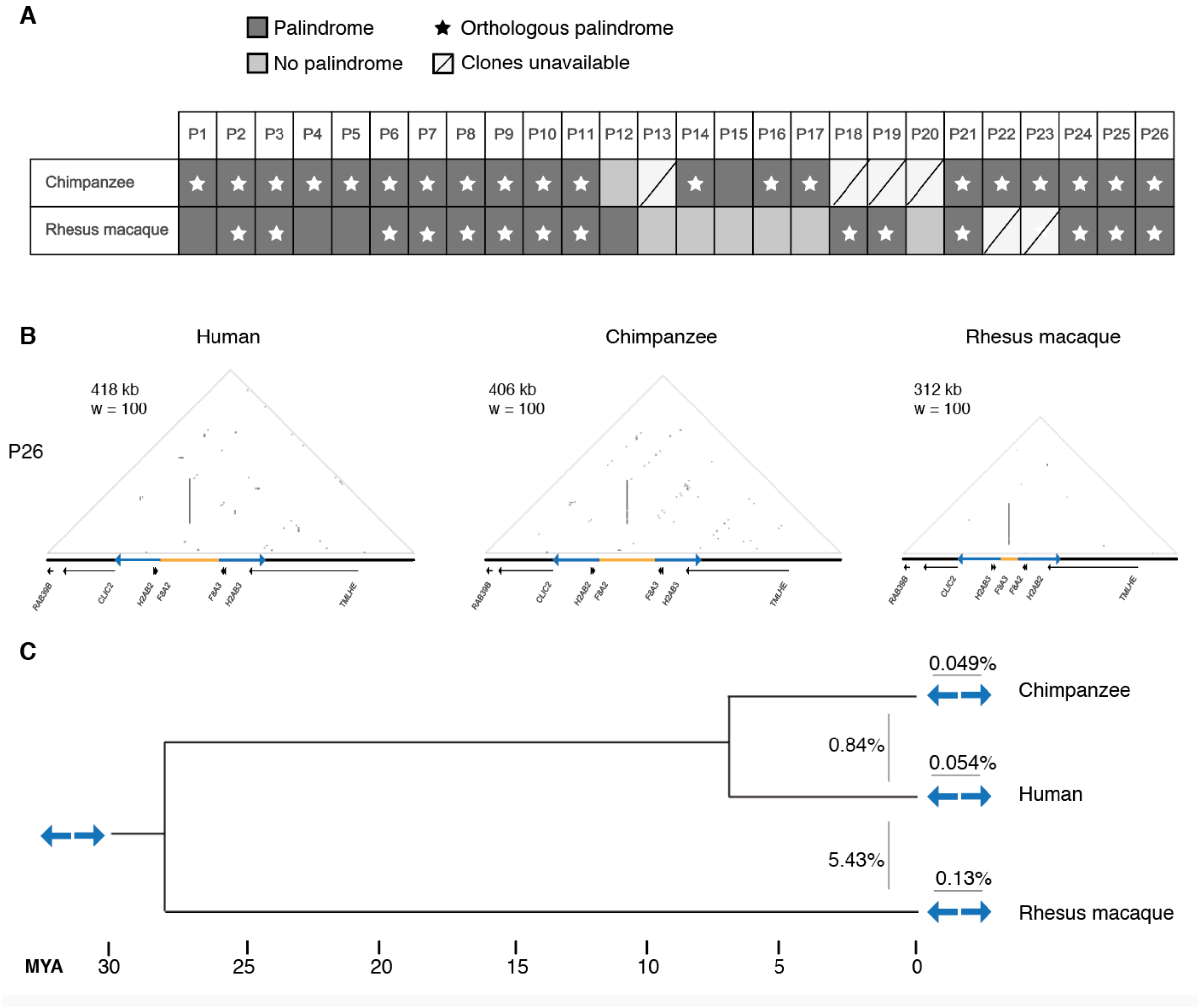
Figure 2. Improvements to prior reference assemblies for chimpanzee and rhesus macaque. A) Sequencing approach. Top bar: Locations of 26 human X palindromes (blue bands). Gray band shows centromere location. Below: Expansion of a single region containing a human X palindrome (solid black box). One or more clones were selected to span orthologous regions in chimpanzee and rhesus macaque (dashed black boxes). Tree shows estimated divergence times from TimeTree (Kumar et al. 2017). B) Full-length nanopore reads supporting the structure of a single finished chimpanzee clone (CH251-385I8). C,D) Two primate X palindromes resolved using SHIMS 3.0 that were missing or misassembled in existing X-Chromosome assemblies. D) Triangular dot plots from the region orthologous to human P9 in chimpanzee assemblies Pan_tro_3.0 (Kuderna et al. 2017), Clint_PTRv2 (Kronenberg et al. 2018), and SHIMS 3.0. D) Triangular dot plots from the region orthologous to human P8 in rhesus macaque assemblies Mmul_8.0.1 (Zimin et al. 2014), Mmul_10 (Warren et al. 2020), and SHIMS 3.0.

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 S2).

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 (WGS) approaches (chimpanzee, Pan_tro_3.0; rhesus macaque, Mmul_8.0.1) (Fig. 2C,D; Supplemental Fig. S3; Supplemental Tables S3,S4) (Kuderna et al. 2017, Zimin et al. 2014). We also compared our SHIMS 3.0 assemblies to two assemblies generated using long-read WGS approaches incorporating PacBio reads (chimpanzee, Clint_PTRv2; rhesus macaque, Mmul_10) (Kronenberg et al. 2018, Warren et al. 2020). Here, we found that while 18 palindromes (46.2%) were represented accurately, 21 palindromes (53.8%) remained missing or incomplete (Fig. 2C,D; Supplemental Fig. S3; Supplemental Tables S3, S4). Palindromes that were missing or incomplete in existing long-read WGS assemblies had longer arms than palindromes that were represented accurately (Clint_PTRv2: 63 kb versus 20 kb; Mmul_10: 65 kb versus 22 kb) ($p<0.01$ for both comparisons, Mann-Whitney U), suggesting that large palindromes remain particularly intractable to whole-genome shotgun approaches (Supplemental Tables S3, S4). We therefore carried out all subsequent analyses using our newly generated SHIMS 3.0 assemblies.

Conservation of X palindromes in two non-human 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).



	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	P22	P23	P24	P25	P26
Chimpanzee	★	★	★	★	★	★	★	★	★	★	★	□	⧄	★	■	★	★	⧄	⧄	⧄	★	★	★	★	★	★
Rhesus macaque	■	★	★	■	■	★	★	★	★	★	★	■	□	□	□	□	□	★	★	□	★	⧄	⧄	★	★	★

Figure 3. Conservation of X-Chromosome palindromes across primates. A) Conservation status of 26 human palindromes in chimpanzee and rhesus macaque. B) Triangular dot plots for a palindrome (P26) conserved between human, chimpanzee, and rhesus macaque. Images are to scale. C) Arm-to-arm divergence within species (above blue arrows) versus between species (left of blue arrows). Values are the average percent divergence across 12 palindromes shared by human, chimpanzee, and macaque. Divergence times estimated using TimeTree (Kumar et al. 2017).

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 non-human 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 (Skaletsky et al. 2003, Rozen 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 versus chimpanzee and human versus 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 macaque to further analyses to uncover the processes governing their evolution and unexpectedly deep conservation.

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 non-allelic 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, while 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).

Unexpectedly, 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

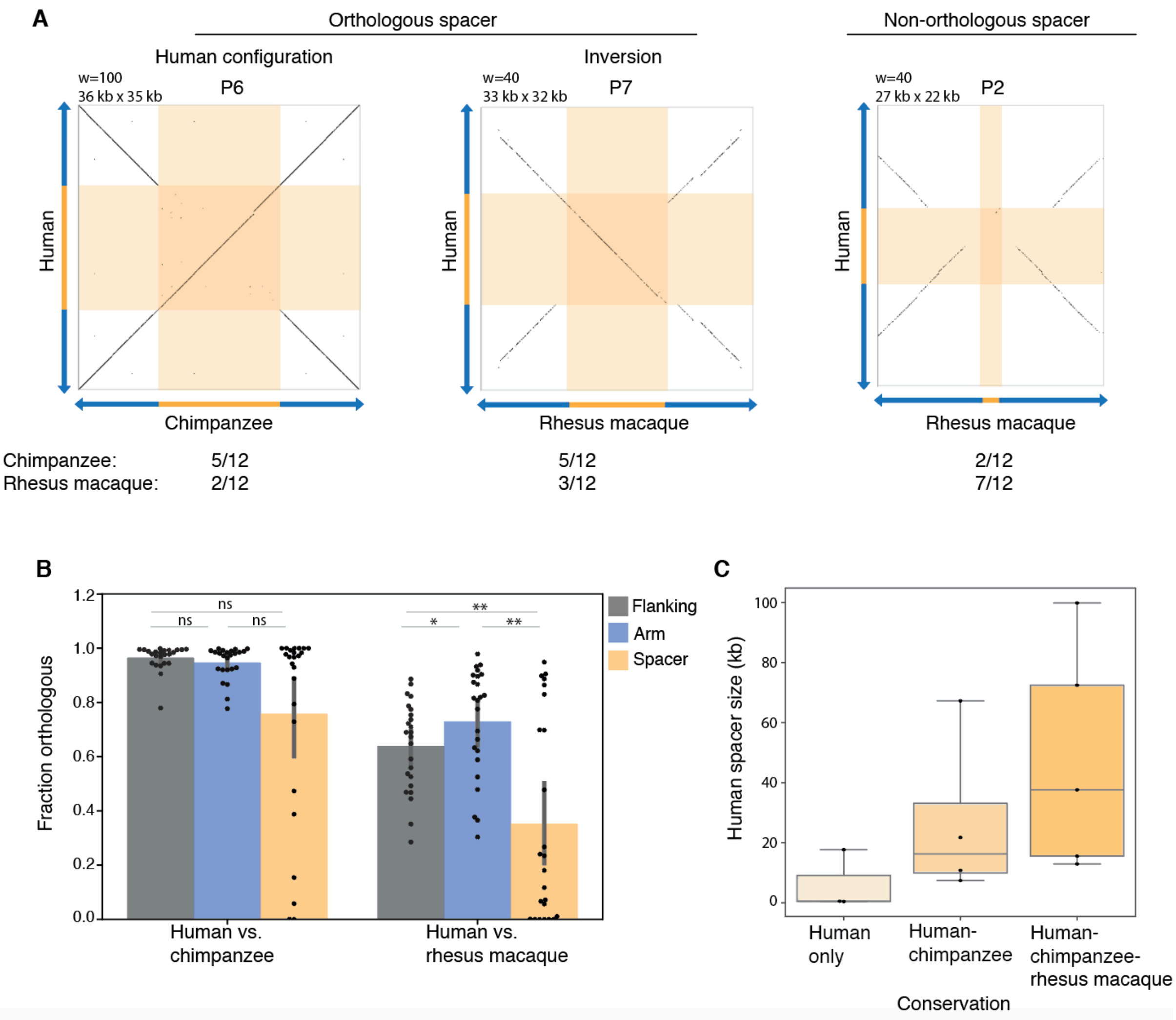


Figure 4. Structural changes between orthologous X-chromosome palindromes are concentrated around the center of symmetry. A) Square dot plots comparing the center of the palindrome, including the spacer and 10 kb of inner arm sequence on each side, between the indicated species. "Orthologous spacer": > 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. C) Sizes of human spacers, binned according to the species between which they are conserved.

Fig. S6). This phenomenon was observed in 2 of 12 spacer comparisons between human and chimpanzee, and 7 of 12 spacer comparisons between human and rhesus macaque (Fig. 4A). Non-orthologous 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 versus 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) (Fig. 4B). Consistent with our observations in Fig. 4A, these results were driven by a subset of palindromes in 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, palindromes 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 5 protein-coding genes in palindrome spacers (Supplemental Table S5). The functions of these gene families are poorly characterized in humans: Only 1 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 four-fold 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 ($p<0.05$, hypergeometric test). Out of 5 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 S5). Among spacer genes, 4 of 5 have at least one intact gene copy in chimpanzee, and 3 of 5 in rhesus macaque; the two genes with ascribed OMIM phenotypes are conserved in all three species (Supplemental Table S5). 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 non-synonymous to synonymous substitution rates (dN/dS) was below 1.0 for 17 of 18 arm and spacer gene families conserved between all three species (Supplemental Table S6), twelve of which were significant using a likelihood ratio test (Supplemental Table S6). Nonetheless, the median dN/dS value for 18 X-palindrome gene families is 0.36, compared to a median dN/dS value of 0.12 for protein-coding genes in the genome (Gayà-Vidal and Albà 2014, Biswas et al. 2016). Elevated dN/dS 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 2 gene families (Supplemental Table S6). We note that with only three species for comparison, we were likely under-powered 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 kmer-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 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 dataset 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 palindrome 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

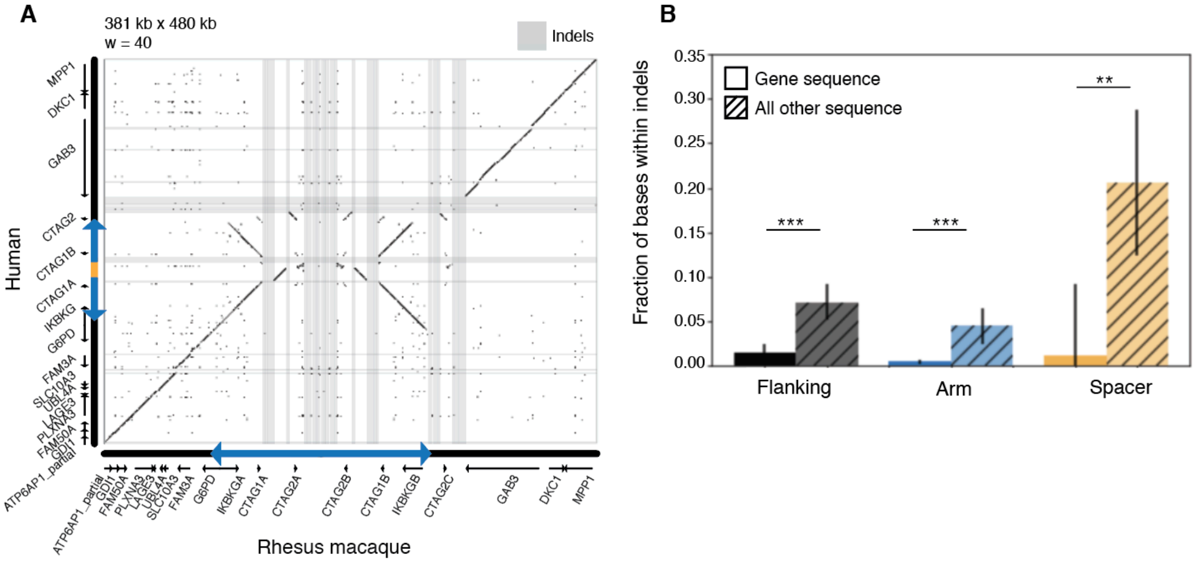


Figure 5. Natural selection has preserved X-palindrome gene families. Results based on 12 palindromes conserved between human, chimpanzee and rhesus macaque. A) Square dot plot comparing structure of P25 between human and rhesus macaque. Indels highlighted in gray; nearly all fall between protein-coding 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.

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 remarkable total of 149 palindrome spacer deletions across 9 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 Dataset (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

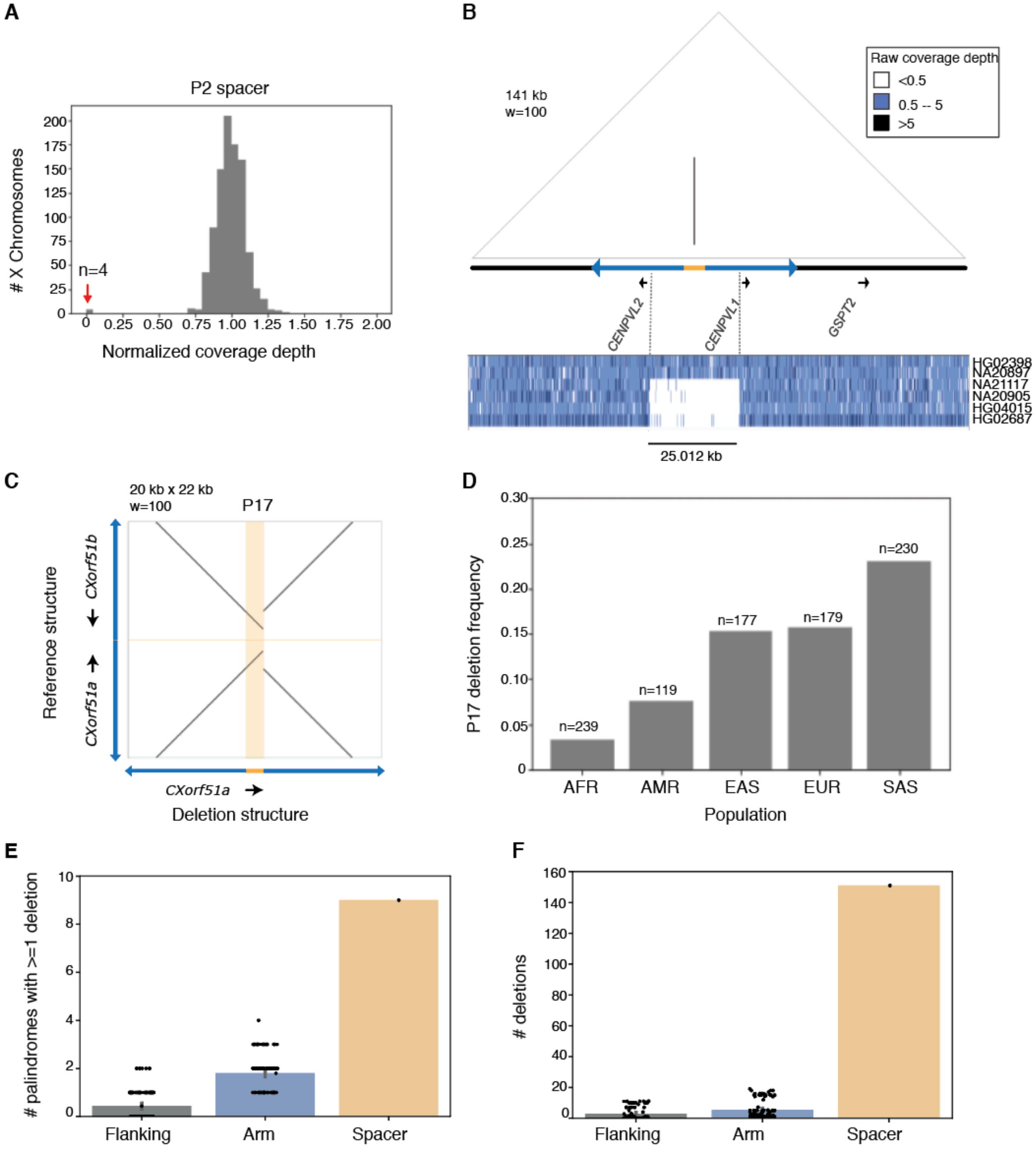
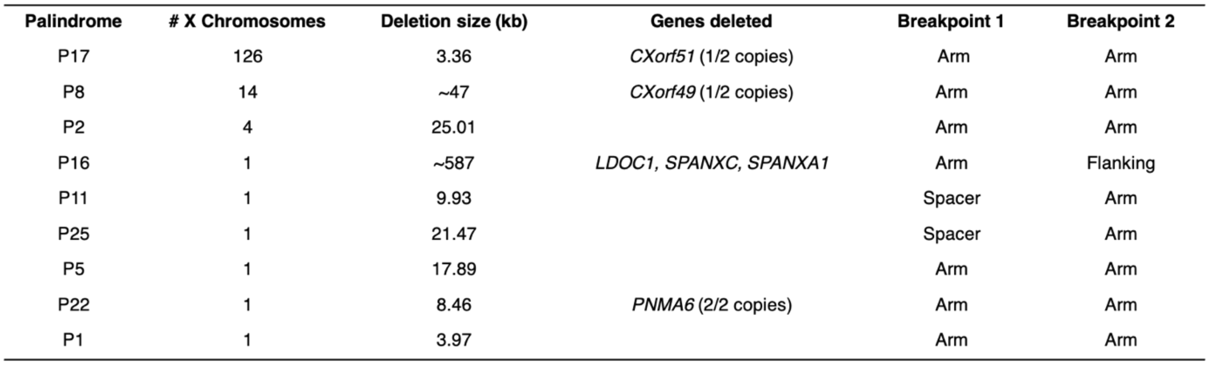


Figure 6. Deletions are enriched in human X-Chromosome spacers. A) Normalized coverage depths for P2 spacer. Red arrow indicates four X Chromosomes with depth near 0. B) Coverage depths across P2 and flanking sequence for two individuals with reference structure (HG02398, NA20897) and four with spacer deletions (NA21117, NA20905, HG04015, HG02687). C) Square dot plot comparing palindrome centers (spacer + 10 kb inner arm on

each side) for P17 reference structure and P17 deletion. D) Frequency of P17 spacer deletions across five superpopulations from 1000 Genomes. EUR = European, AFR = African, AMR = Admixed Americas, EAS = East Asian, SAS = South Asian. E,F) Frequency of deletions detected in palindrome spacers compared to palindrome arms and flanking sequence. Size-matched regions from palindrome arms and single-copy sequence were selected at random; results from 100 iterations are shown.

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.

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 *CXorf51b*, a testis-expressed gene not associated with any phenotypes reported in OMIM (Table 1). The P17 spacer deletion is found in all five super-populations in 1000 Genomes, with frequencies ranging from 3% (Africa) to 23% (South Asia) of X Chromosomes (Fig. 6D). The 1000 Genomes dataset 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 9 different palindromes; in contrast, we never observed more than 4 deletions in size-matched regions from palindrome arms and flanking sequence ($p<0.01$, bootstrapping analysis) (Fig. 6E). The difference was more dramatic 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. Importantly, 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). Since 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 dataset containing capture-based targeted sequencing for 301 azoospermia cases and 300 normospermic controls (dbGaP ID: phs001023).

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 exact test) (Supplemental Table S7). Interestingly, 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 562 oligozoospermic men (sperm counts 0.1–20 million per cubic cm). We found *CXorf51* deletions in 68 men (12.1%), which is not significantly different from the percentage of men from the 1000 Genomes Project (13.3%, Fisher exact test), suggesting that *CXorf51* deletions are not enriched in oligozoospermic men (Supplemental Table S7). 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% versus 1.4%, $p<0.05$, Fisher exact test). While 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 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). While the chimpanzee Y chromosome contains nineteen 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 macaque have clear orthology between arms and between flanking sequences, unambiguously establishing a common origin. Notably, each of these palindromes is found in a single copy on the X chromosome, distinguishing them from previous reports of co-amplified gene families on the X and Y chromosomes of Drosophila (Ellison and Bachtrog 2019), mouse (Cocquet et al. 2009, Cocquet et al. 2012, Soh et al. 2014), and bull (Hughes et al 2020). Our data provides strong evidence that palindromes can be maintained over tens of millions of years of evolution, in some cases with minimal structural change (e.g., P6; Supplementary Fig. 4). We note that 25 million years represents a lower bound on the age of the X-Chromosome palindromes, and that future high-resolution sequencing of mammalian X Chromosomes may reveal that these palindromes are conserved in more distantly related species.

Primate X palindromes could be conserved because they are inherently stable structures, or because they are preserved by natural selection. 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, 7 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. Notably, all twelve 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 crosslinking 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, 4 of 9 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 (*PNMA6*: 2/2 copies; *LDOC1:* 1/1 copies). We speculate that deletions that remove one or more X-palindrome genes 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, Aradhya 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, Carvalho et al. 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 (Akgun et al. 1997). While 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). While 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 than the previous human genome assembly, yet was still missing nearly 20% of segmental duplications and other hard-to-sequence regions of the genome that had been previously sequenced using large-insert clones (Miga et al. 2020).

Going forward, we propose that 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 while 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.

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 bioinformatics tools, such as those used for quantifying gene expression or detecting mutations, automatically discard multi-mapping 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 multi-mapping 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 palindromes 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-base-pair windows, then aligned these windows back to the candidate region using bowtie2, with settings to return up to 10 alignments with alignment scores >-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 Interactive Genome Viewer (IGV) (Robinson et al. 2011). Putative palindromes boundaries were annotated manually based on the start and end of long stretches of multi-mapping 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. 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 as *TMSB15BA* and *TMSB15BB*, respectively.

2. The palindrome arm genes *PNMA6B* and *AC152010.1* 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).

3. 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 re-normalized 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 SRA files from Jan et al. 2017 (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://bacpacresources.org), which derives from a single male individual ("Clint") used in initial sequencing of the chimpanzee genome (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 (Bellott 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 7 chimpanzee clones previously sequenced and deposited in 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 S8).

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 (CM009261.2) 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

Primate gene annotations were performed manually using alignment of human exons and alignment of testis RNA-Seq reads for guidance. For each SHIMS 3.0 assembly from chimpanzee and rhesus macaque, we aligned exons from the corresponding human genomic region using BLAT (Kent 2002), and verified that splice sites were conserved (acceptor: AG; donor: GT or GC). In instances where splice sites were not conserved, we aligned testis RNA-Seq from the appropriate species. If we identified reads supporting the existence of an alternative splice site, we selected the alternative splice site; otherwise, we selected the original position and annotated the transcript as a pseudogene. In a subset of cases, part or all of the transcript fell into a problem area supported by nanopore reads but not Illumina reads. In these instances, gene annotations were modified using alignment of testis RNA-Seq and/or whole-genome sequencing (WGS) Illumina reads from a single male of the appropriate species. Chimpanzee testis RNA-Seq: SRR2040591, Rhesus macaque testis RNA-Seq: SRR2040595. Chimpanzee WGS Illumina: SRR490084 and SRR490117; Rhesus macaque WGS Illumina: SRR10693566.

Primate gene expression

The latest transcriptomes for chimpanzee and rhesus macaque were downloaded from ensembl.org (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 datasets containing at least 5 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 non-human 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 non-human primate palindrome arm was required to align to the putative human ortholog. Second, the alignable portion of the human palindrome arm was BLASTed against the complete non-human 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 versus human; >1 kb, 90% for rhesus macaque versus 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 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 (OMIM) 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 dN/dS

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. dN/dS values were calculated using the basic model in PAML (model = 0, NSsites = 0) (Yang 2007). To test the significance of calculated dN/dS values, we compared the likelihood of calculated values against a model where dN/dS 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 kmers with step size=1 and aligned to orthologous sequence from the other species using bowtie2 with settings to return up to 10 alignments with alignment scores >-11. Kmer 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 kmers from the other species.

1000 Genomes data analysis

We analyzed whole-genome sequencing data from 1225 males from the 1000 Genomes Project (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 dataset reads (1000genomes.sequence.index), we used a deeper 1000 Genomes dataset (1000G_2504_high_coverage.sequence.index).

PCR verification of human palindrome spacer deletions

Patient genomic DNAs were purchased from Coriell Cell Repositories (HG01872, HG02070, HG02398, HG02687, HG03295, HG04015, HG04219, NA11919, NA18645, NA19086, NA19652, NA20351, NA20897, NA20905, NA21116, NA21117, NA21133). DNAs were tested for the presence or absence of palindrome spacers using primer pairs described in Supplemental Table S9. 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 s), 61°C (30 s), 72°C (1 min) for 35 cycles. Long range PCR was performed using Advantage 2 Polymerase following the manufacturer's protocol (Clontech Laboratories, Mountain View CA).

Association of *CXorf51* and *CXorf49* deletions with azoospermia

Palindrome spacer deletions removing one copy of *CXorf51* or *CXorf49* in dbGAP dataset phs001023 were detected based on reduced coverage depth, as described above for 1000 Genomes, with modifications as follows. Dataset phs001023 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 2 targeted 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* deletions with oligozoospermia

We analyzed 562 DNA samples from oligozoospermic men previously collected by our lab. We excluded samples from men with Y chromosome deletions, varicocele, undescended testicles, or other known risk factors for oligozoospermia. 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 S9).

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 GenBank (https://www.ncbi.nlm.nih.gov/) under accession numbers AC280414 through AC280580 (Supplemental Table S10). Codes for replicating these analyses are included in Supplemental Code as well as on GitHub (https://github.com/ejackson054/primate-X-palindromes).

REFERENCES

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526**: 68–74.

Alkan C, Sajjadian S, Eichler EE. 2011. Limitations of next-generation genome sequencing assembly. *Nat Methods* **8**: 61–65.

Akgun E, Zahn J, Baumes S, Brown G, Liang F, Romanienko PJ, Lewis S, Jasin M. 1997. Palindrome resolution and recombination in the mammalian germline. *Mol Cell Biol* **17**: 5559–5570.

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.

Aradhya S, Bardaro T, Galgoczy 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 35kb duplication involving the NEMO and LAGE2 genes. *Hum Mol Genet* **10**: 2557–2567.

Bellott DW, Cho T-J, Jackson E, Skaletsky H, Hughes JF, Page DC. 2020. Highly efficient single-haplotype iterative mapping and sequencing using ultra-long nanopore reads. *bioRxiv* doi: https://doi.org/10.1101/2020.09.18.303735.

Bellott DW, Cho TJ, Hughes JF, Skaletsky H, Page DC. 2018. Cost-effective high-throughput single-haplotype iterative mapping and sequencing for complex genomic structures. *Nat Protoc* **13**: 787-809.

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**: 643–650.

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.

Biswas K, Chakraborty S, Podder S, Ghosh TC. 2016. Insights into the dN/dS ratio heterogeneity between brain specific genes and widely expressed genes in species of different complexity. *Genomics* **108:** 11–17.

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. 2011. The evolution of gene expression in mammalian organs. *Nature* **478:** 373–348.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**: 525–527.

Caceres M, McDowell JC, Gupta J, Brooks S, Bouffard GG, Blakesley RW, Green ED, Sullivan RT, Thomas JW. 2007. A recurrent inversion on the eutherian X chromosome. *PNAS* **104**: 18571–18576.

Carvalho CMB, Zhang F, Liu P, Patel A, Sahoo T, Bacino CA, Shaw C, Peacock S, Pursley A, Tavyev YJ, et al. 2009. Complex rearrangements in patients with duplications of MECP2 can occur by fork stalling and template switching. *Hum Mol Genet* **18**: 2188–2203.

Carvalho CMB, Ramocki MB, Pehlivan D, Franco LM, Gonzaga-Jauregui C, Fang P, McCall A, Pivnick EK, Hines-Dowell S, Seaver L, et al. 2011. Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. *Nat Genet* **43**: 1074–1082.

The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**: 69–87.

Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, Bult CJ, Agarwala R, Cherry JL, Dicuccio M et al. 2009. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol* **7**: e1000112.

Clapham KR, Yu TW, Ganesh VS, Barry B, ChanY, Mei D, Parrini E, Funalot B, Dupuis L, Nezarati MM, et al. 2012. *FLNA* genomic rearrangements cause periventricular nodular heterotopia. *Neurology* **78**: 269–278.

Cocquet J, Ellis PJI, Mahadevaiah SK, Affara NA, Vaiman D, Burgoyne PS. 2012. A genetic basis for a postmeiotic X versus Y chromosome intragenomic conflict in the mouse. *PLoS Genet* **8**: e1002900.

Cocquet J, Ellis PJI, Yamauchi Y, Mahadevaiah SK, Affara NA, Ward MA, Burgoyne PS. 2009. The multicopy gene Sly represses the sex chromosomes in the male mouse germline after meiosis. *PLoS Biol* **7**: e1000244.

Eichler EE. 2001. Segmental duplications: what's missing, misassigned, and misassembled—and should we care? *Genome Res* **11**: 653–656.

Ellison C, Bachtrog D. 2019. Recurrent gene co-amplification on Drosophila X and Y chromosomes. *PLoS Genet* **15:** e1008251.

Fox JW, Lamperti ED, Ekşioğlu YZ, Hong SE, Feng Y, Graham DA, Scheffer IE, Dobyns WB, Hirsch BA, Radtke RA. 1998. Mutations in Filamin 1 Prevent Migration of Cerebral Cortical Neurons in Human Periventricular Heterotopia. *Neuron* **21**: 1315–1325.

Gayà-Vidal M, Albà MM. 2014. Uncovering adaptive evolution in the human lineage. *BMC Genomics* **15:** 599.

Godfrey AK, Naqvi S, Chmátal L, Chick JM, Mitchell RN, Gygi SP, Skaletsky H, Page DC. 2020. Quantitative analysis of Y-Chromosome gene expression across 36 human tissues. *Genome Res* **30**: 1–14.

Gordon D, Huddleston J, Chaisson MJP, Hill CM, Kronenberg ZN, Munson KM, Malig M, Raja A, Fiddes I, Hillier LDW, et al. 2016. Long-read sequence assembly of the gorilla genome. *Science* **352**: aae0344.

The GTEx Consortium. 2017. Genetic effects on gene expression across human tissues. *Nature* **550**: 204–213

Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Graves T, Fulton RS, Dugan S, Ding Y, Buhay CJ, Kremitzki C, et al. 2012. Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* **483**: 82–86.

Hughes JF, Skaletsky H, Pyntikova T, Graves TA, van Daalen SKM, Minx PJ, Fulton RS, McGrath SD, Locke DP, Friedman C, et al. 2010. Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* **463**: 536–539.

Hughes JF, Skaletsky H, Pyntikova T, Koutseva N, Raudsepp T, Brown LG, Bellott DW, Cho T-J, Dugan-Rocha S, et al. 2020. Sequence analysis in *Bos taurus* reveals pervasiveness of X-Y arms races in mammalian lineages. *Genome Res* **30**: 1716–1726.

Jan SZ, Vormer TL, Jongejan A, Röling MD, Silber SJ, de Rooij DG, Hamer G, Repping S, van Pelt AMM. 2017. Unraveling transcriptome dynamics in human spermatogenesis. *Development* **144:** 3659–3673.

Jensen-Seaman MI, Furey TS, Payseur BA, Lu Y, 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**: 528–538

Kent WJ. 2002. BLAT–the BLAST-like alignment tool. *Genome Res* **12**: 656–664.

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.

Kronenberg ZN, Fiddes IT, Gordon D, Murali S, Cantsilieris S, Meyerson OS, Underwood JG, Nelson BJ, Chaisson MJP, Dougherty ML, et al. 2018. High-resolution comparative analysis of great ape genomes. *Science* **360:** eaar6343.

Kruger AN, Ellison Q, Brogley MA, Gerlinger ER, Mueller JL. 2018. Male mice with large inversions or deletions of X-chromosome palindrome arms are fertile and express their associated genes during post-meiosis. *Sci Rep* **8**: 8985.

Kuderna LFK, Tomlinson C, Hillier LW, Tran A, Fiddes IT, Armstrong J, Laayouni H, Gordon D, Huddleson J, Perez RG, et al. 2017. A 3-way hybrid approach to generate a new high-quality chimpanzee reference genome (Pan tro 3.0). *GigaScience* **6**: 1–6.

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.

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: A resource for timelines, timetrees, and divergence times. *Mol Biol Evol* **34:** 1812–1819.

Kuroda-Kawaguchi T, Skaletsky H, Brown LG, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Silber S, Oates R, Rozen S, et al. 2001. The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. *Nat Genet* **29**: 279–186.

Lakich D, Kazazian HH, Antonarakis SE, Gitschier J. 1993. Inversions disrupting the factor VIII gene are a common cause of several haemophilia A. *Nat Genet* **5**: 236–241.

Lange J, Skaletsky H, van Daalen SKM, Embry SL, Cindy M, Brown LG, Oates RD, Silber S, Repping S, Page DC. 2009. Isodicentric Y chromosomes and sex disorders as byproducts of homologous recombination that maintains palindromes. *Cell* **138**: 855–869.

Low WY, Tearle R, Bickhart DM, Rosen BD, Kingan SB, Swale T, Thibaud-Nissen F, Murphy TD, Young R, Lefevre L, et al. 2019. Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity. *Nat Commun* **10**: 1–11.

Merkin J, Russell C, Chen P, Burge CB. 2012. Evolutionary dynamics of gene and isoform regulation mammalian tissues. *Science* **338:** 1593–1599.

McKusick-Nathans Institute of Genetic Medicine. 2020. *Online Mendelian Inheritance in Man, OMIM*. Johns Hopkins University, Baltimore, MD.

Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA, et al. 2020. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585:** 79–84.

Mueller JL, Mahadevaiah SK, Park PJ, Warburton 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.

Mueller JL, Skaletsky H, Brown LG, Zaghlul S, Rock S, Graves T, Auger K, Warren WC, Wilson RK, Page DC. 2013. Independent specialization of the human and mouse X chromosomes for the male germline. *Nat Genet* **45**: 1083–1087.

Ohno S. 1967. *Sex chromosomes and sex-linked genes*. Springer, Berlin.

Porubsky D, Sanders AD, Höps W, Hsieh PH, Sulovari A, Li R, Mercuri L, Sorensen M, Murali SC, Gordon D, et al. 2020. Recurrent inversion toggling and great ape genome evolution. *Nat Genet* **52:** 849–858.

Rice WR. 1984. Sex chromosomes and the evolution of sexual dimorphism. *Evolution* **38:** 735–742.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nature Biotech* **29**: 24–26.

Ross MT, Grafham 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.

Rozen S, Skaletsky H, Marszalek 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.

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.

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

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown 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.

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

Soh YQS, Alfoldi J, Pyntikova T, Brown LG, Minx PJ, Fulton RS, Kremitzki 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.

Stankiewicz P, Lupski JR. 2010. Structural variation in the human genome and its role in disease. *Annu Rev Med* **61:** 437–455.

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 human X chromosome. *BMC Genomics* **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

Teitz LS, Pyntikova T, Skaletsky H, Page DC. 2018. Selection has countered high mutability to preserve ancestral copy number of Y chromosome amplicons in diverse human lineages. *Am J Hum Genet* **103**: 261–275.

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.

Vogt PH, Edelmann A, Kirsch S, Henegariu 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.

Warburton PE, Giordano J, Cheung F, Gelfand Y, Benson G. 2004. Inverted repeat structures 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.

Warren WC, Harris RA, Haukness M, Fiddes 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.

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586–1591.

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.

SUPPLEMENTAL FIGURES AND TABLES

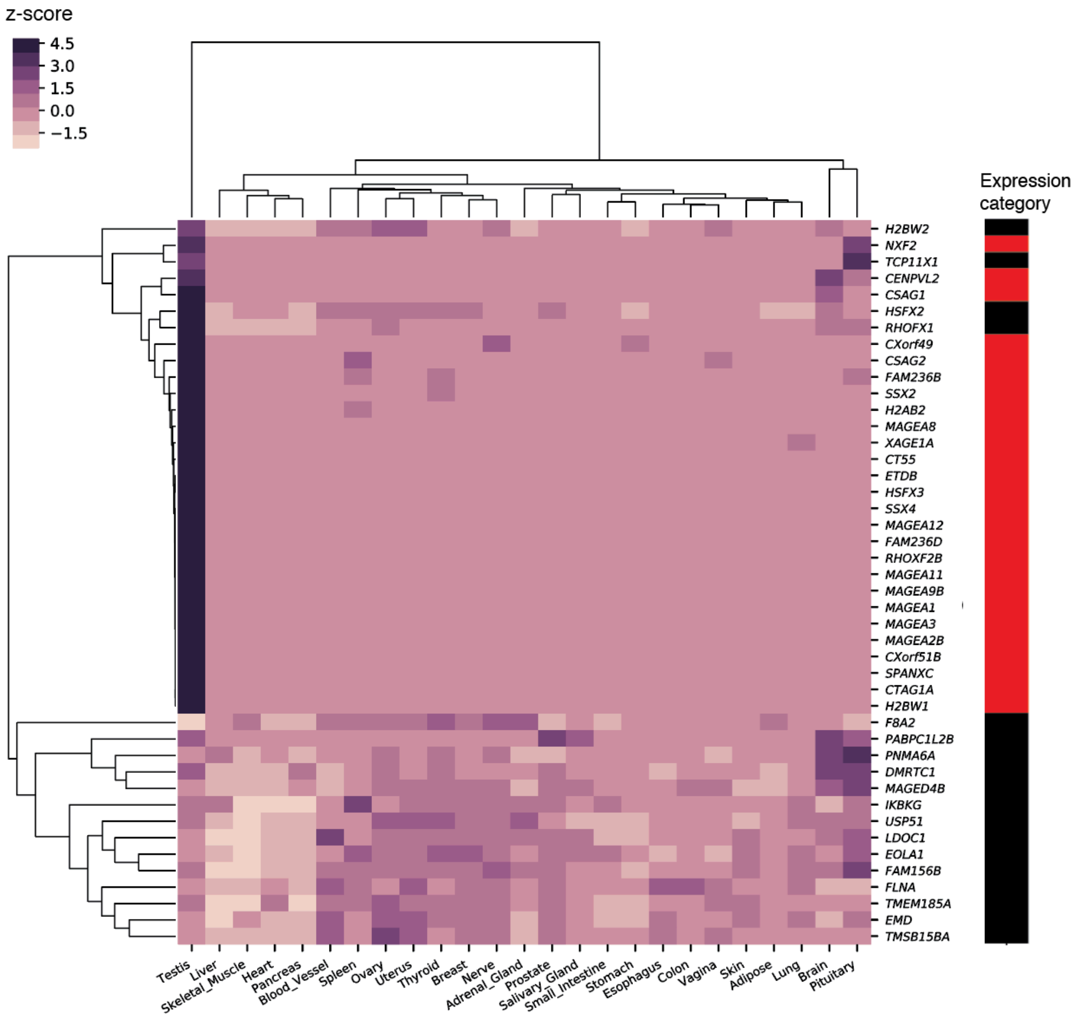
Supplemental Note 1: Treatment of human X-palindrome genes with conflicting annotations

There were three instances in which a gene in one arm of a palindrome was designated as protein-coding while the homologous sequence in the other arm was designated a pseudogene: *IKBKG* (protein-coding) and *IKBKGP1* (unprocessed pseudogene); *PNMA6A* (protein-coding) and *PNMA6B* (unprocessed pseudogene), and *AC236972.4* (protein-coding) and *AC152010.1* (processed pseudogene). We decided whether to include gene copies marked as pseudogenes in downstream analyses, i.e., whether their expression should be averaged with that of the corresponding protein-coding gene, as follows:

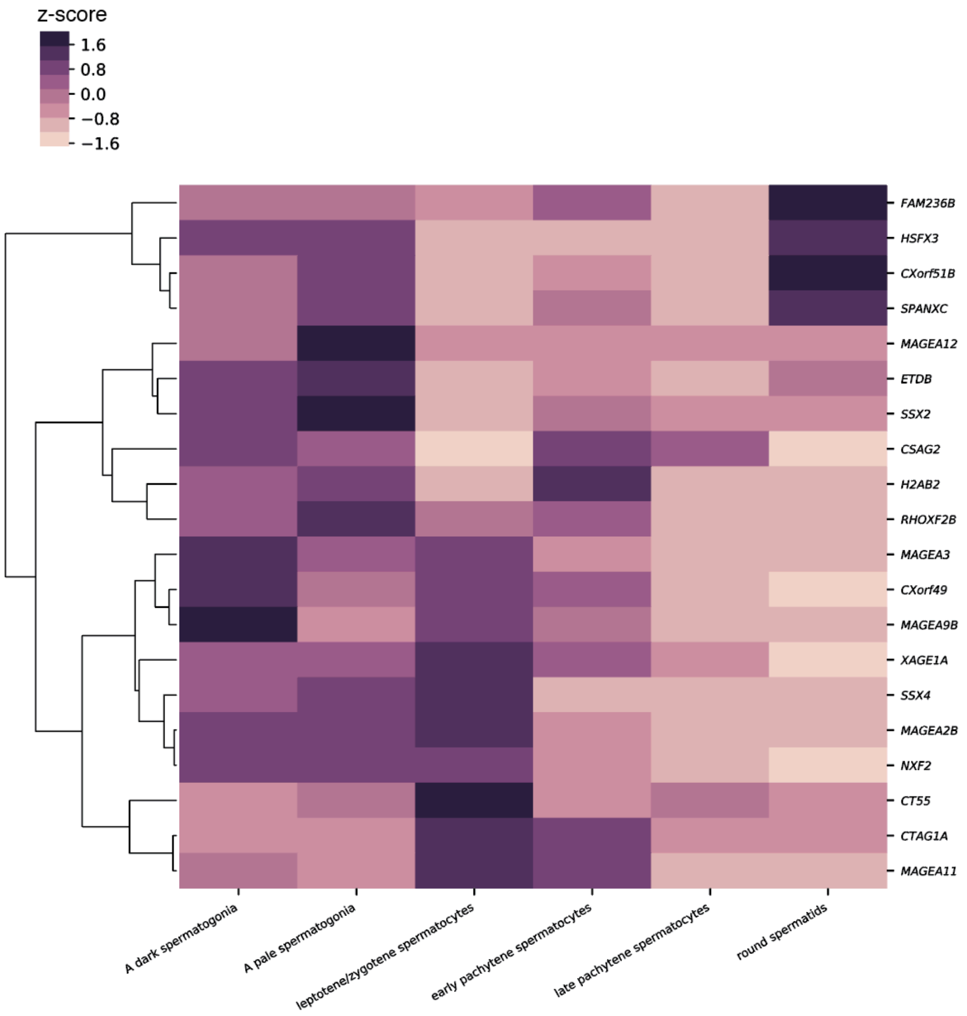
1) *PNMA6A* encodes a protein of 399 amino acids. *PNMA6A* and *PNMA6B* differ in their coding sequence by only a single missense substitution. The 3' UTR of *PNMA6B* is truncated, but the significance of this is unclear. Given that *PNMA6B* encodes an intact protein-coding sequence, we chose to include *PNMA6B* in downstream analyses.

2) *AC236972.4* encodes a protein of 2061 amino acids. *AC152010.1* has a nonsense substitution, but contains a downstream start codon that would lead to translation of the terminal 1253 amino acids of *AC236972.4.* Given that *AC152010.1* encodes a protein encompassing more than half the length of the original protein, we chose to include *AC152010.1* in downstream analyses.

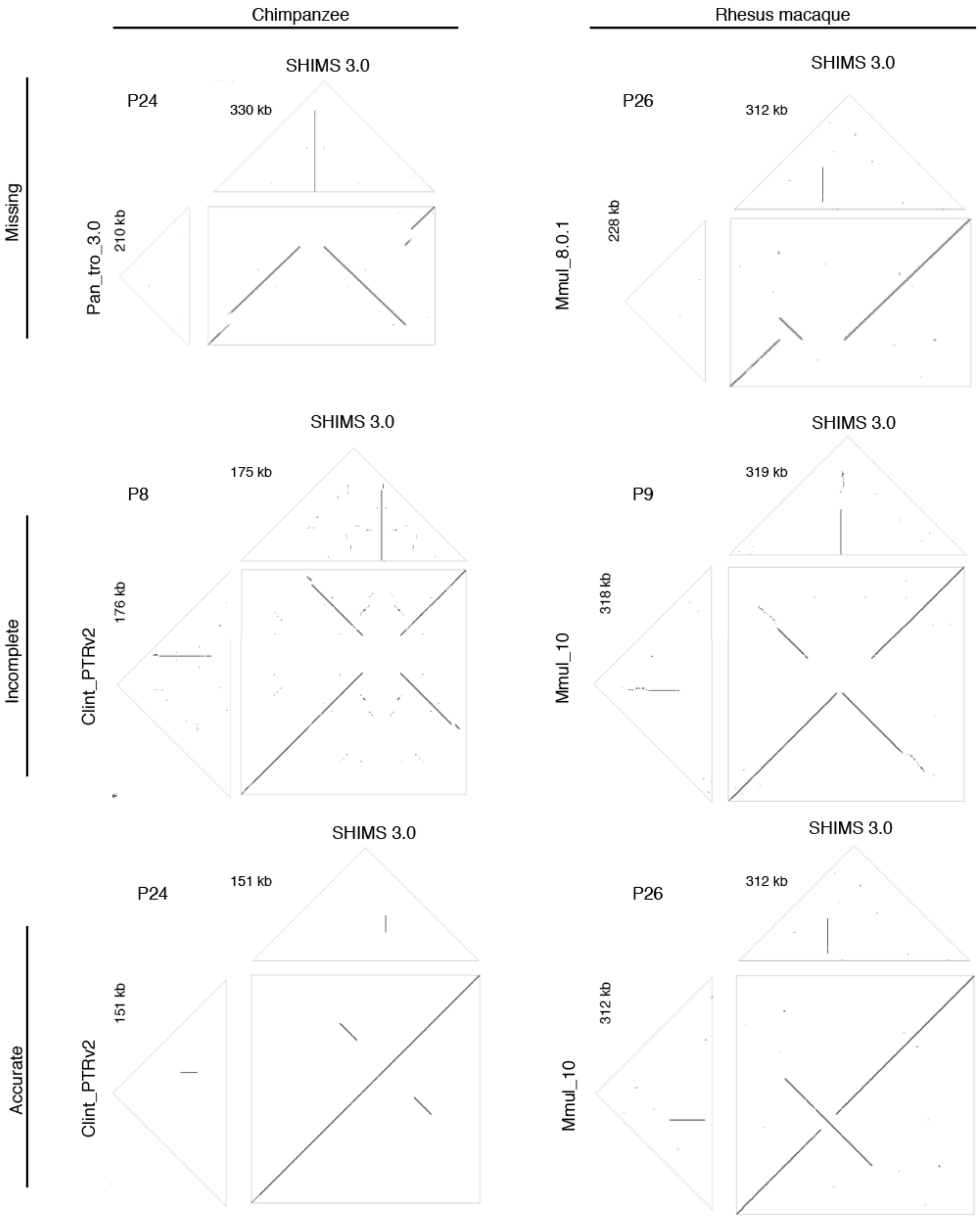
3) *IKBKG* encodes a protein of 419 amino acids. *IKBKGP1* is a well-characterized pseudogene lacking the promoter and first four exons of *IKBKG* (Aradhya et al. 2001); we therefore chose not to include it in downstream analyses.



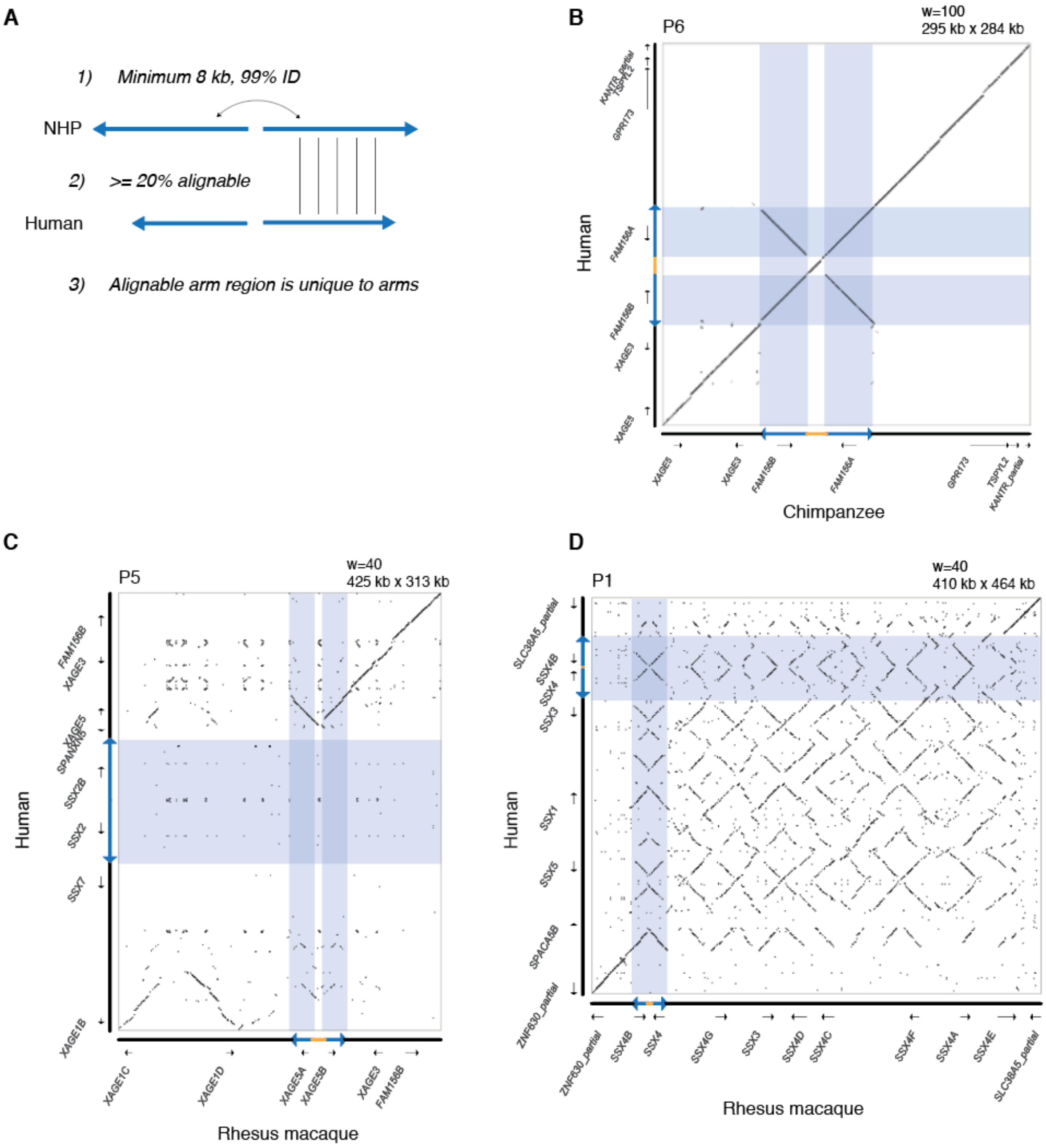
Supplemental Figure S1. Expression of human X-palindrome gene families. Heatmap shows z-score for each expressed gene (>2 TPM in at least one tissue) across 25 tissues from GTEx, with row and column order determined by hierarchical clustering. Expression category: Shows whether expression is testis-biased (red) or broad (black). Testis-biased: Minimum 2 TPM in testis, and testis accounts for >25% of log2 normalized expression summed across all tissues. Broad: All other expressed genes.



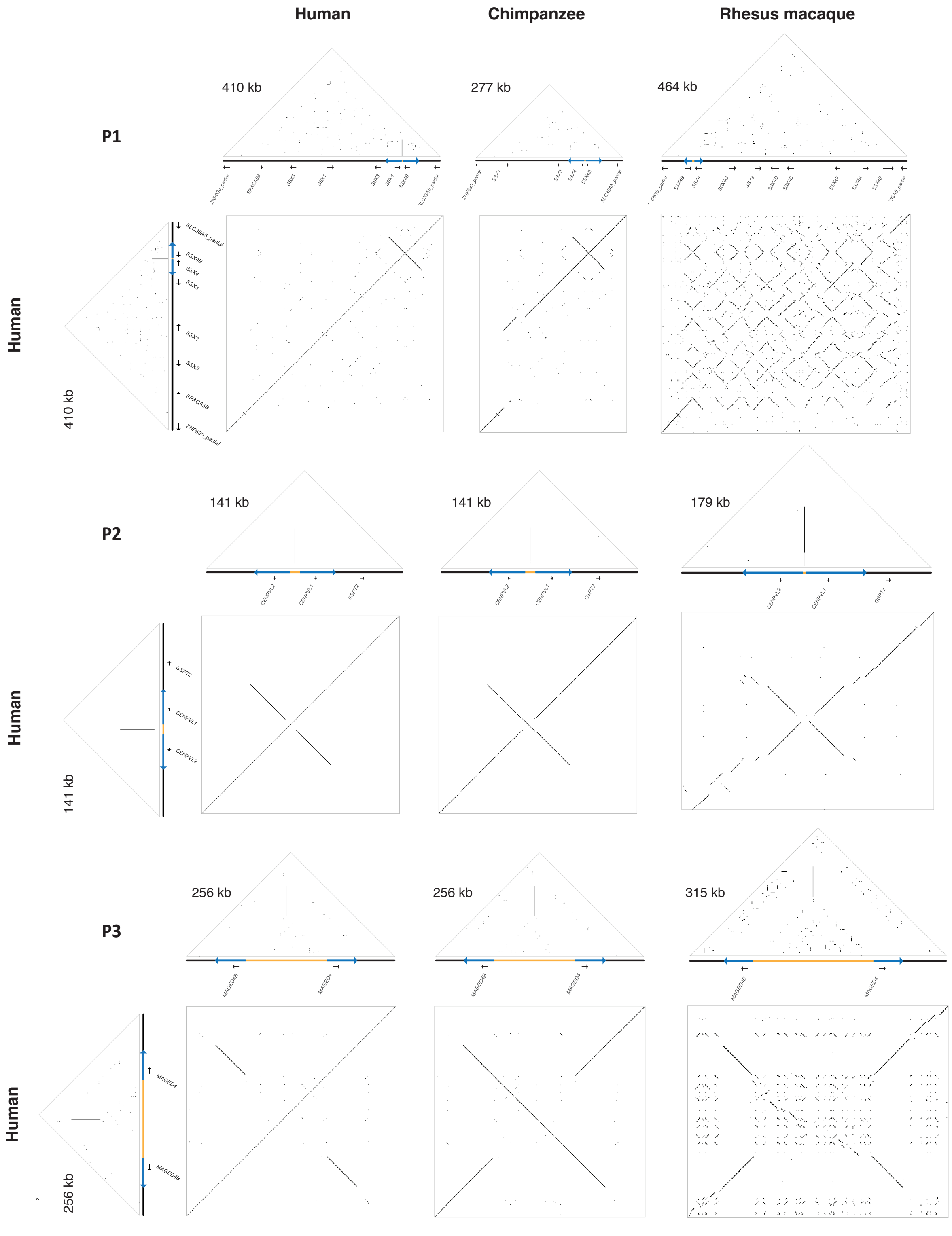
Supplemental Figure S2. Expression of human testis-biased X-palindrome gene families during spermatogenesis. Heatmap shows z-score for each expressed gene (>2 TPM in at least one spermatogenic stage) across six spermatogenic stages from Jan et al. 2017, with row order determined by hierarchical clustering.



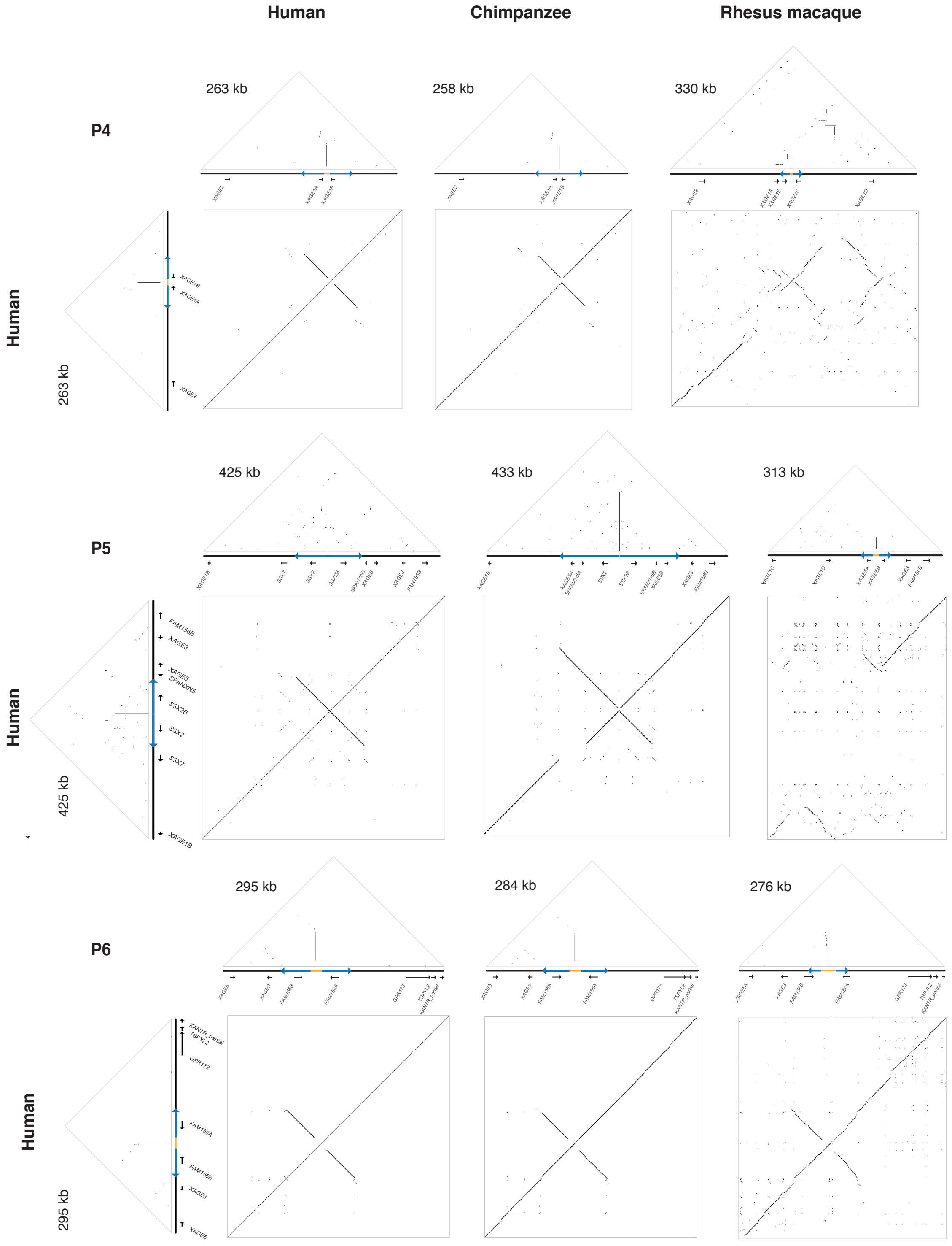
Supplemental Figure S3. Structural comparisons between palindromes in SHIMS 3.0 assemblies and existing X-Chromosome assemblies. Missing: No palindrome present in non-SHIMS 3.0 assembly. Incomplete: Part of palindrome present in non-SHIMS 3.0 assembly. Accurate: Full palindrome present in non-SHIMS 3.0 assembly. For accurate palindrome assemblies, note the presence in the square dot plot of an uninterrupted diagonal line and two arms that each map twice to the SHIMS 3.0 assembly. w=100 for all triangular and square dot plots.



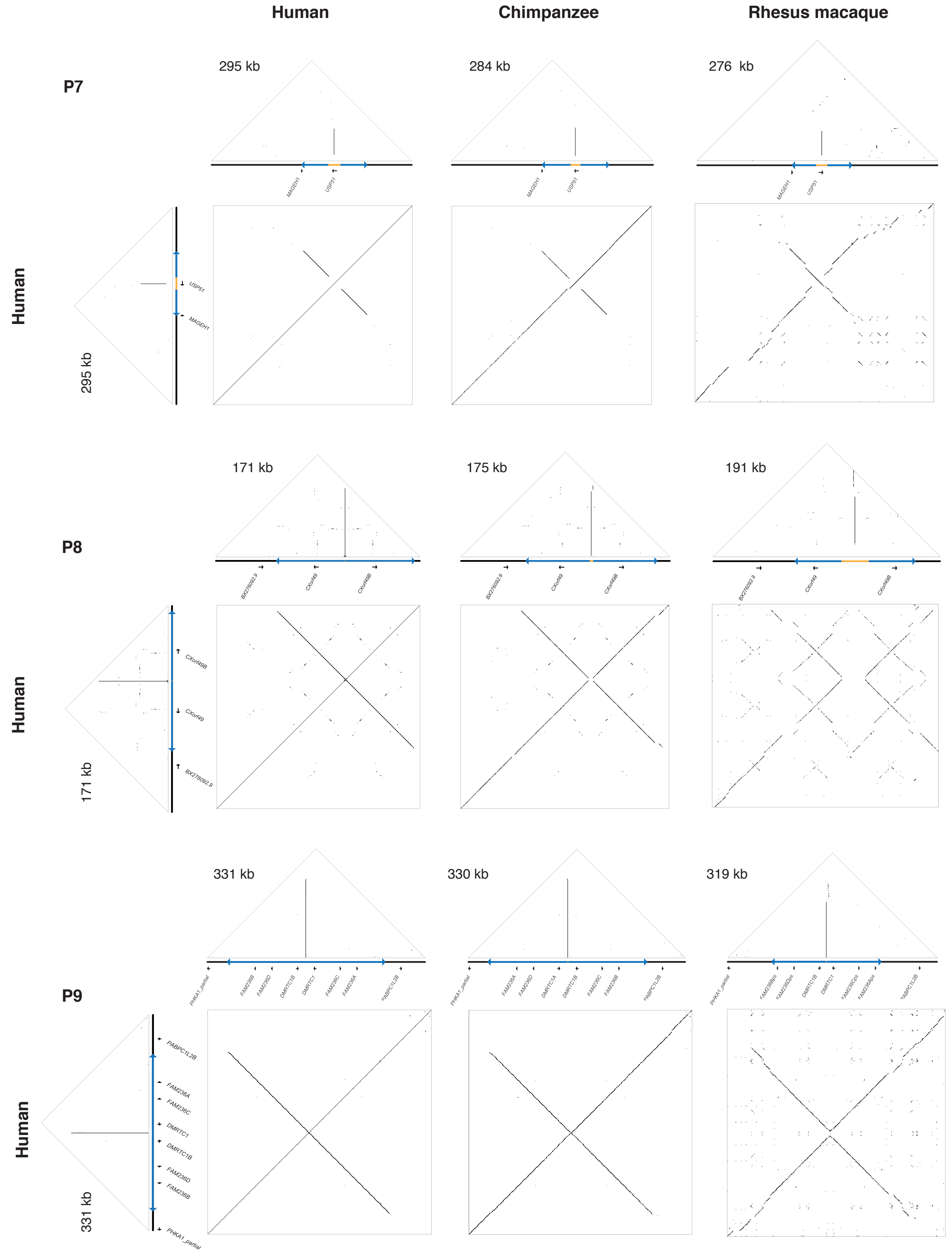
Supplemental Figure S4. Definition of orthologous palindromes. a) Criteria for defining orthologous palindromes. NHP = non-human primate (chimpanzee or rhesus macaque). Human and NHP palindrome arms were aligned with ClustalW, and required to have at least 20% alignment between species. Palindromes were excluded if the alignable region between palindrome arms mapped equally well to flanking sequence using reciprocal BLAST hits (>10% positions in high-quality hits mapping outside of palindrome arms). b) Example of an orthologous palindrome. Human palindrome arms map exactly twice to chimpanzee palindrome arms, and vice versa. c, d) Examples of palindromes that are not orthologous. c) Human palindrome arms have no orthologous sequence in rhesus macaque. Rhesus macaque palindrome arms correspond to flanking sequence in human. d) Rhesus macaque palindrome arms correspond equally well to more than two positions in human, and vice versa. Note that the region with the strongest orthology to rhesus macaque palindrome arms corresponds to flanking sequence in human.



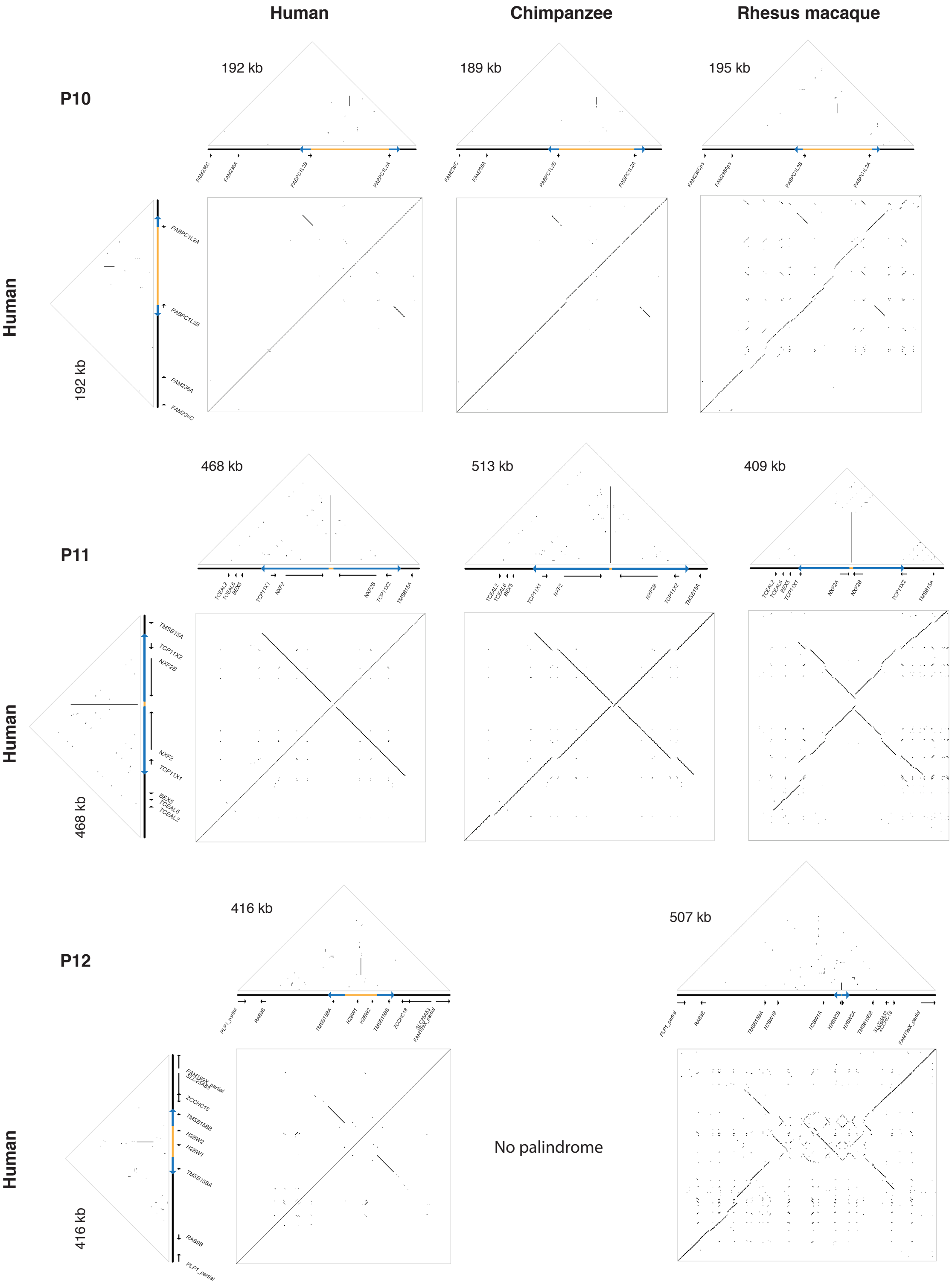
Supplemental Figure S5. Annotated square and triangular dot plots of primate X palindromes. w=100 for all triangle plots. w=100 for human vs. human square plots, human vs. chimpanzee square plots; w=40 for human vs. rhesus macaque square plots.



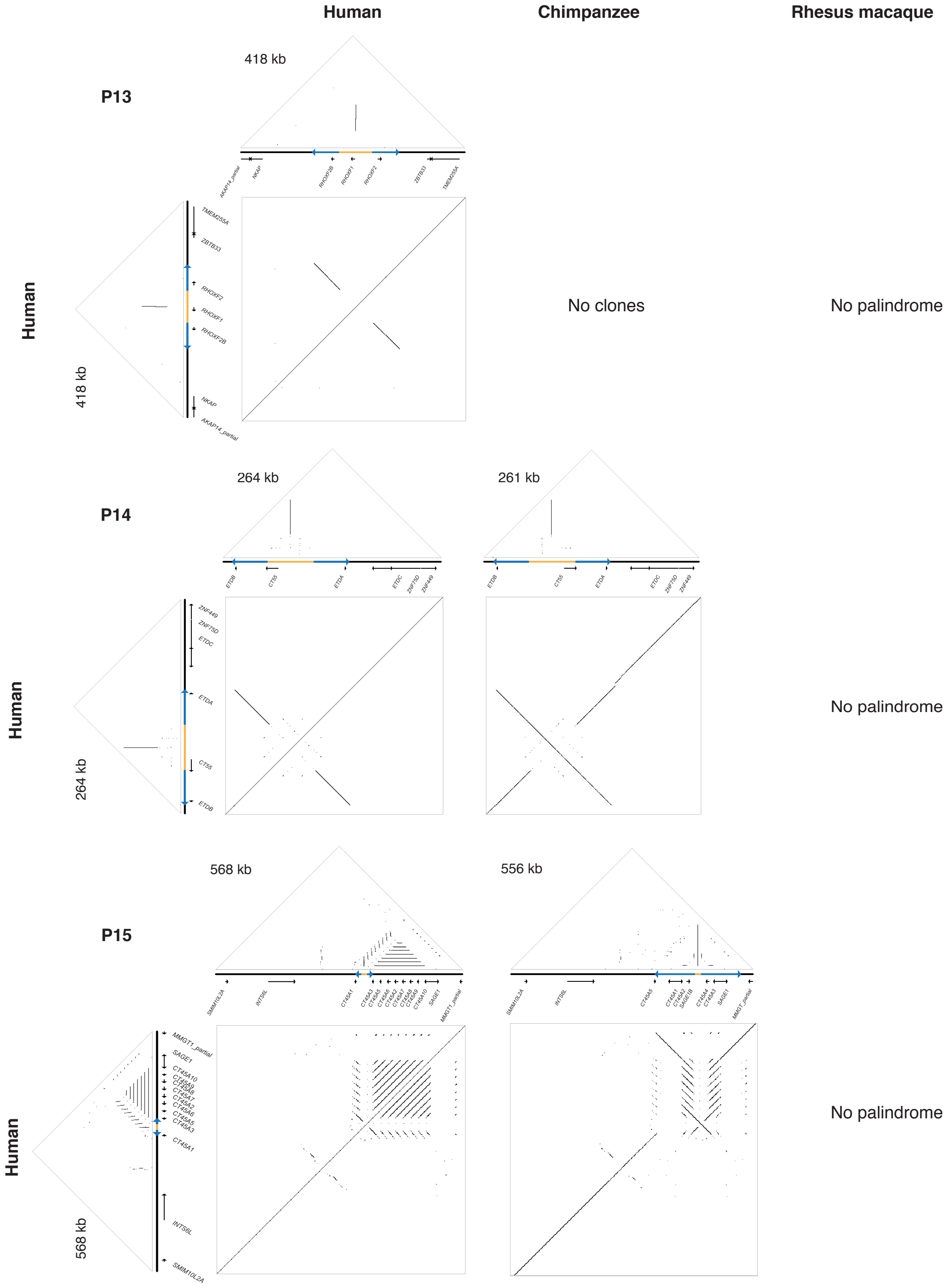
Supplemental Figure S5. Annotated square and triangular dot plots of primate X palindromes. w=100 for all triangle plots. w=100 for human vs. human square plots, human vs. chimpanzee square plots; w=40 for human vs. rhesus macaque square plots.



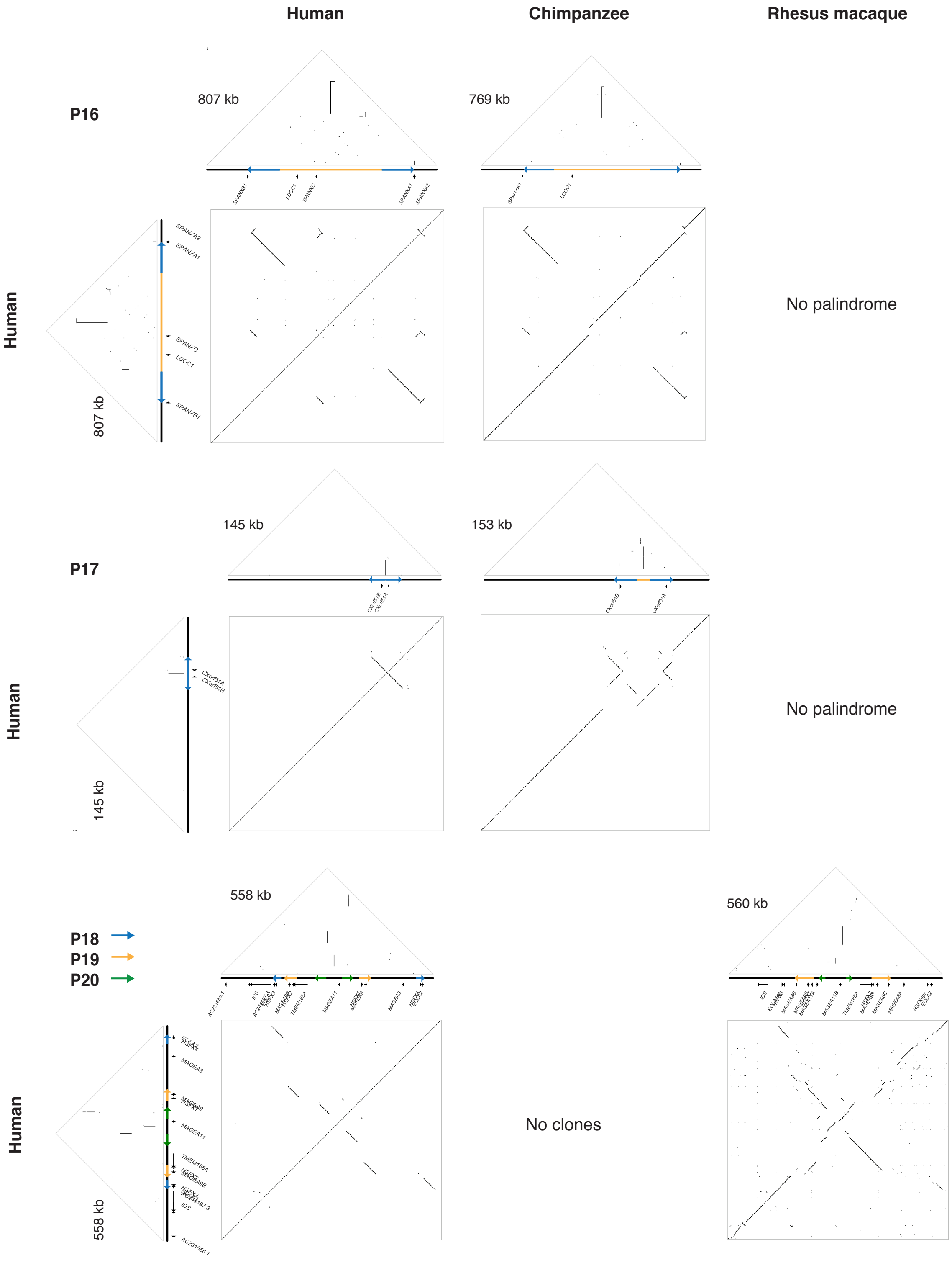
Supplemental Figure S5. Annotated square and triangular dot plots of primate X palindromes. w=100 for all triangle plots. w=100 for human vs. human square plots, human vs. chimpanzee square plots; w=40 for human vs. rhesus macaque square plots.



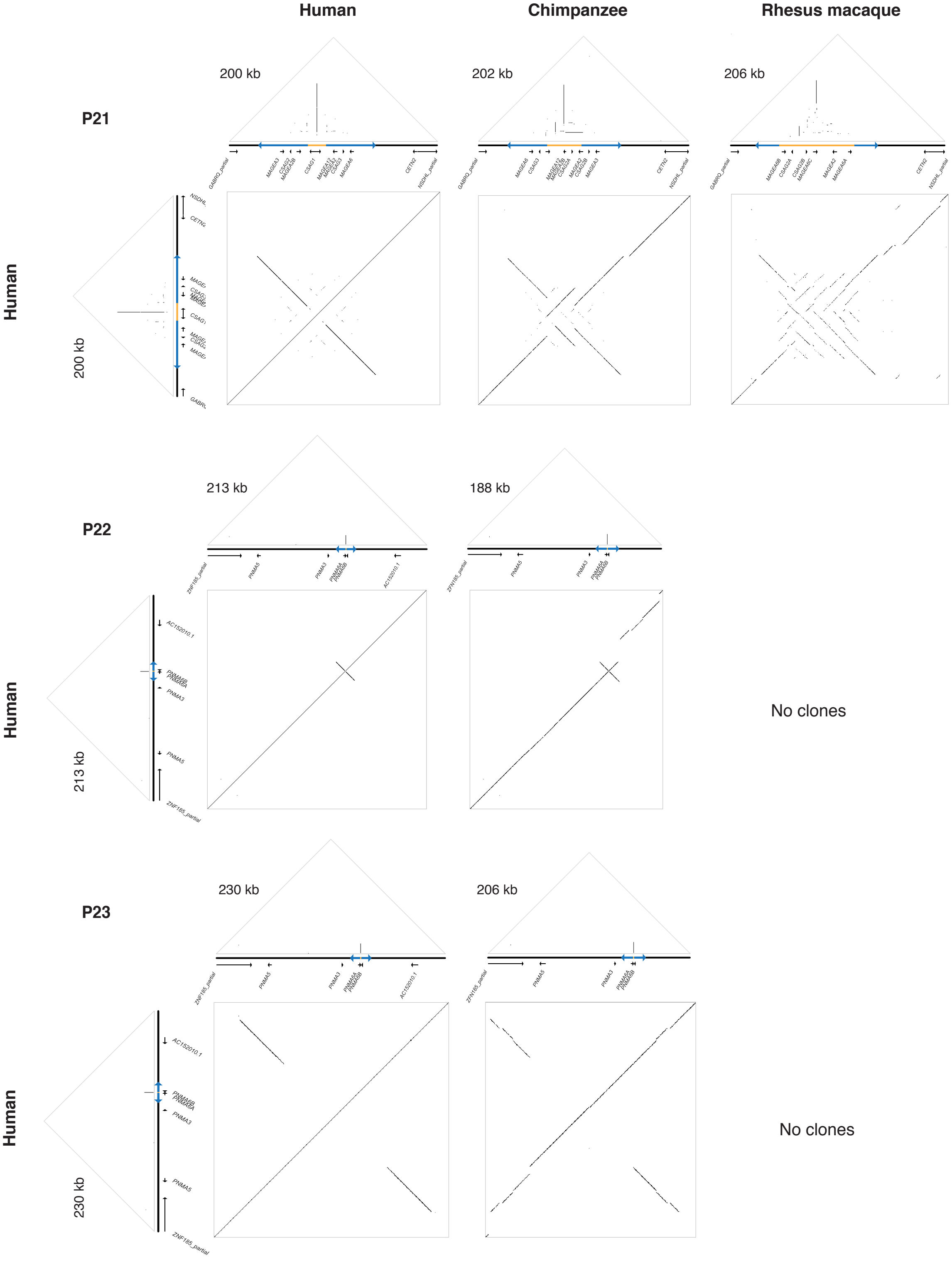
Supplemental Figure S5. Annotated square and triangular dot plots of primate X palindromes. w=100 for all triangle plots. w=100 for human vs. human square plots, human vs. chimpanzee square plots; w=40 for human vs. rhesus macaque square plots.



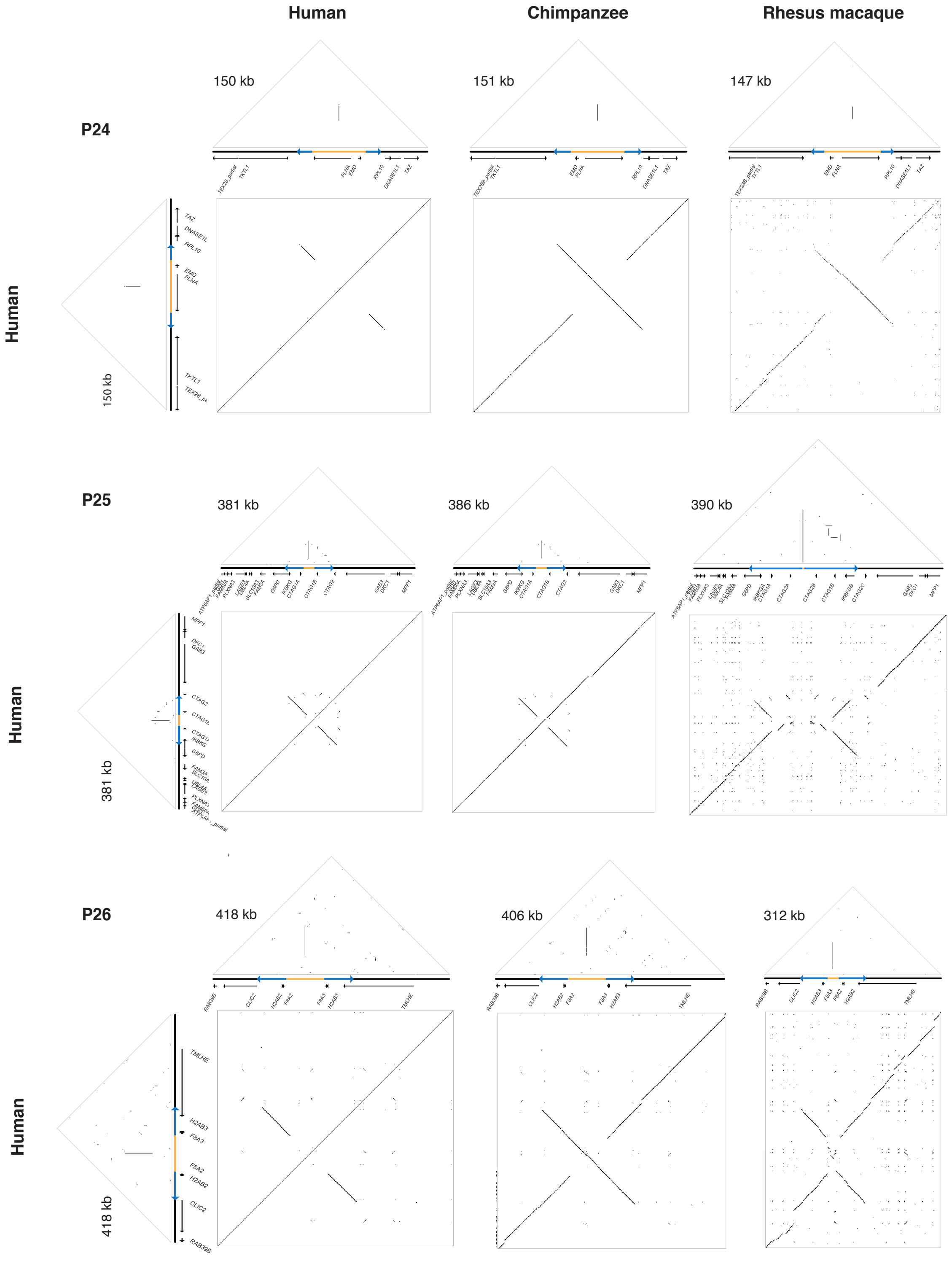
Supplemental Figure S5. Annotated square and triangular dot plots of primate X palindromes. w=100 for all triangle plots. w=100 for human vs. human square plots, human vs. chimpanzee square plots; w=40 for human vs. rhesus macaque square plots.



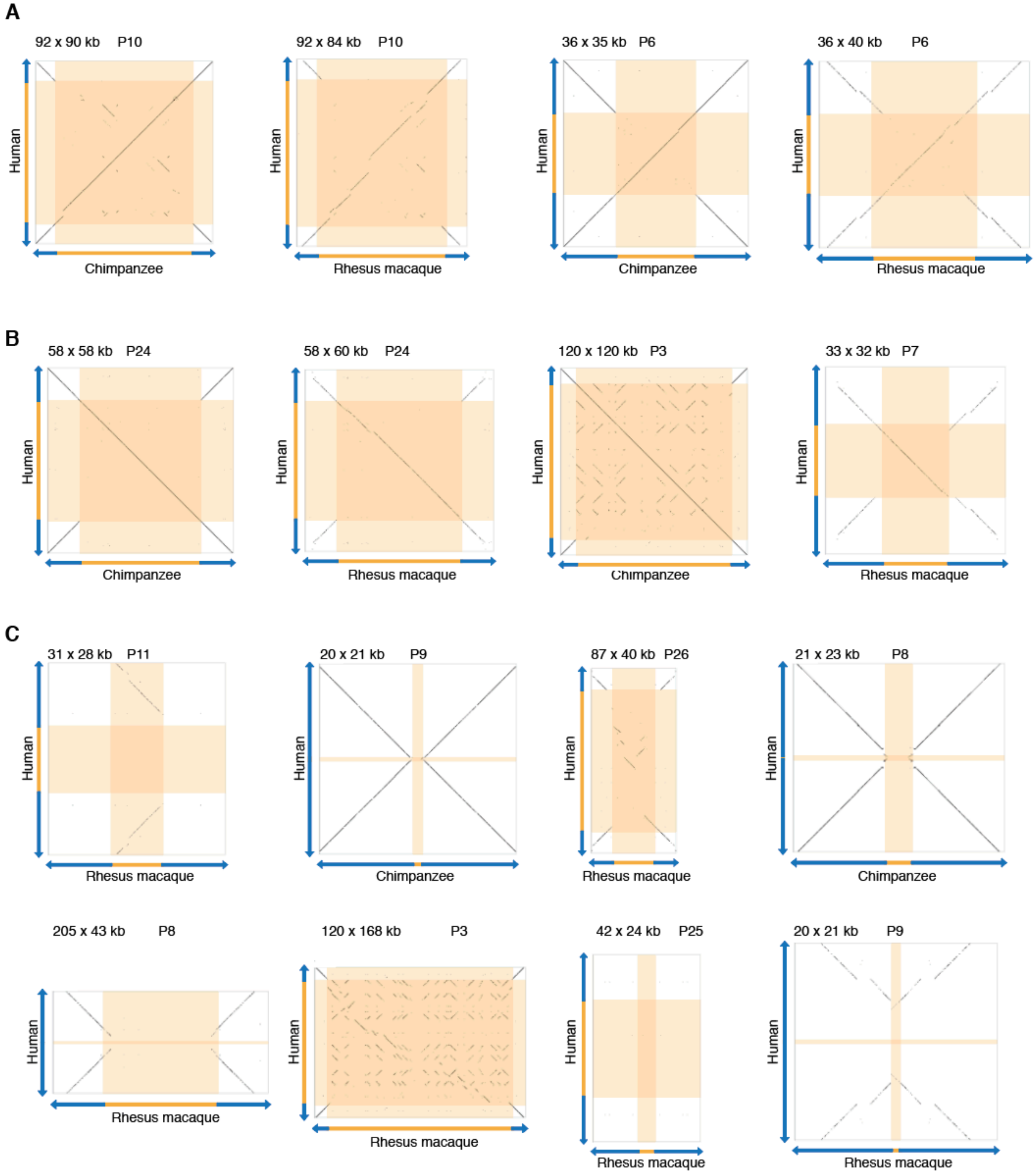
Supplemental Figure S5. Annotated square and triangular dot plots of primate X palindromes. w=100 for all triangle plots. w=100 for human vs. human square plots, human vs. chimpanzee square plots; w=40 for human vs. rhesus macaque square plots.



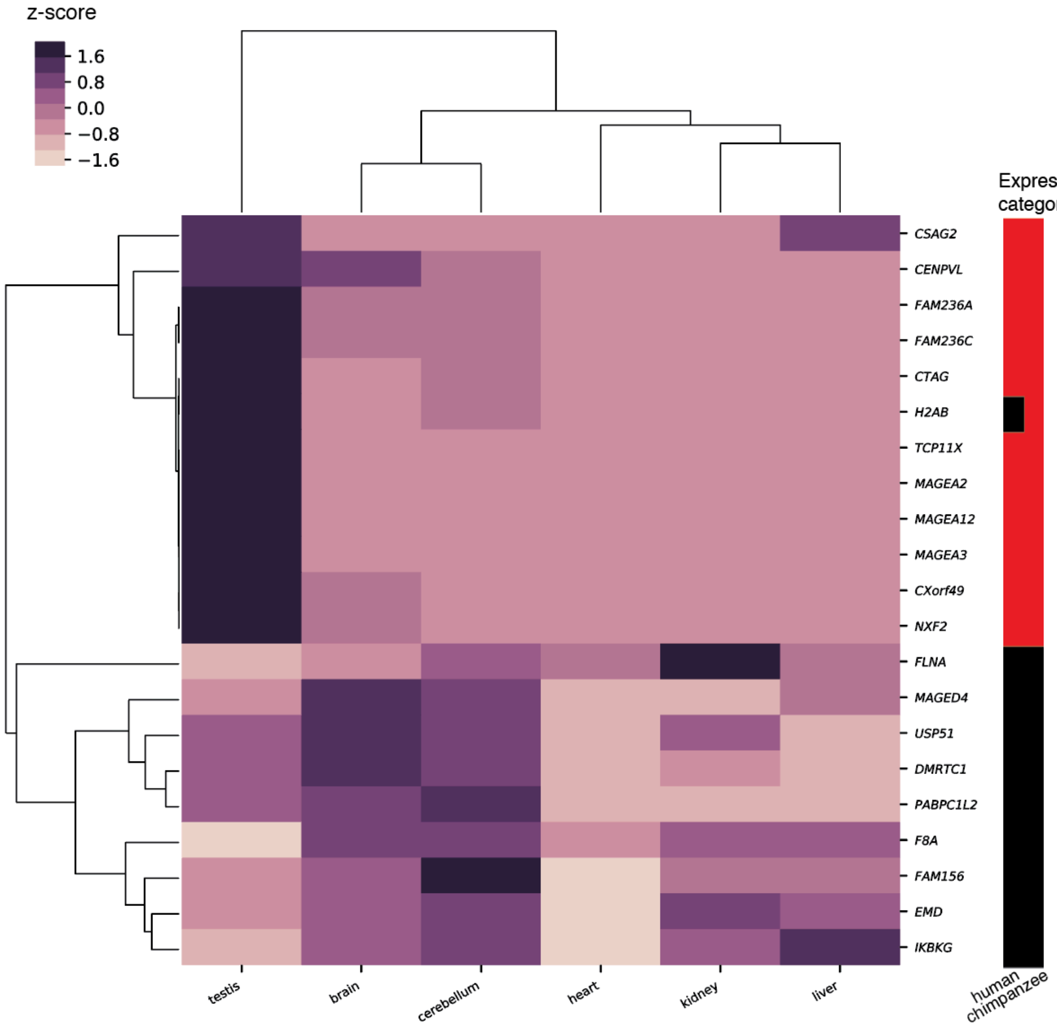
Supplemental Figure S5. Annotated square and triangular dot plots of primate X palindromes. w=100 for all triangle plots. w=100 for human vs. human square plots, human vs. chimpanzee square plots; w=40 for human vs. rhesus macaque square plots.



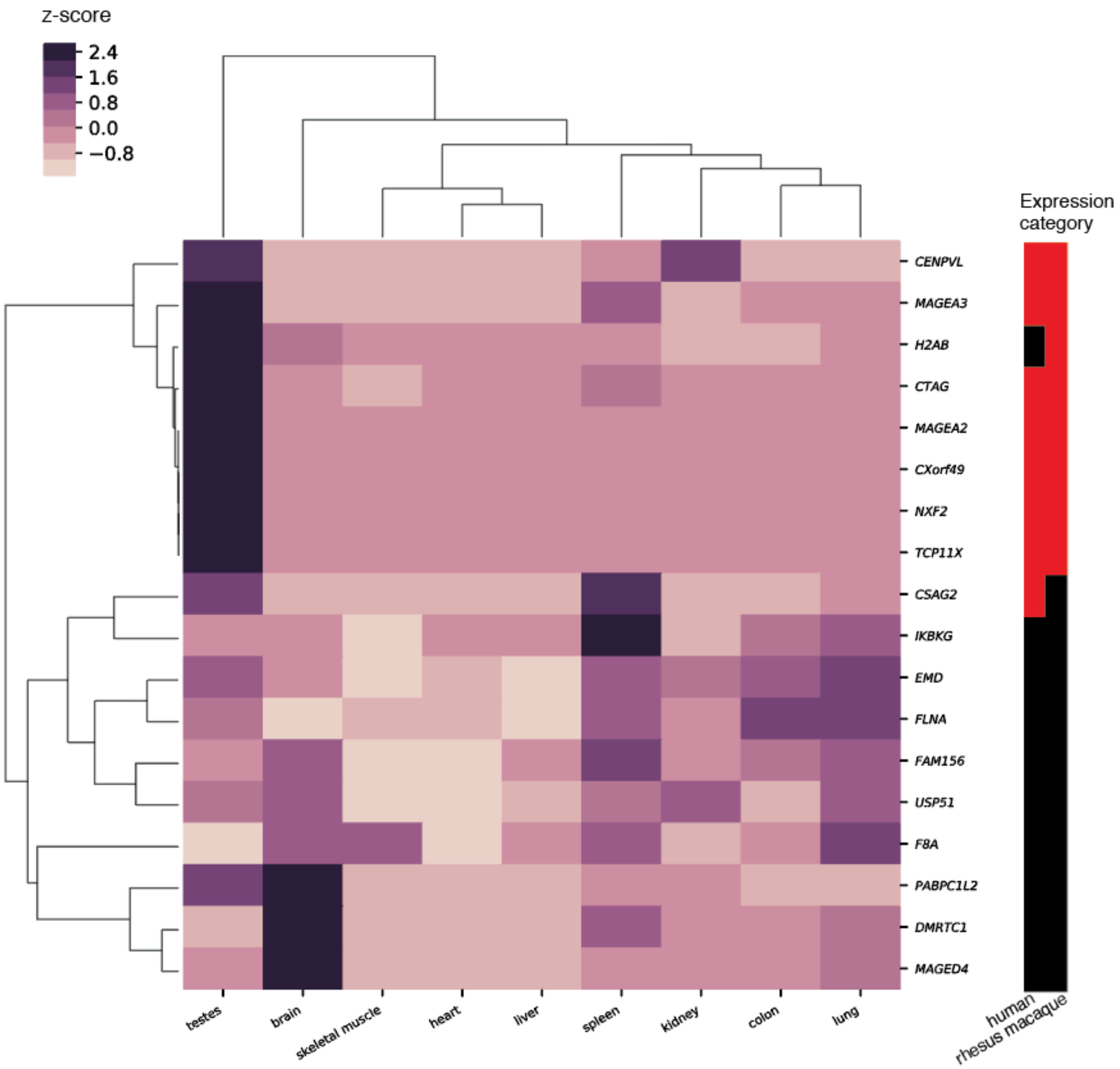
Supplemental Figure S5. Annotated square and triangular dot plots of primate X palindromes. w=100 for all triangle plots. w=100 for human vs. human square plots, human vs. chimpanzee square plots; w=40 for human vs. rhesus macaque square plots.



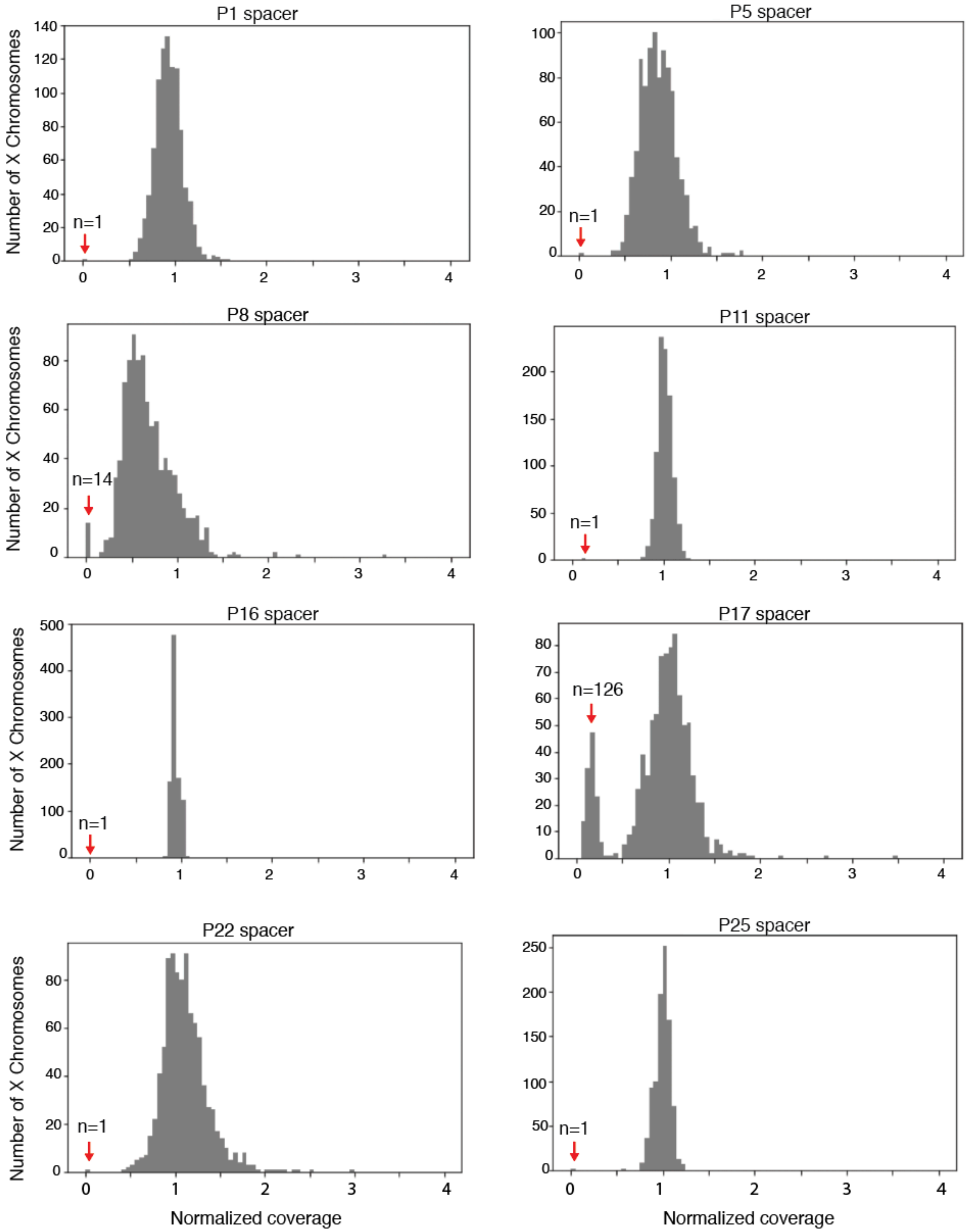
Supplemental Figure S6. Additional examples of spacer configurations in orthologous palindromes. All plots show the inner 10 kb of palindrome arms plus the spacer. w=40 for human vs. rhesus macaque comparisons; w=100 for human vs. chimpanzee comparisons. a) Human configuration, b) Inversions, c) Non-orthologous spacers.



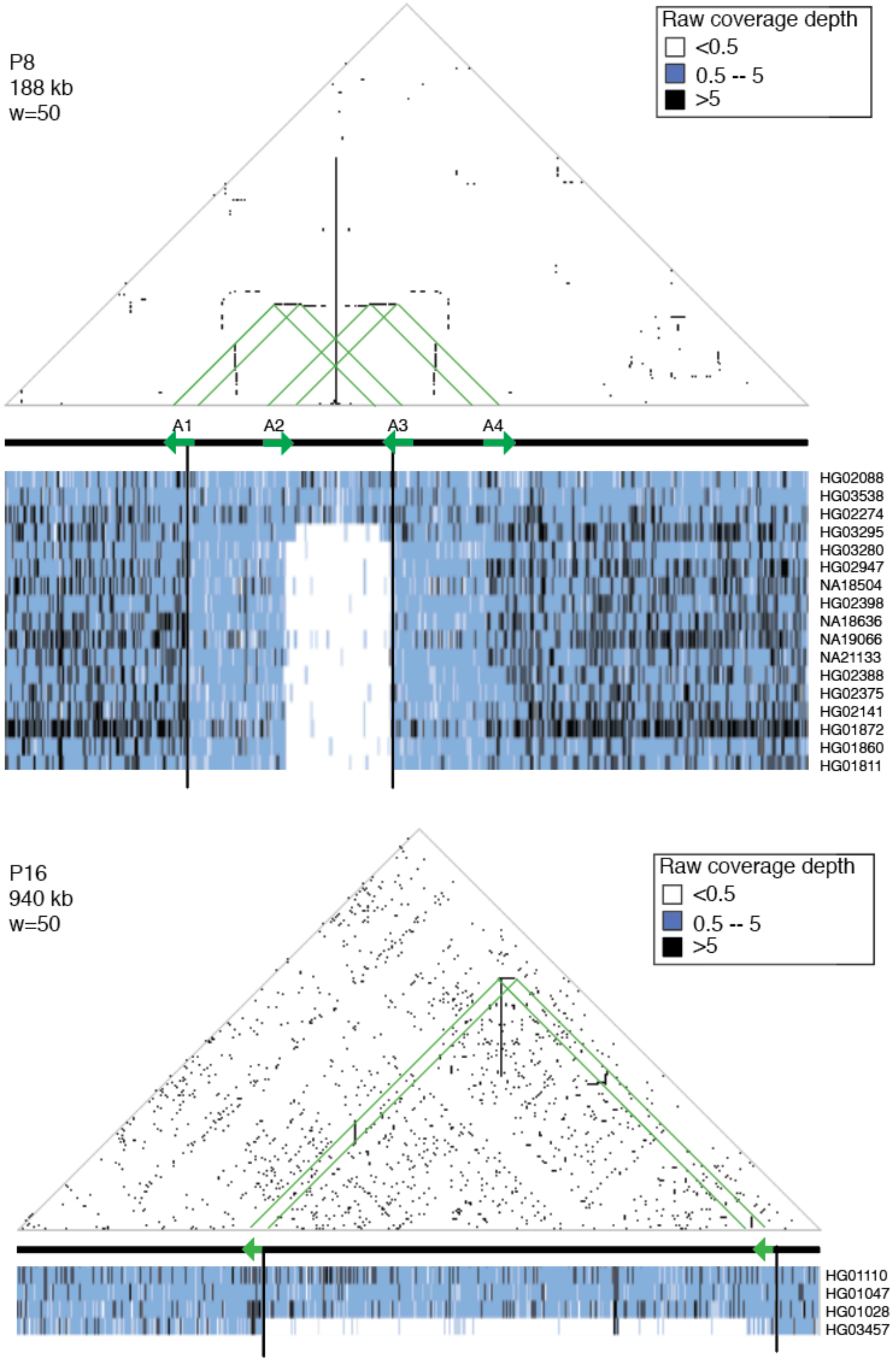
Supplemental Figure S7. Expression of gene families from palindromes shared by human, chimpanzee, and macaque in chimpanzee. Data from Brawand et al. 2011 was re-analyzed with kallisto. Each row shows averaged expression from one gene family. Row and column orders were determined by hierarchical clustering. Expression category: Shows whether expression is testis-biased (red) or broad (black) in the indicated species. Testis-biased: Minimum 2 TPM in testis, and testis accounts for >25% of log2 normalized expression summed across all tissues. Broad: All other expressed genes.



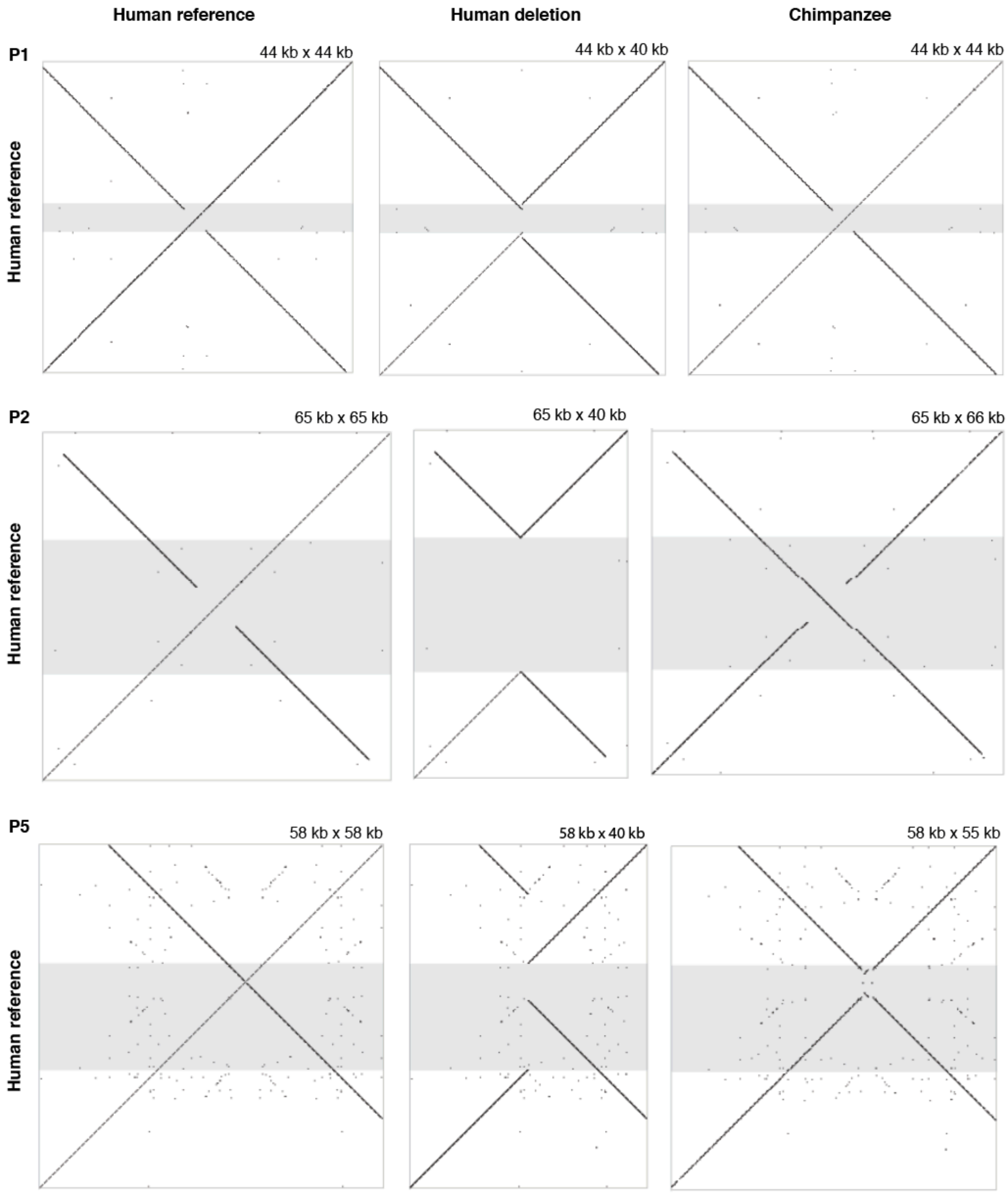
Supplemental Figure S8. Expression of gene families from palindromes shared by human, chimpanzee, and macaque in macaque. Data from Merkin et al. 2012 was re-analyzed with kallisto. Each row shows averaged expression from one gene family. Row and column orders were determined by hierarchical clustering. Expression category: Shows whether expression is testis-biased (red) or broad (black) in the indicated species. Testis-biased: Minimum 2 TPM in testis, and testis accounts for >25% of log2 normalized expression summed across all tissues. Broad: All other expressed genes.



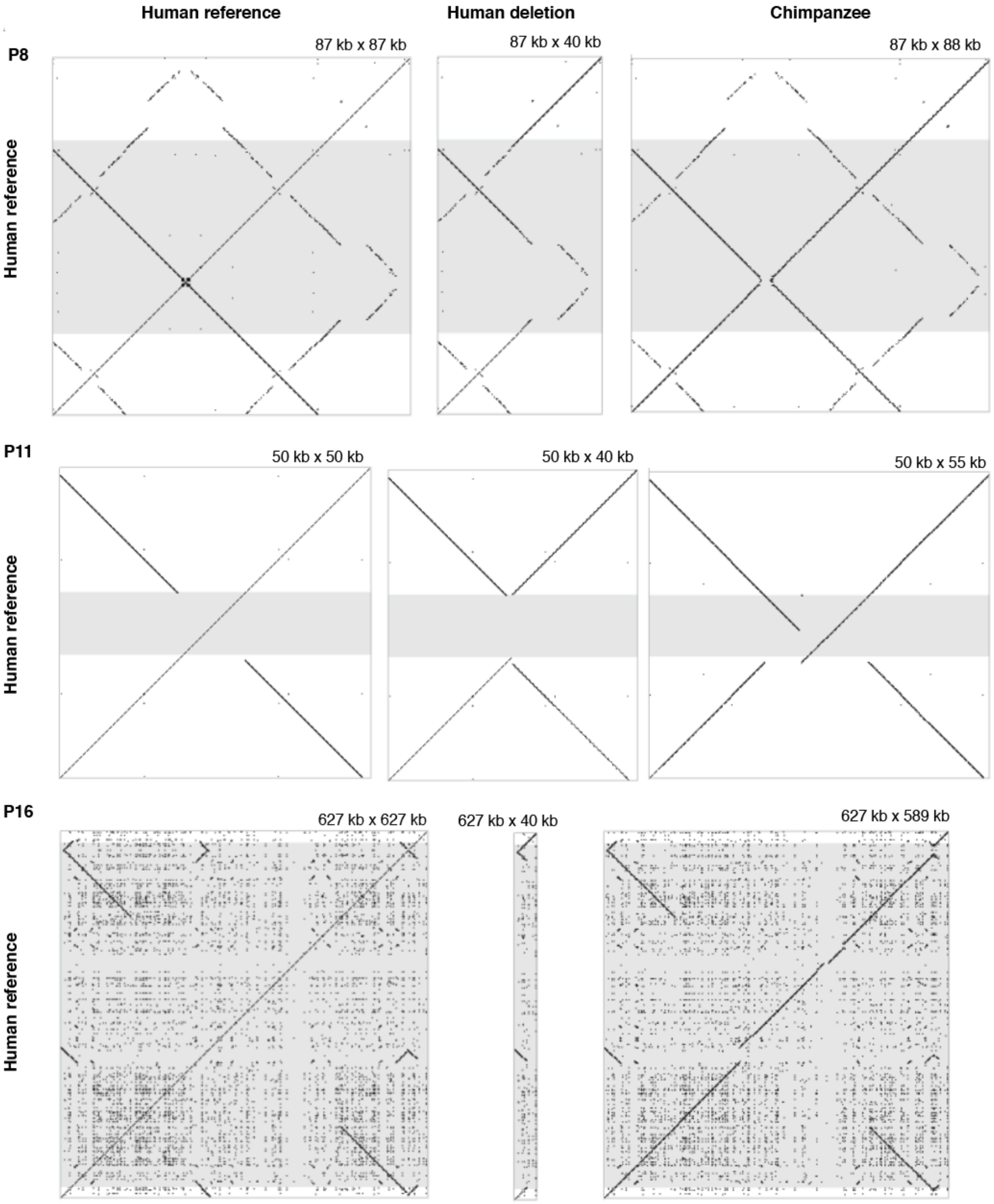
Supplemental Figure S9. Normalized coverage depths for eight palindrome spacers with at least one deletion in the 1000 Genomes dataset. Coverage depth for the ninth palindrome with spacer deletions, P2, is shown in Figure 6A.



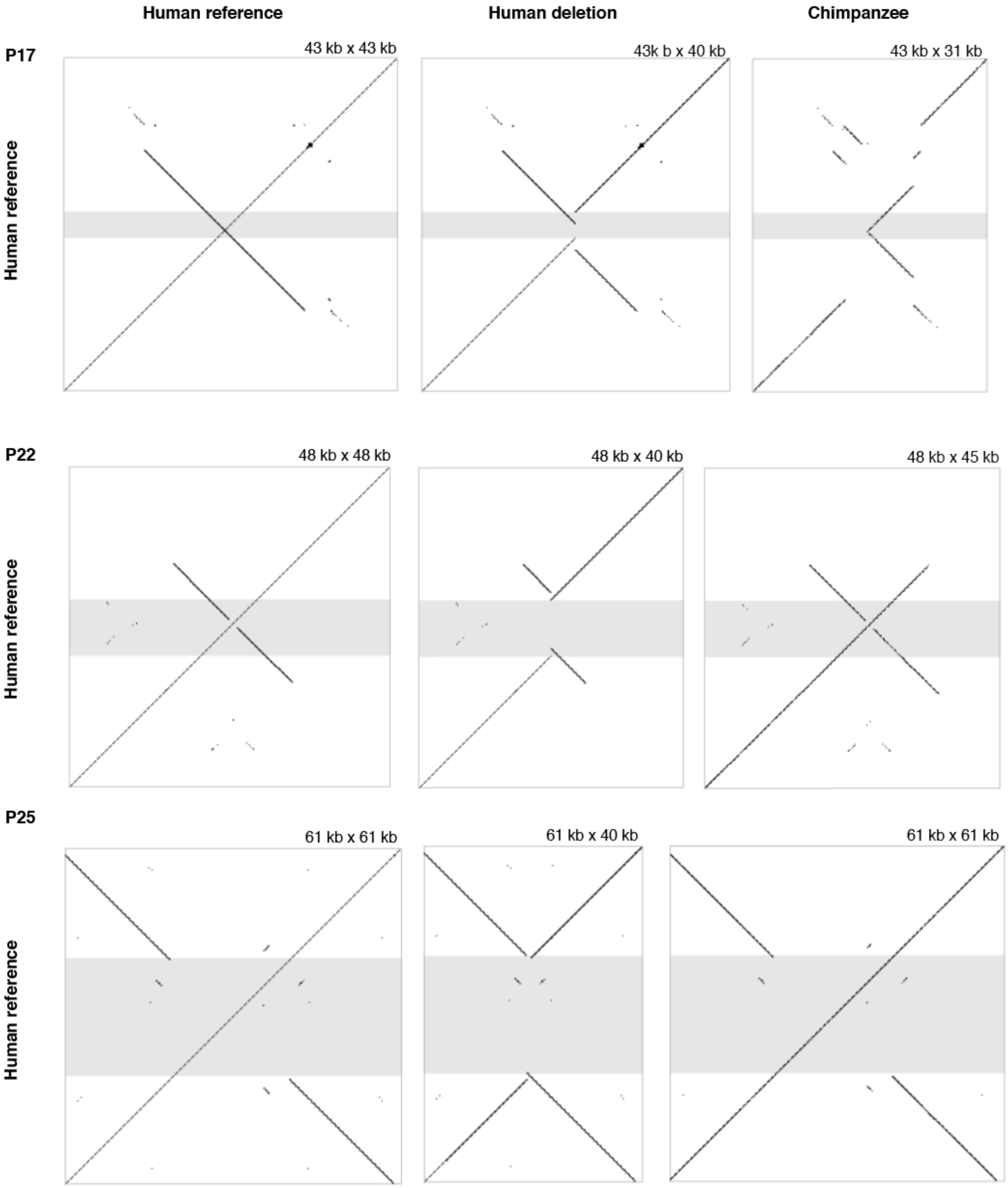
Supplemental Figure S10. Human spacer deletions with breakpoints within tandem repeats. Green arrows: Tandem repeats suspected to cause deletions through NAHR. In the case of P8, the deletion could have occurred with equal probability between arrows A1 & A3, or A2 & A4. Note that the suspected P8 deletion spans areas of no coverage (white) and reduced coverage (lighter blue, from approximately A1 to A2); the copy number of the lighter blue region is reduced from 2 to 1.



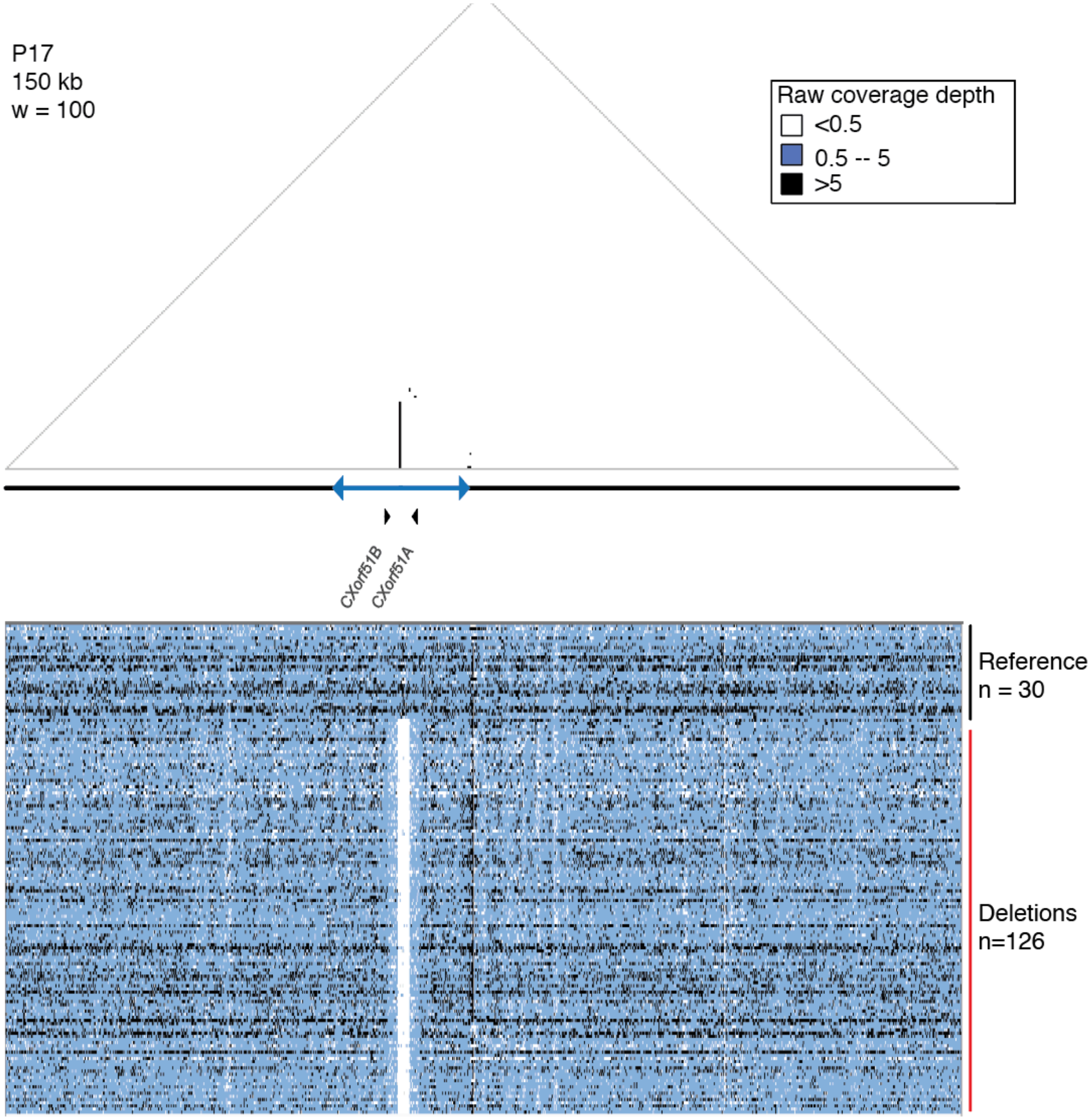
Supplemental Figure S11. Structural comparisons between human reference, human deletion, and chimpanzee for nine X palindromes with at least one spacer deletion. w=30 for all square dot plots. Position of the human X spacer deletion is highlighted in gray. For all nine palindromes, most or all of the of sequence absent in the human deletion is present in chimpanzee, confirming that the human structural polymorphism results from deletion rather than insertion.



Supplemental Figure S11. Structural comparisons between human reference, human deletion, and chimpanzee for nine X palindromes with at least one spacer deletion. w=30 for all square dot plots. Position of the human X spacer deletion is highlighted in gray. For all nine palindromes, most or all of the of sequence absent in the human deletion is present in chimpanzee, confirming that the human structural polymorphism results from deletion rather than insertion.



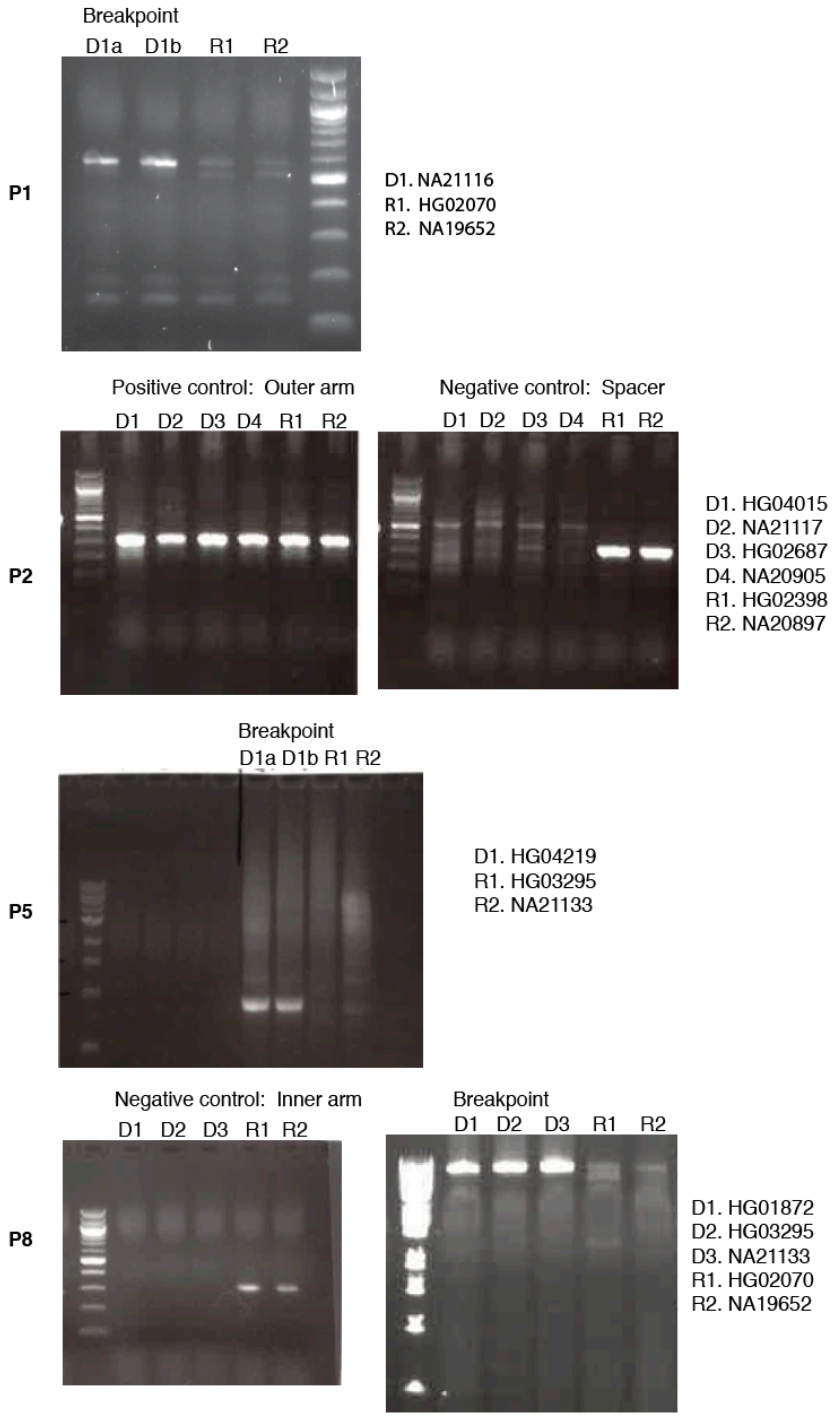
Supplemental Figure S11. Structural comparisons between human reference, human deletion, and chimpanzee for nine X palindromes with at least one spacer deletion. w=30 for all square dot plots. Position of the human X spacer deletion is highlighted in gray. For all nine palindromes, most or all of the of sequence absent in the human deletion is present in chimpanzee, confirming that the human structural polymorphism results from deletion rather than insertion.



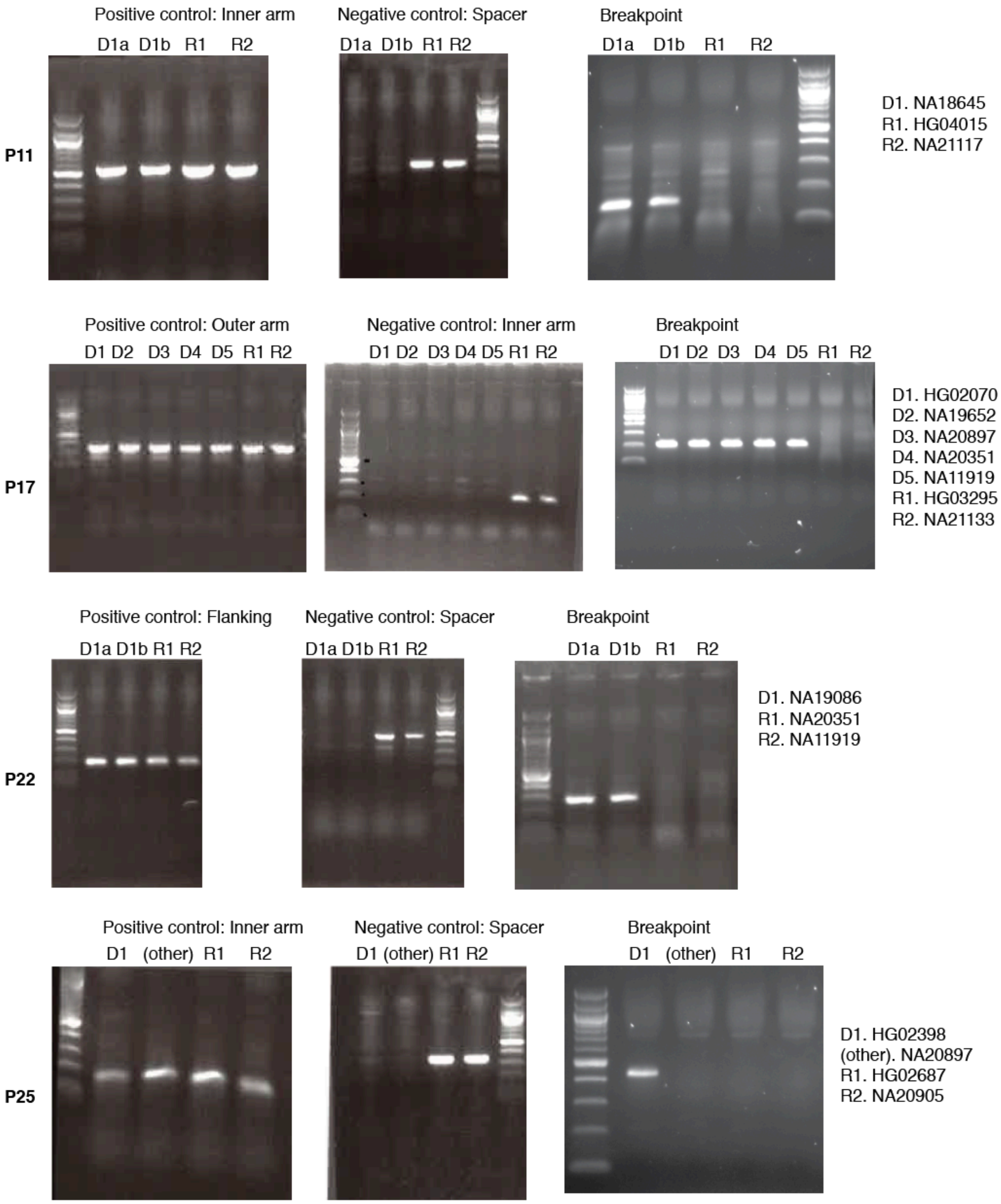
Supplemental Figure S12. Coverage depth for X Chromosomes with P17 spacer deletions. Tracks are shown for all 126 X Chromosomes with P17 spacer deletions, plus 30 randomly chosen X Chromosomes with the reference structure.

```
Proximal reference  GCTTTCTGTCACTATCTGTTTCTATTTTCTGGAGGTTTGTAAGAATGGAATCATAAGATATGTACTCTTT
                    ...................................
P17 deletion        GCTTTCTGTCACTATCTGTTTCTATTTTCTGGAGGCTCACTGGTCACCTTTGCCATACTGAATTGTCCC
                                                     ....................................
Distal reference     TCTGTTCGTCCACATCTTCATTAGGCTTCTGTGGCTCACTGGTCACCTTTGCCATACTGAATTGTCCC
```

Supplemental Figure S13. Junction for P17 spacer deletion. Proximal reference: chrX: 146811296-146811365. Distal reference: chrX: 146814668-146814736. Two base pairs (GG, purple) overlap between the breakpoints.



Supplemental Figure S14. Verification of human X-palindrome spacer deletions. PCR primers were designed based on deletion breakpoints from split reads or, in cases where reads spanning the breakpoint could not be found, based on the estimated deletion breakpoints from visualization of coverage depth. Positive control: Sequence expected to be present in both reference samples and deletion samples. Negative control: Sequence expected to be present in reference samples, and absent in deletion samples. Breakpoint: Sequence expected to be present in deletion samples, and absent in reference samples. D = deletion sample, R = reference sample.



Supplemental Figure S14. Verification of human X-palindrome spacer deletions. PCR primers were designed based on deletion breakpoints from split reads or, in cases where reads spanning the breakpoint could not be found, based on the estimated deletion breakpoints from visualization of coverage depth. Positive control: Sequence expected to be present in both reference samples and deletion samples. Negative control: Sequence expected to be present in reference samples, and absent in deletion samples. Breakpoint: Sequence expected to be present in deletion samples, and absent in reference samples. D = deletion sample, R = reference sample.

Supplemental Table S1: Coordinates of human X palindromes in hg38. Palindromes were identified using a kmer-based method (see Methods).

Palindrome	**Arm 1 coordinates**	**Spacer coordinates**	**Arm 2 coordinates**
P1	chrX: 48367308 - 48396305	chrX: 48396306 - 48399118	chrX: 48399119 - 48428123
P2	chrX: 51668115 - 51692963	chrX: 51692964 - 51700361	chrX: 51700362 - 51725226
P3	chrX: 52040713 - 52077110	chrX: 52077111 - 52176982	chrX: 52176983 - 52213380
P4	chrX: 52473449 - 52502219	chrX: 52502220 - 52510667	chrX: 52510668 - 52539437
P5	chrX: 52670289 - 52728969	chrX: 52728970 - 52729454	chrX: 52729455 - 52788214
P6	chrX: 52882200 - 52920138	chrX: 52920139 - 52935673	chrX: 52935674 - 52973611
P7	chrX: 55453561 - 55480141	chrX: 55480142 - 55493061	chrX: 55493062 - 55519639
P8	chrX: 71683199 - 71740534	chrX: 71740535 - 71741054	chrX: 71741055 - 71798363
P9	chrX: 72741066 - 72860190	chrX: 72860191 - 72860605	chrX: 72860606 - 72979767
P10	chrX: 72996086 - 73005278	chrX: 73005279 - 73077748	chrX: 73077749 - 73086935
P11	chrX: 102197589 - 102338170	chrX: 102338171 - 102348932	chrX: 102348933 - 102489526
P12	chrX: 103955950 - 103987907	chrX: 103987908 - 104050879	chrX: 104050880 - 104082911
P13	chrX: 120038201 - 120086776	chrX: 120086777 - 120149524	chrX: 120149525 - 120198163
P14	chrX: 135116122 - 135158129	chrX: 135158130 - 135214412	chrX: 135214413 - 135256414
P15	chrX: 135723708 - 135734732	chrX: 135734733 - 135748584	chrX: 135748585 - 135759685
P16	chrX: 141005250 - 141114518	chrX: 141114519 - 141472891	chrX: 141472892 - 141582141
P17	chrX: 146801856 - 146812219	chrX: 146812220 - 146812295	chrX: 146812296 - 146822659
P18	chrX: 149542167 - 149562176	chrX: 149562177 - 149917319	chrX: 149917320 - 149937339
P19	chrX: 149573130 - 149602230	chrX: 149602231 - 149767142	chrX: 149767143 - 149796257
P20	chrX: 149654519 - 149681126	chrX: 149681127 - 149722142	chrX: 149722143 - 149748749
P21	chrX: 152678580 - 152725390	chrX: 152725391 - 152743079	chrX: 152743080 - 152789894
P22	chrX: 153066010 - 153074417	chrX: 153074418 - 153075611	chrX: 153075612 - 153084043
P23	chrX: 153106025 - 153149489	chrX: 153149490 - 153250484	chrX: 153250485 - 153293919
P24	chrX: 154337197 - 154347246	chrX: 154347247 - 154384865	chrX: 154384866 - 154394951
P25	chrX: 154555880 - 154591327	chrX: 154591328 - 154613094	chrX: 154613095 - 154648556
P26	chrX: 155336691 - 155386727	chrX: 155386728 - 155453980	chrX: 155453981 - 155504550

Supplemental Table S2: Clones sequenced for this project. Some clones contained portions of multiple palindromes. "Finished": Clone sequence was supported by complete Illumina coverage; "Prefinished": Clone had at least one stretch of sequence supported by one or more nanopore reads, but no Illumina reads.

Clone	**Palindrome**	**Status**	**# full length reads**
CH250-106M20	P7	Prefinished	5
CH250-114J18	P9/P10	Prefinished	0
CH250-119L11	P16	Finished	0
CH250-120L20	P18/P19/P20	Prefinished	29
CH250-136N6	P3	Finished	12
CH250-137I15	P26	Prefinished	8
CH250-138B21	P12	Prefinished	2
CH250-149O24	P1	Prefinished	8
CH250-150I6	P15	Prefinished	0
CH250-163K20	P18/P19/P20	Prefinished	0
CH250-168E3	P21	Prefinished	3
CH250-174F12	P21	Prefinished	19
CH250-184A21	P18/P19/P20	Prefinished	3
CH250-191K20	P2	Finished	13
CH250-197O3	P6	Prefinished	6
CH250-214O8	P17	Finished	79
CH250-228D11	P11	Prefinished	3
CH250-234D7	P16	Finished	0
CH250-236O7	P8	Prefinished	1
CH250-240H14	P14	Finished	6
CH250-257F3	P1	Finished	4
CH250-257M3	P1	Prefinished	1
CH250-25I12	P13	Prefinished	3
CH250-273C12	P15	Prefinished	2
CH250-280C5	P16	Prefinished	6
CH250-300J22	P4/P5	Prefinished	20
CH250-312L23	P16	Finished	0
CH250-313D10	P13	Prefinished	11
CH250-318K15	P16	Finished	0
CH250-371L16	P9	Prefinished	4
CH250-396M7	P1	Prefinished	3
CH250-397P11	P15	Prefinished	23
CH250-398K19	P21	Prefinished	5
CH250-412K19	P5/P6	Prefinished	16

CH250-417G7	P26	Prefinished	2
CH250-420A18	P14	Finished	15
CH250-424H13	P25	Prefinished	23
CH250-436M9	P4/P5	Prefinished	0
CH250-440K2	P13	Prefinished	0
CH250-462M8	P21	Prefinished	4
CH250-486E21	P1	Prefinished	0
CH250-487N16	P26	Finished	2
CH250-491H11	P24	Prefinished	2
CH250-493M11	P25	Prefinished	2
CH250-498I16	P11	Prefinished	8
CH250-499B10	P1	Finished	8
CH250-503C21	P25	Prefinished	10
CH250-503N19	P25	Prefinished	2
CH250-504P11	P18/P19/P20	Prefinished	5
CH250-516N14	P12	Prefinished	0
CH250-530N5	P4/P5	Finished	5
CH250-540J3	P11	Prefinished	10
CH250-541H5	P3	Prefinished	0
CH250-547J16	P18/P19/P20	Prefinished	4
CH250-563M7	P3	Prefinished	22
CH250-57C9	P16	Finished	0
CH250-80G22	P9	Prefinished	6
CH250-87B7	P7	Prefinished	7
CH250-92B13	P12	Prefinished	1
CH250-94G2	P3	Prefinished	4
CH250-95D17	P1	Prefinished	2
CH251-130O9	P4/P5	Finished	0
CH251-160A4	P16	Prefinished	7
CH251-161L14	P7	Prefinished	2
CH251-172F20	P16	Finished	0
CH251-177B21	P2	Prefinished	21
CH251-183G21	P8	Prefinished	0
CH251-189G13	P16	Prefinished	0
CH251-239P10	P26	Prefinished	26
CH251-240O17	P5/P6	Prefinished	61
CH251-261H21	P15	Prefinished	4
CH251-277H18	P7	Prefinished	33
CH251-285D14	P26	Prefinished	1
CH251-292E19	P22/P23	Prefinished	8

CH251-316L7	P17	Prefinished	4
CH251-346A10	P25	Prefinished	12
CH251-34N14	P4/P5	Prefinished	3
CH251-385I8	P1	Prefinished	4
CH251-389B7	P14	Prefinished	20
CH251-397P16	P3	Prefinished	2
CH251-4M24	P11	Finished	21
CH251-504H5	P10	Finished	0
CH251-506D4	P6	Prefinished	0
CH251-50L15	P15	Prefinished	5
CH251-514B7	P8	Prefinished	3
CH251-542A6	P10	Prefinished	12
CH251-542D16	P12	Finished	17
CH251-542E16	P22	Prefinished	4
CH251-550E20	P12	Finished	0
CH251-565G15	P21	Prefinished	3
CH251-571K4	P15	Finished	23
CH251-58J24	P16	Finished	0
CH251-635P13	P11	Prefinished	15
CH251-639F23	P15	Prefinished	3
CH251-64D22	P16	Prefinished	4
CH251-651H9	P11	Prefinished	16
CH251-654E24	P16	Prefinished	25
CH251-657L4	P9/P10	Prefinished	17
CH251-658J15	P8	Prefinished	7
CH251-65E21	P24	Prefinished	1
CH251-671I19	P3	Prefinished	5
CH251-673E12	P16	Prefinished	2
CH251-677L24	P24/P25	Prefinished	8
CH251-702N4	P16	Finished	0
CH251-737G9	P9/P10	Prefinished	1
CH251-73C22	P2	Prefinished	0
CH251-83H5	P4	Finished	0

Supplemental Table S3: Status of palindromes identified in SHIMS 3.0 in other chimpanzee X assemblies. Missing: No palindrome present in non-SHIMS 3.0 assembly; Incomplete: Part of palindrome present in non-SHIMS 3.0 assembly; Accurate: Full palindrome present in non-SHIMS 3.0 assembly.

Palindrome	**Pan_tro_3.0**	**Clint_PTRv2**	**Arm length (SHIMS 3.0)**
P1	Accurate	Incomplete	28844
P2	Accurate	Accurate	25534
P3	Incomplete	Missing	36270
P4	Missing	Missing	29842
P5	Missing	Missing	105400
P6	Missing	Incomplete	34928
P7	Accurate	Missing	28530
P8	Missing	Incomplete	53154
P9	Missing	Incomplete	119578
P10	Missing	Accurate	9184
P11	Incomplete	Incomplete	160682
P14	Incomplete	Incomplete	41640
P15	Incomplete	Missing	90284
P16	Missing	Incomplete	102504
P17	Accurate	Accurate	14737
P21	Incomplete	Incomplete	37360
P22	Incomplete	Accurate	9934
P23	Missing	Incomplete	38530
P24	Missing	Accurate	11228
P25	Accurate	Missing	35368
P26	Accurate	Accurate	49024

Supplemental Table S4: Status of palindromes identified in SHIMS 3.0 in other rhesus macaque X assemblies. Missing: No palindrome present in non-SHIMS 3.0 assembly; Incomplete: Part of palindrome present in non-SHIMS 3.0 assembly; Accurate: Full palindrome present in non-SHIMS 3.0 assembly.

Palindrome	Mmul_8.0.1	Mmul_10	Arm length (SHIMS 3.0)
P1	Missing	Accurate	14749
P2	Incomplete	Accurate	42783
P3	Missing	Accurate	35266
P4	Missing	Accurate	11323
P5	Missing	Accurate	19290
P6	Missing	Accurate	15572
P7	Incomplete	Accurate	24490
P8	Incomplete	Missing	38496
P9	Incomplete	Incomplete	81971
P10	Accurate	Accurate	6574
P11	Incomplete	Missing	106186
P12	Missing	Accurate	12978
P18	Missing	Incomplete	47531
P19	Missing	Incomplete	14175
P21	Accurate	Accurate	21963
P24	Missing	Accurate	8398
P25	Missing	Missing	103488
P26	Missing	Accurate	46855

Supplemental Table S5: Conservation of X-palindrome gene families. Values shown are the number of intact gene copies found in the given species. "Region" refers to the human region in which the gene family is found.

Gene	Palindrome	Region	Human	Chimpanzee	Rhesus macaque
CENPVL	P2	Arm	2	2	2
MAGED4	P3	Arm	2	2	2
FAM156	P6	Arm	2	2	2
USP51	P7	Spacer	1	1	1
CXorf49	P8	Arm	2	2	2
DMRTC1	P9	Arm	2	2	2
FAM236B	P9	Arm	2	2	0
FAM236D	P9	Arm	2	2	0
PABPC1L2	P10	Arm	2	2	2
NXF2	P11	Arm	2	2	2
TCP11X2	P11	Arm	2	2	2
MAGEA12	P21	Spacer	1	1	0
CSAG1	P21	Spacer	1	0	0
MAGEA2	P21	Arm	2	2	1
MAGEA3	P21	Arm	2	2	3
CSAG2	P21	Arm	2	3	2
FLNA	P24	Spacer	1	1	1
EMD	P24	Spacer	1	1	1
CTAG1	P25	Arm	2	2	4
IKBKG	P25	Arm	1	1	2
F8A	P26	Arm	2	2	2
H2AB	P26	Arm	2	2	2

Supplemental Table S6: Purifying selection on X-palindrome genes. dN/dS values were calculated using the basic model in PAML (model=0,NSites=0).

2ΔL (observed vs. 1): Likelihood ratio test for observed dN/dS value versus null hypothesis (dN/dS = 1)

2ΔL (M1a vs. M2a): Likelihood ratio test for neutral evolution versus positive selection

Gene	dN	dS	dN/dS	2ΔL (observed vs. 1)	2ΔL (M1a vs. M2a)
PABPC1L2	0	0.0472	0.0001	16.94***	0
FLNA	0.0037	0.1483	0.02472	584.08***	0
IKBKG	0.0062	0.1403	0.04421	68.52***	0.18
F8A	0.0055	0.0832	0.06665	32.38***	0
MAGED4	0.0099	0.0695	0.14308	46.38***	0
CENPVL	0.0242	0.0998	0.24204	12.5***	0
FAM156	0.0202	0.0825	0.24445	10.56**	0
USP51	0.0125	0.039	0.31986	12.16***	0.64
H2AB	0.066	0.1925	0.34267	6.70**	0
EMD	0.0369	0.096	0.38423	7.68**	4.18*
NXF2	0.0334	0.0649	0.51501	7.28**	0
MAGEA2	0.063	0.1078	0.58423	3.84	0
MAGEA3	0.1432	0.2091	0.68475	4.12*	3.18
CXorf49	0.0725	0.1036	0.69933	3.06	0.86
CTAG1	0.1309	0.1751	0.74787	1.26	2.72
TCP11X	0.0406	0.0525	0.77242	0.6	0
DMRTC1	0.0165	0.0197	0.83446	0.08	0
CSAG2	0.0958	0.0866	1.10635	0.06	8.86**

*p<0.05
**p<0.01
***p<0.001

Supplemental Table S7: Rates of P17 spacer deletions in azoospermic and oligozoospermic men

Dataset	**# men**	**# P17 deletions**	**% P17 deletions**
1000 Genomes	944	126	13.3
dbGaP phs001023 (control)	292	54	18.5
dbGaP phs001023 (azoospermia)	286	47	16.5
Oligozoospermia	562	68	12.1

Supplemental Table S8: Chimpanzee clones used in this project that were previously sequenced and deposited in GenBank.

Clone	**Palindrome**
CH251-26J9	P1
CH251-17J20	P11
CH251-98I1	P12
CH251-55N11	P13
CH251-52P5	P21
CH251-498N14	P25
CH251-25B20	P26

Supplemental Table S9: PCR primers for gels shown in Supplemental Fig. 14.

Feature	Forward primer (5' to 3')	Reverse primer (5' to 3')	Notes
P1 breakpoint	CCTCCTCCGTGTTTTTCTGA	CACAAGACAGGTGCAAGGAA	
P2 outer arm	CCCTCATCAAAAGGTAGGGG	CTGGGTAAGGAGATGGGGAT	
P2 spacer	CGTGCGTGTGTACCATCTTT	AGCTGACTTACATGGAGGGG	
P5 breakpoint	AGAAGGAGTCTCACTTTTGTCGCCCAAG	GCCTCCCAAAGTGTCTTTGTTCAGTTCA	Long range PCR
P8 breakpoint	AGACTGGGTGTTGCGAACAGACAAAAAC	GGATTTGTCTGAGAACTCATTCTTGGCG	Long-range PCR
P8 inner arm	TCCCACTGCTCTGCATCC	CTGGAAGAAGATCTTTATCCTGC	
P11 breakpoint	AATCCACAGGGGACAGCTC	TGTGGGGATAGGAAGTGACA	
P11 inner arm	GCAGGAGTTGCTTCTGTTACTG	TTTGAGTTTGGCTTTCCTGG	
P11 spacer	TCTGTTGAATATGCTCCACACC	TAGTGCAAATTGCTTTCCAGTC	
P17 breakpoint	TCAAAGTTGAAGGGTGTGGC	TTTGGCAATTCTTCCCTGTC	
P17 inner arm	AAAGCAAGCTCCTAAGGATGTG	GGCATCATCCAAACAAGTGG	
P17 outer arm	ATTCGAATGCTGACTCCCAC	GGGAGCTGAACTGCTGTACC	
P22 breakpoint	AGTACCACACAGAGAGGGAGC	GAGGTCAGGCAAGGAAAGAG	
P22 flanking	AACCATGGTCCCAAAATTCA	TCAGCAGTCAACCAGCATTC	
P22 spacer	TGACCATGACTGTGGGAGAA	CAGCCCCTGCTCAAGACTAC	
P25 breakpoint	TCATAGGCTGTTGATGACGG	CGTGATCCCCAAAGGTTG	
P25 inner arm	CACTGTGTCCGGCAACATAC	TCTGTTCTGAGACCCTGTGC	
P25 spacer	TCACACGCTGGTAATTGCAT	CAGCCCTCAGAAGAATTTGC	

Supplemental Table S10: Chimpanzee and rhesus macaque clones sequenced for this project and deposited in GenBank

Clone	**Accession**
CH250-106M20	AC280444
CH250-114J18	AC280531
CH250-119L11	AC280481
CH250-120L20	AC280562
CH250-136N6	AC280430
CH250-137I15	AC280580
CH250-138B21	AC280536
CH250-149O24	AC280566
CH250-150I6	AC280452
CH250-163K20	AC280436
CH250-168E3	AC280520
CH250-174F12	AC280455
CH250-184A21	AC280508
CH250-191K20	AC280440
CH250-197O3	AC280457
CH250-214O8	AC280424
CH250-228D11	AC280538
CH250-234D7	AC280571
CH250-236O7	AC280541
CH250-240H14	AC280414
CH250-257F3	AC280575
CH250-257M3	AC280477
CH250-25I12	AC280437
CH250-273C12	AC280539
CH250-280C5	AC280451
CH250-300J22	AC280417
CH250-312L23	AC280517
CH250-313D10	AC280454
CH250-318K15	AC280486
CH250-371L16	AC280564
CH250-396M7	AC280543
CH250-397P11	AC280441
CH250-398K19	AC280504
CH250-412K19	AC280483
CH250-417G7	AC280442
CH250-420A18	AC280569
CH250-424H13	AC280467

CH250-436M9	AC280526
CH250-440K2	AC280563
CH250-462M8	AC280473
CH250-486E21	AC280527
CH250-487N16	AC280453
CH250-491H11	AC280464
CH250-493M11	AC280503
CH250-498I16	AC280468
CH250-499B10	AC280432
CH250-503C21	AC280489
CH250-503N19	AC280524
CH250-504P11	AC280429
CH250-516N14	AC280555
CH250-530N5	AC280476
CH250-540J3	AC280456
CH250-541H5	AC280425
CH250-547J16	AC280475
CH250-563M7	AC280492
CH250-57C9	AC280498
CH250-80G22	AC280568
CH250-87B7	AC280549
CH250-92B13	AC280534
CH250-94G2	AC280518
CH250-95D17	AC280449
CH251-130O9	AC280561
CH251-160A4	AC280458
CH251-161L14	AC280465
CH251-172F20	AC280557
CH251-177B21	AC280525
CH251-183G21	AC280578
CH251-189G13	AC280560
CH251-239P10	AC280533
CH251-240O17	AC280544
CH251-261H21	AC280545
CH251-277H18	AC280556
CH251-285D14	AC280499
CH251-292E19	AC280434
CH251-316L7	AC280500
CH251-346A10	AC280446
CH251-34N14	AC280480

CH251-385I8	AC280416
CH251-389B7	AC280426
CH251-397P16	AC280507
CH251-4M24	AC280567
CH251-504H5	AC280540
CH251-506D4	AC280415
CH251-50L15	AC280523
CH251-514B7	AC280488
CH251-542A6	AC280462
CH251-542D16	AC280579
CH251-542E16	AC280512
CH251-550E20	AC280495
CH251-565G15	AC280459
CH251-571K4	AC280521
CH251-58J24	AC280448
CH251-635P13	AC280558
CH251-639F23	AC280574
CH251-64D22	AC280469
CH251-651H9	AC280522
CH251-654E24	AC280553
CH251-657L4	AC280546
CH251-658J15	AC280445
CH251-65E21	AC280530
CH251-671I19	AC280576
CH251-673E12	AC280463
CH251-677L24	AC280565
CH251-702N4	AC280482
CH251-737G9	AC280491
CH251-73C22	AC280423
CH251-83H5	AC280548

CHAPTER 3:

GC-biased gene conversion in X-chromosome palindromes conserved in human, chimpanzee, and rhesus macaque

Emily K. Jackson, Daniel W. Bellott, Helen Skaletsky, and David C. Page

Author contributions
E.K.J., D.W.B, H.S., and D.C.P designed the study. 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.

Acknowledgements
This work was supported by the Howard Hughes Medical Institute, the Whitehead Institute, and generous gifts from Brit and Alexander d'Arbeloff and Arthur W. and Carol Tobin Brill.

Adapted from: Jackson E, Bellott DW, Skaletsky H, and Page DC (submitted to *G3 (Bethesda)*). 2021. GC-biased gene conversion in X-Chromosome palindromes conserved in human, chimpanzee, and rhesus macaque. *bioRxiv* doi: https://doi.org/10.1101/2021.04.21.440828.

ABSTRACT

Gene conversion is GC-biased across a wide range of taxa. Large palindromes on mammalian sex chromosomes undergo frequent gene conversion that maintains arm-to-arm sequence identity greater than 99%, which may increase their susceptibility to the effects of GC-biased gene conversion. Here, we demonstrate a striking history of GC-biased gene conversion in 12 palindromes conserved on the X chromosomes of human, chimpanzee, and rhesus macaque. Primate X-chromosome palindrome arms have significantly higher GC content than flanking single-copy sequences. Nucleotide replacements that occurred in human and chimpanzee palindrome arms over the past 7 million years are one-and-a-half times as GC-rich as the ancestral bases they replaced. Using simulations, we show that our observed pattern of nucleotide replacements is consistent with GC-biased gene conversion with a magnitude of 70%, similar to previously reported values based on analyses of human meioses. However, GC-biased gene conversion explains only a fraction of the observed difference in GC content between palindrome arms and flanking sequence, suggesting that additional factors are required to explain elevated GC content in palindrome arms. This work supports a greater than 2:1 preference for GC bases over AT bases during gene conversion, and demonstrates that the evolution and composition of mammalian sex-chromosome palindromes is strongly influenced by GC-biased gene conversion.

INTRODUCTION

Homologous recombination maintains genome integrity through the repair of double-stranded DNA breaks, while also promoting genetic innovation through programmed reshuffling during meiosis. Homologous recombination can produce crossover events, in which genetic material is exchanged between two DNA molecules, or non-crossover events. Crossover events and non-crossover events both result in gene conversion, the non-reciprocal transfer of DNA sequence from one homologous template to another. When the templates involved in gene conversion are not identical, gene conversion can be biased, resulting in the preferential transmission of one allele over another (reviewed in Galtier et al. 2001, Marais 2003, Duret and Galtier 2009). In particular, GC alleles are generally favored over AT alleles, leading to a strong correlation between GC content and recombination rates across the genome. GC-biased gene conversion is widespread across taxa, including plants (Muyle et al. 2011), yeast (Mancera et al. 2008), birds (Smeds et al. 2016), rodents (Montoya-Burgos et al. 2003, Clément and Arndt 2011), humans (Odenthal-Hesse et al. 2014, Williams et al. 2015, Halldorsson et al. 2016), and other primates (Galtier et al. 2009, Borges et al. 2019).

While early evidence for GC-biased gene conversion was indirect (Galtier et al. 2001, Marais 2003), two recent studies identified gene conversion events in humans directly using three-generation pedigrees (Williams et al. 2015, Halldorsson et al. 2016). This approach enabled calculation of the magnitude of GC bias, defined as the frequency at which gene conversion at a locus containing one GC allele and one AT allele results in transmission of the GC allele. Williams et al. identified 98 autosomal non-crossover gene conversion events at loci with one GC allele and one AT allele, and found that 63 (68%) transmitted the GC allele (Williams et al. 2015). Halldorsson et al. analyzed autosomal crossover and non-crossover gene conversion events separately, and found GC biases of 70.1% and 67.6%, respectively (Halldorsson et al. 2016). The magnitude of GC bias may vary across different genomic positions: Another study used sperm typing to examine allele transmission at six autosomal recombination hotspots, and found evidence for GC-biased transmission at two hotspots, but unbiased transmission at the other four hotspots (Odenthal-Hesse et al. 2014).

Mammalian sex chromosomes contain large, highly identical palindromes, with arms that can exceed 1 Mb in length and arm-to-arm identities greater than 99% (Skaletsky et al. 2003, Warburton et al. 2004, Hughes et al. 2010, Hughes et al. 2012, Mueller et al. 2013, Soh et al. 2014, Hughes et al. 2020, Jackson et al. 2020). Near-perfect identity between palindrome arms is maintained by high rates of ongoing gene conversion (Rozen et al. 2003), which may make palindromes uniquely susceptible to the effects of GC-biased gene conversion (Hallast et al. 2013, Skov et al. 2017). Recently, we generated high-quality reference sequence for twelve large palindromes that are conserved on the X chromosomes of human, chimpanzee, and rhesus macaque, demonstrating a common origin at least 25 million years ago (Jackson et al. 2020). Here, we use a comparative genomic approach combined with evolutionary simulations to analyze the impact and magnitude of GC-biased gene conversion in primate X-chromosome palindromes. We find that GC content is elevated in palindrome arms relative to flanking sequence, and that recent nucleotide replacements in human and chimpanzee palindrome arms are approximately one-and-a-half times as GC-rich as the ancestral bases that they replace. Using simulations of palindrome evolution, we show that our observed pattern of nucleotide replacements is consistent with a magnitude of GC bias of about 70%, which supports recent estimates derived from analyses of human meioses using an orthogonal approach.

RESULTS

High rates of intrachromosomal gene conversion in arms of primate X-chromosome palindromes

To understand the role of GC-biased gene conversion in the evolution of primate X-chromosome palindromes, we first calculated the rate of intrachromosomal gene conversion between palindrome arms. Sequence identity between palindrome arms depends on the balance between two evolutionary forces: The rate at which new mutations arise in each arm, and the rate at which gene conversion between arms homogenizes the resulting sequence differences. The rate of intrachromosomal gene conversion can therefore be calculated using the formula $c = 2\mu /d$, where μ represents the mutation rate, and d represents the fraction divergence between arms (Rozen et al. 2003). Among twelve X-chromosome palindromes conserved between human, chimpanzee, and rhesus macaque, we found a median divergence between arms of 4.7 x 10^{-4} differences per nucleotide, or around one difference per 2200 nucleotides. Assuming a mutation rate of 1.06 x 10^{-8} mutations per nucleotide per generation (Roach et al. 2010, Kong et al. 2012, Jónsson et al. 2017, see Methods), we calculated a gene conversion rate of 4.5 x 10^{-5} events per nucleotide per generation for primate X-chromosome palindromes. This value is nearly eight times the recent estimate of 5.9 x 10^{-6} gene conversion events per nucleotide per generation across human autosomes (Williams et al. 2015, Halldorsson et al. 2016), highlighting the rapid pace of genetic exchange between sex-chromosome palindrome arms.

GC content is elevated in primate X-chromosome palindrome arms compared to flanking sequence

Previous studies have proposed that high rates of gene conversion in sex-chromosome palindromes could lead to elevated GC content in palindrome arms (Caceres et al. 2007, Hallast et al. 2013). We calculated GC content for primate X-chromosome palindrome arms relative to flanking sequence, and found significantly higher median GC content in palindrome arms than in flanking sequence across all three species: 46.3% versus 41.2% (human), 46.3% versus 40.9% (chimpanzee), and 45.2% versus 41.0% (rhesus macaque) ($p<0.05$ for all three species, Mann-Whitney U) (Figure 1A). The GC content of flanking sequences is slightly elevated compared to the overall GC content of the human X chromosome (39.5%), while the GC content of palindrome arms is markedly higher. The trend of elevated GC content in palindrome arms was highly consistent across different palindromes, with at least eleven out of twelve palindromes having significantly higher GC content in palindrome arms than flanking sequence within each species ($p<1 \times 10^{-6}$ for each significant palindrome, chi-square test with Yates correction, Supplemental Table 1). Given that ten out of twelve conserved primate X-chromosome palindrome arms contain one or more protein-coding genes (Jackson et al. 2020), which tend to be GC-rich, we considered the possibility that elevated GC content in primate X-chromosome palindrome arms results from an enrichment of protein-coding genes. However, the difference between GC content in palindrome arms

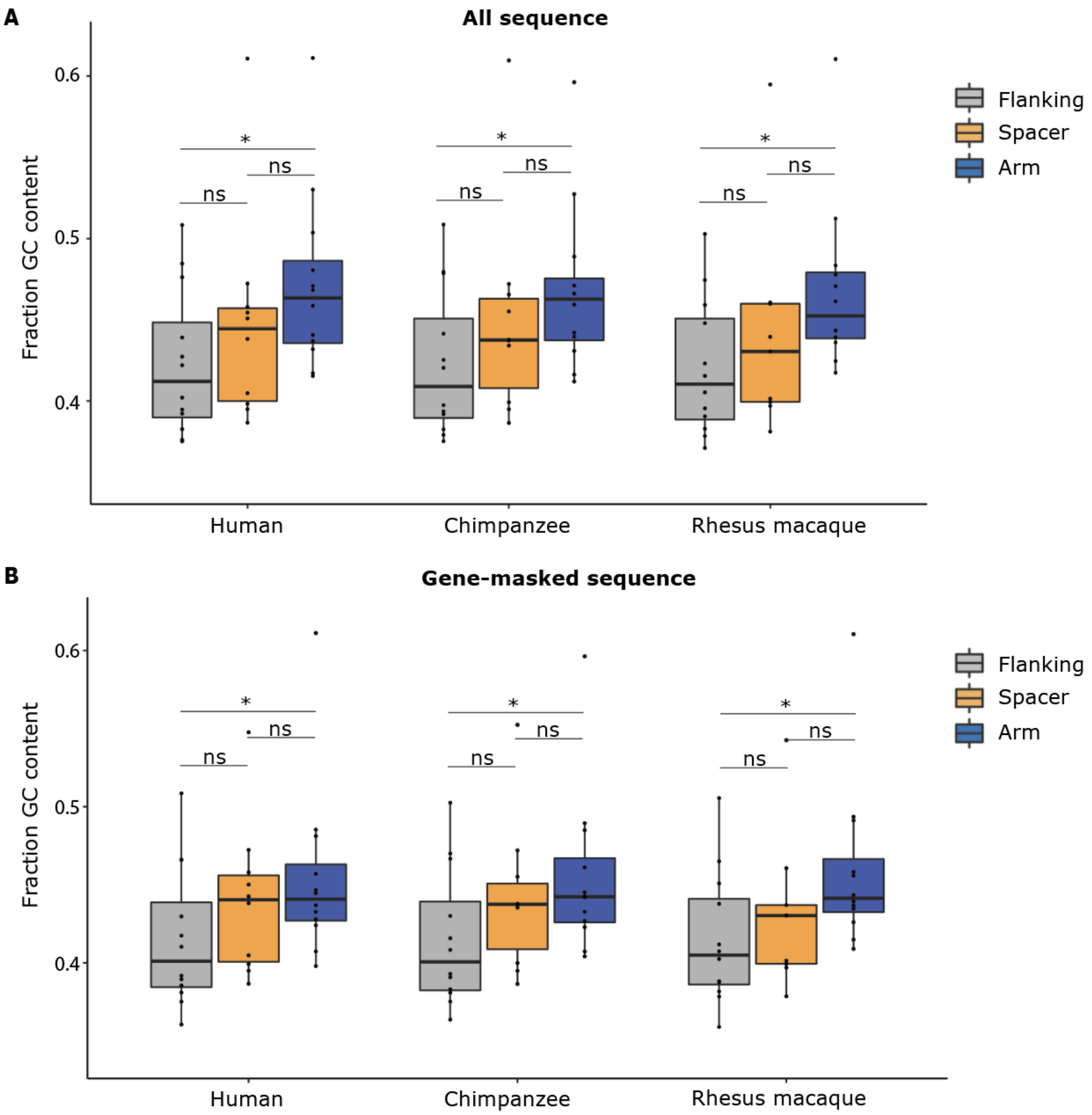


Figure 1. GC content is elevated in primate X-chromosome palindrome arms compared to flanking sequence. GC content measured in 12 palindromes conserved between human, chimpanzee, and rhesus macaque. Small spacers (<5 kb) excluded from analysis. Results A) for all sequence and B) after masking protein-coding genes (gene body plus 1 kb upstream). *p<0.05, ns = not significant, Mann-Whitney U.

and flanking sequence remained significant after masking protein-coding genes plus their promoters (defined as 1 kb upstream): 44.1% versus 40.1% (human), 44.2% versus 40.1% (chimpanzee), and 44.1%

versus 40.5% (rhesus macaque) (p<0.05 for all three species, Mann-Whitney U) (Figure 1B). We conclude that high gene conversion rates in primate X-chromosome palindrome arms are associated with elevated GC content, consistent with the hypothesis that frequent gene conversion causes an increase in GC content over time.

Previous studies of molecular evolution in sex-chromosome palindromes have used two different genomic regions as controls for comparison to palindrome arms: Flanking sequence (Caceres et al. 2007, Swanepoel et al. 2020), or the unique sequence that separates palindrome arms, called the spacer (Rozen et al. 2003, Geraldes et al. 2010, Hallast et al. 2013). Given that both spacers and flanking sequence comprise unique sequence, their GC content might be expected to be similar. However, we found that the GC content of spacers occupied an intermediate range between arms and flanking sequence, and did not differ significantly from palindrome arms (Figure 1A, B). This finding may be explained by a recent observation that palindrome spacers are structurally unstable on the timescale of primate evolution: For 7/12 palindromes conserved between human and rhesus macaque, spacer sequence could not be aligned between species, and for five palindromes, part of the spacer from one species corresponded to arm sequence in the other (Jackson et al. 2020). We suggest that palindrome spacers display an intermediate level of GC content because some spacers spent part of their evolutionary history in the palindrome arm, where they were subject to higher levels of gene conversion. There were also examples of X-chromosome palindromes for which part of the arm in one species corresponded to flanking sequence in another (e.g., P9 in human and rhesus macaque, Jackson et al. 2020); this phenomenon may explain why flanking sequence has slightly higher GC content than the X chromosome average, as noted above.

Nucleotide replacement patterns in human and chimpanzee X-chromosome palindrome arms demonstrate that GC content has increased in the past seven million years

We next looked for evidence of GC-biased gene conversion based on nucleotide replacement patterns in palindrome arms. For each conserved X-chromosome palindrome, we generated a six-way alignment using both palindrome arms from human, chimpanzee, and rhesus macaque. We then identified nucleotide replacements that occurred in the human lineage by searching for sites with the same base in both human arms (e.g. G/G) and a different base in rhesus macaque and chimpanzee arms (e.g. A/A in both species) (Figure 2A). Such fixed differences can be inferred to have arisen through a substitution in the human lineage, followed by gene conversion between human arms (Hallast et al. 2013, Supplemental Note 1). We compared the base composition of the ancestral base at each site of inferred gene conversion to the derived base. If gene conversion is GC-biased, then derived bases should have a higher GC content than ancestral bases. Indeed, we found that the median GC content of derived bases was 64.5%, compared to 41.5% for ancestral bases ($p<0.0001$, Mann-Whitney U) (Figure 2B). We repeated the same analysis for nucleotide replacements in the chimpanzee lineage, with similar results (62.7% vs 39.4%, $p<0.0001$, Mann-Whitney U) (Figure 2B). In contrast, a comparable analysis examining the GC content of ancestral versus derived sequence for flanking sequence, using three-way alignments between species, revealed little or no significant difference in base-pair composition (Figure 2C). We conclude that GC-biased gene conversion in human and chimpanzee palindrome arms over the past 7 million years has skewed nucleotide replacement patterns, resulting in derived bases being more than one-and-a-half times more GC rich than the ancestral bases that they replaced.

Simulations of palindrome gene conversion are consistent with GC bias of about 0.7

Our interpretation of the results shown in Figure 2B assumes that all nucleotide replacements result from the same series of evolutionary events, i.e., a substitution followed by gene conversion. Although we consider this the most parsimonious explanation for fixed differences found in a single species, other explanations cannot be excluded (see Supplemental Note 1). We therefore devised a series of Markov chain Monte Carlo (MCMC) simulations to model palindrome evolution under different magnitudes of GC-biased gene conversion. These simulations allowed us to examine the expected behaviors of palindrome evolution within reasonable parameters for substitution rate, neutral substitution patterns, gene conversion rate, and the magnitude of GC bias, without requiring assumptions about the specific evolutionary trajectory of each site. Our simulations were designed to achieve three objectives: 1)

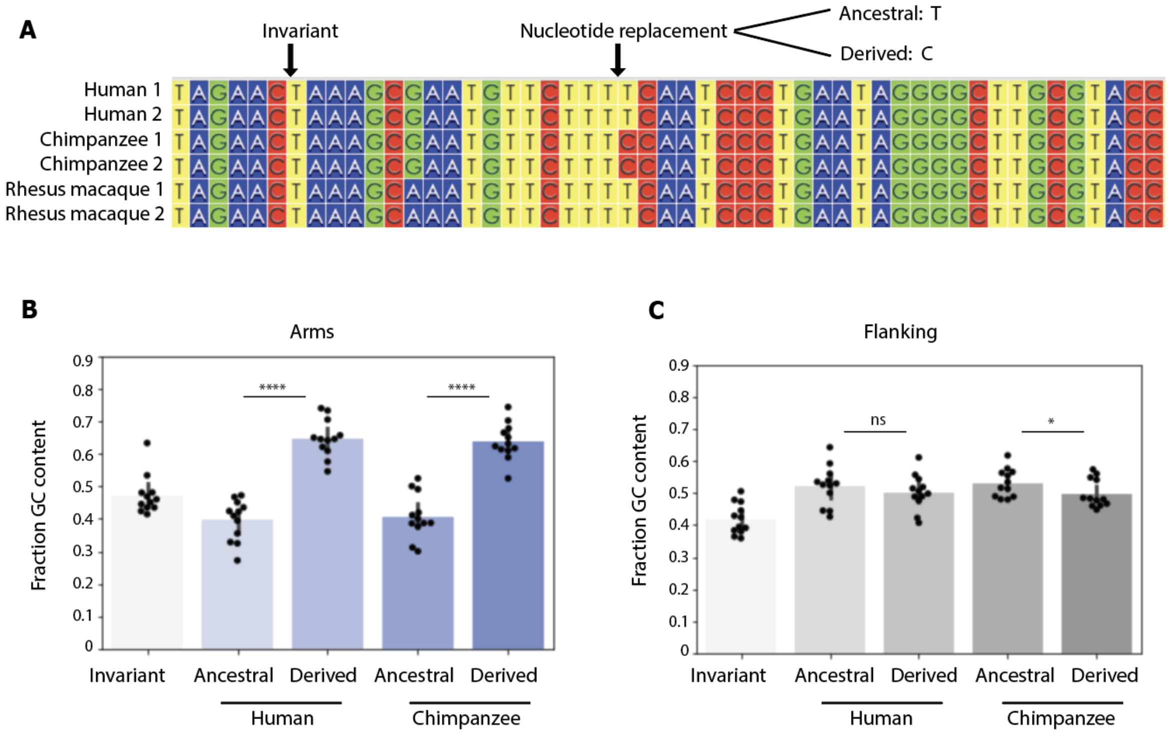


Figure 2: Nucleotide replacements in human and chimpanzee X-chromosome palindrome arms in the past 7 million years have been GC-biased. A) Identification of nucleotide replacements from six-way arm alignments from palindromes conserved between human, chimpanzee, and rhesus macaque. Invariant sites are identical in human, chimpanzee, and rhesus macaque. Alignments generated with ClustalW and visualized using Wasabi (Veidenberg et al. 2016). B,C) Fraction GC content for ancestral versus derived bases. ****p<0.0001, *p<0.05, Mann-Whitney U.

determine the likelihood of observing the pattern of nucleotide replacements shown in Figure 2B in the absence of GC-biased gene conversion, 2) find the magnitude of GC-biased gene conversion most consistent with our results in Figure 2B, and 3) determine what fraction of the elevated GC content seen in primate X-chromosome palindrome arms relative to flanking sequence can be attributed to GC-biased gene conversion. While the simulations shown in Figure 3 were run using identical evolutionary parameters except for the magnitude of GC bias, the effects of altering other parameters are explored in Supplemental Notes 2 and 3; none of these parameter modifications altered the major conclusions of these analyses.

Our simulations model the evolution of a palindrome that was present in the common ancestor of human, chimpanzee, and rhesus macaque, and maintained in all three lineages over 29 million years until the present (see Methods). Briefly, for each iteration, we initialized an ancestral palindrome conforming to the median characteristics of twelve conserved primate X palindromes, including arm length, total GC content, and arm-to-arm identity. We then subjected the ancestral palindrome to rounds of nucleotide substitution followed by gene conversion, with each round representing one generation (Figure 3A). We determined neutral substitution patterns based on alignments of 3.8 Mb gene-masked flanking sequence; our observed pattern showed a strong preference for transitions over transversions, as well as a preference for GC→AT substitutions over AT→GC substitutions, consistent with previous reports (Petrov and Hartl 1999, Zhang and Gerstein 2003, Duret and Arndt 2008; see Methods). We included two branching events to account for the divergence of each lineage, resulting in three evolved palindromes representing those present today in human, chimpanzee, and rhesus macaque. Each simulation described below represents one hundred trials, each simulating twelve independent palindromes, representative of the twelve palindromes described in Figures 1 and 2.

We first used our simulations to determine the likelihood of observing a median difference in GC content between ancestral bases and derived bases as large as that observed in Figure 2B in the absence of GC-biased gene conversion (GC bias = 0.50). For simplicity, we report only the results of evolved human palindromes, given that the palindromes designated as "human" and "chimpanzee" underwent equivalent evolutionary trajectories in our simulations. Out of 100 simulations run without GC-biased gene conversion, we never observed a median difference in GC content between ancestral and derived bases as large as the true median difference of ~23% in primate X-chromosome palindromes (Figure 3B,C, Figure 2B). Indeed, all observed differences were less than zero, demonstrating that in the absence of GC bias, ancestral bases are expected to be more GC-rich than derived bases, reflecting the higher rate of GC→AT substitutions versus AT→GC substitutions (Figure 3B, C). We conclude that our observed pattern of nucleotide replacements in Figure 2B is unlikely ($p<0.01$, bootstrapping) in the absence of GC-biased gene conversion.

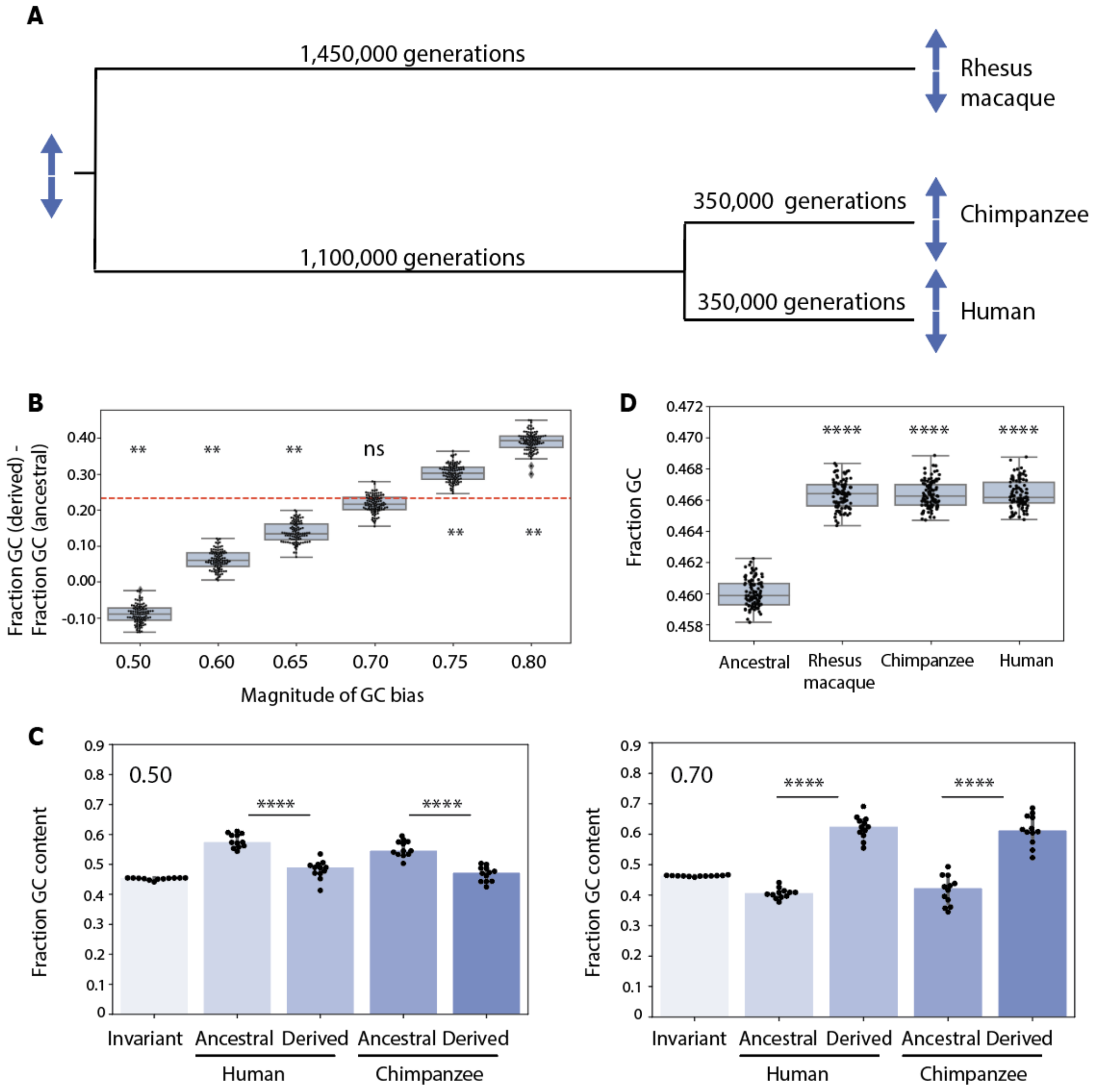


Figure 3: Simulating palindrome evolution with different degrees of GC bias. A) Schematic of simulations. B) Simulated differences between GC content of ancestral and derived bases for six different magnitudes of GC bias. Each dot (n=100 for each magnitude of GC bias) represents the median difference for a set of 12 simulated palindromes. Dashed red line represents true value observed in Figure 2B. **p<0.01, ns = not significant, bootstrapping. C) Fraction GC content for ancestral versus derived bases in simulated palindromes. Results shown for one representative set of 12 palindromes from simulations in Figure 3B. Upper left corner: Magnitude of GC bias. ****p<0.0001, *p<0.05, Mann-Whitney U. D) Fraction GC content for simulated palindrome arms and

ancestral sequence. Magnitude of GC bias = 0.70. Each dot (n=100 for each category) represents median GC content for a set of 12 simulated palindromes. ****p<0.0001, *p<0.05, Mann-Whitney U.

We next asked what magnitude of GC-biased gene conversion could best explain our observed results in Figure 2B. We repeated our simulations using magnitudes of GC bias ranging from 0.60 to 0.80. Simulations using GC bias of 0.75 and 0.80 both produced median differences in GC content between ancestral and derived bases that were significantly larger than our observed value of 23% (39.0% and 31.8%, respectively, p<0.01 for both), while simulations using GC bias of 0.60 and 0.65 produced values that were significantly smaller (6.8% and 13.8%, respectively, p<.01 and p<0.01) (Figure 3C). We found that an intermediate value of 0.70 produced results highly consistent with our observations, with a median difference in GC content between ancestral and derived bases of 21.8% (ns, Figure 3C). We conclude that our results in Figure 2B are best explained by a magnitude of GC bias of approximately 0.70, consistent with previous estimates derived from analyses of human meioses (Williams et al. 2015, Halldorsson et al. 2016).

Finally, we used our simulations to explore the increase in GC content in palindrome arms that would be produced by GC-biased gene conversion of our inferred magnitude, 0.70, over 29 million years of evolution. In particular, we asked what fraction of the difference in GC content observed between palindrome arms and flanking sequence—ranging from 3.6% in rhesus macaque to 4.1% in chimpanzee, after masking protein-coding genes (Figure 1)—could be explained by GC-biased gene conversion over this time scale. We compared the GC content in simulated human, chimpanzee, and rhesus macaque arms to the GC content of the ancestral palindrome. While the three evolved palindromes had significantly higher GC content than the ancestral palindrome, it was by a median magnitude of 0.68%, explaining at most 19% of our observed difference from primate X-chromosome palindromes (Figure 3D). While GC-biased gene conversion leads to a significant increase in GC content over time, our results suggest that an increase of the magnitude we observed in Figure 1 is unlikely to have occurred since the divergence of human, chimpanzee and rhesus macaque. We conclude that either primate X-chromosome palindromes are considerably older than 29 million years, or that other factors contribute to the difference (see Discussion).

DISCUSSION

GC-biased gene conversion is a powerful force that shapes nucleotide composition across mammalian genomes (Galtier et al. 2001, Marais 2003, Duret and Galtier 2009). Previous reports have estimated the magnitude of GC bias in humans to be around 68%, based on the detection of autosomal gene conversion events from three-generation pedigrees (Williams et al. 2015, Halldorsson et al. 2016). Here, we inferred a magnitude of GC bias of around 70% in a unique system of twelve large palindromes conserved on the X chromosome, using a comparative genomic approach combined with evolutionary simulations. The concordance between our results and those of previous studies, including investigations of GC-biased gene conversion in human Y-chromosome palindromes (Hallast et al. 2013, Skov et al. 2017), suggests that the magnitude of GC bias in humans is relatively constant across diverse genomic contexts. From this, we further infer that regional differences in the effects of GC-biased gene conversion—such as the GC-skewed nucleotide replacements that we detect in primate X-chromosome palindrome arms—stem from regional differences in the rate of gene conversion, rather than in the strength of GC bias.

Despite the prediction that high rates of gene conversion could amplify the effects of GC-biased gene conversion, few previous studies have examined the GC content of sex-chromosome palindrome arms. One human X-chromosome palindrome with putative orthologs in other mammals was found to have higher GC content in palindrome arms compared to flanking sequence in all sixteen species studied (Caceres et al. 2007). Results based on six human Y-chromosome palindromes were mixed, with two palindromes showing significantly higher GC content in arms than in spacer, and the other four palindromes showing no significant difference (Hallast et al. 2013). The selection of the spacer for comparison may have reduced the significance of the latter findings, given that we found significant results only from comparing GC content between palindrome arms and flanking sequence. In general, we propose that flanking sequence represents a stronger comparison than spacers for molecular analyses of palindrome evolution, due to the fact some X-chromosome palindrome spacers have mixed evolutionary histories that may include time spent within the palindrome arm (Jackson et al. 2020).

Although we found that GC content in primate X-chromosome palindromes is robustly elevated in palindrome arms versus flanking sequence, simulations show that less than 20% of this increase can be attributed to GC-biased gene conversion since the divergence of the human and rhesus macaque lineages. One possible explanation is that palindromes arose much earlier in primate or mammalian evolution, resulting in additional time to accumulate GC content. However, given the order-of-magnitude difference between our observed results and simulations, we consider under-estimation of palindrome age unlikely to explain the entire discrepancy. We instead propose two mutually compatible possibilities: that GC-rich sequence is more susceptible to palindrome formation, and/or that GC-rich palindromes are more likely to survive over long evolutionary timescales. Both possibilities are bolstered by the fact that although high rates of recombination can elevate GC content over time (Montoya-Burgos et al. 2003, Li et al. 2016), elevated GC content can also increase local rates of recombination (Petes and Merker 2002, Kiktev et al. 2018). Given that palindrome formation is believed to require two recombination events (Kuroda-Kawaguchi et al. 2001), recombinogenic GC-rich sequence may be more likely than AT-rich sequence to form palindromes. Palindromes with high GC content may also have a survival advantage over palindromes with lower GC content, given that high rates of recombination are required to prevent arms from diverging over time. We speculate that both factors—an increased tendency for GC-rich sequence to form and maintain palindromes, combined with further gains in GC content over time from GC-biased gene conversion—contribute to the remarkably GC-rich palindromes we observe in X-chromosome palindromes from human, chimpanzee and rhesus macaque.

MATERIALS AND METHODS

Human mutation rate

Three recent publications used whole-genome shotgun sequencing data from related individuals to calculate human mutation rates of around 1.2×10^{-8} mutations per nucleotide per generation (Roach et al. 2010, Kong et al. 2012, Jónsson et al. 2017). However, these publications used only autosomal data, while the human X chromosome may have a lower mutation rate than autosomes due to its unique evolutionary history (Schaffner 2004). To our knowledge, similarly high-quality estimates of the human X chromosome mutation rate do not exist. To estimate the mutation rate for the human X chromosome, we examined Supplemental Table 4 from Jónsson et al., which provides information for all autosomal and X chromosome mutations detected in their dataset. Supplemental Table 4 reports 2694 X chromosome mutations from 871 probands, or around 3.1 mutations per generation. To calculate the autosomal mutation rate, Jónsson et al divided the number of autosomal mutations per generation by the number of autosomal base pairs with adequate coverage depth in their dataset. We therefore divided 3.1 X-chromosome mutations per generation by the length of the X chromosome in hg38 (156,040,895 base pairs) multiplied by the fraction autosomal coverage (93.3%), which we assume here is similar to the fraction of X chromosome coverage. This approach yielded an estimated human X chromosome mutation rate of 1.06×10^{-8} mutations per nucleotide per generation. This value is about 20% lower than the value calculated by Jónsson et al. for autosomes ($1.28 \text{ x } 10^{-8}$ mutations per nucleotide per generation), consistent with predictions that mutation rates are lower on X chromosomes than on autosomes.

GC content of primate X-chromosome palindromes

We calculated the GC content for each palindrome (Arm 1, spacer, and flanking sequence) using custom Python code. We performed all analyses using clones sequenced by Jackson et al. 2020. For flanking sequence, we used available sequence upstream and downstream of palindrome arms that was present in all three species. For example, if the human clones for a given palindrome contained 3’ sequence that was not sequenced in chimpanzee and rhesus macaque, we trimmed the human sequence to contain only the portion alignable between all three species. Visualizations were generated using ggplot2 in R (Wickham 2016, R Core Team 2020).

Generation of sequence alignments

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 (Jackson et al. 2020).

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 (Jackson et al. 2020).

Simulations

Our simulations were designed to model the evolution of a palindrome present in the common ancestor of human, chimpanzee, and rhesus macaque, and maintained in all three lineages until the present. For each iteration, we initialized an ancestral palindrome with each nucleotide chosen at random based on the median characteristics of conserved primate X-chromosome palindromes (arm length: 37 kb, arm-to-arm identity: 99.953%, GC content: 46%). Each ancestral palindrome then underwent rounds of substitution followed by intra-chromosomal gene conversion, with two branching events to account for the divergence of human, chimpanzee, and rhesus macaque (see below for the calculation of the number of generations in each branch). Simulation parameters included the substitution rate for each evolutionary branch, relative rates for different types of substitutions (i.e., the neutral substitution matrix), and the frequency and GC bias of intra-chromosomal gene conversion, with parameter values selected as described below. Simulations were implemented with custom Python code.

Estimation of generation numbers for simulations

Divergence times for human versus chimpanzee and for human versus rhesus macaque are estimated at about 7 and 29 million years, respectively (Kumar et al. 2017). Generation times for primates vary between species, with estimated generation times around 30 years for humans (Tremblay and Vézina 2000, Matsumura and Forster 2008), 25 years for chimpanzee (Langergraber et al. 2012), and 11 years for rhesus macaque (Gage 1998, Xue et al. 2016). For simplicity, we assumed an intermediate value of 20 years per generation for all branches. Using these values, we estimated a total of 1,450,000 generations for the branch from the common human-chimpanzee-rhesus macaque (HCR) ancestor to rhesus macaque (Branch 1), 1,100,000 generations for the branch from the common HCR ancestor to the common human-chimpanzee (HC) ancestor (Branch 2), and 350,000 generations each for the branches from the common HC ancestor to chimpanzee and to human (Branches 3 and 4, respectively). For a discussion of the impact of generation numbers on our simulations, see Supplemental Note 2.

Estimation of substitution rates for simulations

Substitution rates per generation can be inferred from the nucleotide divergence observed between species of known divergence times. We calculated these rates for each branch of our simulated evolutionary tree as follows:

Substitution rate: Human versus chimpanzee

Palindrome arm divergence: 0.84% (Jackson et al. 2020)

Generations: 350,000 * 2 = 700,000 (see above)

Substitution rate: 1.20 x 10^{-8} substitutions per base per generation.

Substitution rate: Human versus rhesus macaque

Palindrome arm divergence: 5.4% (Jackson et al. 2020)

Generations: 1,450,000 * 2 = 2,900,000 (see above)

Substitution rate: 1.86 x 10^{-8} substitutions per base per generation.

The human-chimpanzee substitution rate was mapped directly onto Branches 3 and 4. The human-rhesus macaque substitution rate was mapped directly onto Branch 1. For Branch 2, we calculated the substitution rate such that the expected divergence along Branch 1 would equal the expected divergence along Branch 2 + Branch 3:

2.7% = 0.42% + (Branch 2 rate * 1,100,000 generations)

Branch 2 rate: 2.07 x 10^{-8} substitutions per base per generation.

Note that for the Branch 2 calculation we assume symmetry of divergence, i.e., divergence between two lineages is divided equally between them.

To confirm that our substitution rates were reasonable, we converted our values to per-year mutation rates assuming a generation time of 20 years, and compared these rates to previously published values. All three of our per-year substitution rates fall within confidence intervals for the same species estimated using autosomal data (Scally and Durbin 2012). Our values fell near the lower end of the confidence intervals, consistent with the prediction that substitution rates on the X chromosome should be slightly lower than on autosomes. Note that our estimated substitution rates per generation differ from the mutation rates reported above for the human X chromosome: Single-generation mutation rates are known to differ from substitution rates over long evolutionary timescales, likely due to a recent slowdown in the mutation rate in humans and great apes (Scally et al. 2012). For a discussion of the impact of mutation rates on our simulations, see Supplemental Note 2.

Estimation of neutral substitution matrix for simulations

Neutral substitution patterns between species do not follow a uniform distribution: Transitions are more common than transversions, and substitutions that replace a strong base (GC) with a weak base (AT) are more common than substitutions in the opposite direction (Petrov and Hartl 1999, Zhang and Gerstein 2003, Duret and Arndt 2008). In addition to branch-specific substitution rates, we therefore also sought to determine a reasonable pattern of neutral substitutions for our simulations. We identified neutral substitutions using alignments from 3.8 Mb of gene-masked sequence flanking X-chromosome palindromes, using parsimony to infer substitution events in human and chimpanzee with rhesus macaque as an outgroup. From this we calculated seven different substitution rates:

Substitution	Substitution rate (substitutions/nt/ generation)
AT →TA	1.64×10^{-9}
AT →CG	1.93×10^{-9}
AT →GC	8.04×10^{-9}
CG → GC	2.98×10^{-9}
CG →AT	3.22×10^{-9}
CG → TA (non-CpG)	1.02×10^{-8}
CG→TA (CpG)	9.58×10^{-8}

The overall neutral substitution rate (K) can be calculated as described in Duret and Arndt 2008:

$$K = F_{GC} (R_{CG\rightarrow GC} + R_{CG\rightarrow AT} + R_{CG\rightarrow TA\ (non\text{-}CpG)}) + F_{AT} (R_{AT\rightarrow TA} + R_{AT\rightarrow CG} + R_{AT\rightarrow GC}) + F_{CpG} (R_{CG\rightarrow TA\ (CpG)})$$

where F_{GC}, F_{AT} and F_{CpG} represent the frequencies of each site and $R_{AA\rightarrow BB}$ represents the frequencies of each substitution. Using the substitution rates above combined with the observed frequencies of each site (F_{GC}: 0.396, F_{AT}: 0.596, F_{CpG}: 0.08), we found that $K = 1.42 \times 10^{-8}$ substitutions per nucleotide per generation. We then combined the categories CG→TA (non-CpG) and CG→TA (CpG) into a single rate CG→TA as follows:

$R_{CG\rightarrow TA} = [F_{GC} (R_{CG\rightarrow TA\ (non\text{-}CpG)}) + F_{CpG} (R_{CG\rightarrow TA\ (CpG)})\ ]/ (F_{GC} + F_{CpG}) = 1.18 \times 10^{-8}$ substitutions per nucleotide per generation

We do not expect combining rates for CpG and non-CpG substitutions to affect either of our simulation output metrics (Figures 3B-C: Fraction GC derived – Fraction GC ancestral at sites of nucleotide replacements; Figure 3D: Fraction GC overall) because these metrics are agnostic to the context in which each fixed nucleotide replacement occurred.

The substitution rates above were calculated using substitutions in flanking sequence since the divergence of chimpanzee and human; however, each evolutionary branch in our simulation has a different overall substitution rate (see section above). For each branch, we therefore divided the substitution rates above by the original overall substitution rate of 1.42×10^{-8} substitutions per nucleotide per generation, then multiplied by the branch-specific overall substitution rate. This kept the relative ratios between different substitution types constant, while accounting for different overall substitution rates in each branch. The effects of reasonable alterations of this neutral substitution matrix, including adjusting for possible under-estimation of the CpG substitution rate due to artifacts of parsimony, are described in Supplemental Note 3.

DATA ACCESS

BAC sequences used for this study are available from GenBank (https://www.ncbi.nlm.nih.gov/) under accession numbers listed in Supplemental Table 2. The authors affirm that all other data necessary for confirming the conclusions of the article are present within the article, figures and tables. Code used to generate the simulated data can be found at https://github.com/ejackson054/GC-biased-gene-conversion.

REFERENCES

Borges R, Szöllősi GJ, Kosiol C. 2019. Quantifying GC-biased gene conversion in great ape genomes using polymorphism-aware models. *Genetics* **212:** 1321–1336.

Caceres M, McDowell JC, Gupta J, Brooks S, Bouffard GG, Blakesley RW, Green ED, Sullivan RT, Thomas JW. 2007. A recurrent inversion on the eutherian X chromosome. *PNAS* **104**: 18571–18576.

Clément Y, Arndt PF. 2011. Substitution patterns are under different influences in primates and rodents. *Genome Biol Evol* **3**: 236–245.

Duret L. 2006. The GC content of primates and rodents genomes is not at equilibrium: A reply to Antezana. *J Mol Evol* **62:** 803–806.

Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet* **4:** e1000071.

Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet* **10:** 285–311.

Gage TB. 1998. The comparative demography of primates: with some comments on the evolution of life histories. *Annu Rev Anthropol* **27:** 197–221.

Galtier N, Piganeau G, Mouchiroud D, Duret L. 2001. GC-content evolution in mammalian genomes: The biased gene conversion hypothesis. *Genetics* **159:** 907–911.

Galtier, N., Duret, L., Glémin, S., and Ranwez, V., 2009 GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. Trends Genet. 25: 1–5.

Geraldes A, Rambo T, Wing RA, Ferrand N, Nachman MW. 2010. Extensive gene conversion drives the concerted evolution of paralogous copies of the SRY gene in European rabbits. *Mol Biol Evol* **27:** 2437–2440.

Hallast P, Balaresqu P, Bowden GR, Ballereau S, Jobling MA. 2013. Recombination dynamics of a human Y-chromosomal palindrome: Rapid GC-biased gene conversion, multikilobase conversion tracts, and rare inversions. *PLoS Genet* **9**: e1003666.

Halldorsson BV, Hardarson MT, Kehr B, Styrkarsdottir U, Gylfason A, Thorleifsson G, Zink F, Jonasdottir A, Jonasdottir A, Sulem P, et al. 2016. The rate of meiotic gene conversion varies by sex and age. *Nat Genet* **48**: 1377–1384.

Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Graves T, Fulton RS, Dugan S, Ding Y, Buhay CJ, Kremitzki C, et al. 2012. Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* **483**: 82–86.

Hughes JF, Skaletsky H, Pyntikova T, Graves TA, van Daalen SKM, Minx PJ, Fulton RS, McGrath SD, Locke DP, Friedman C, et al. 2010. Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* **463**: 536–539.

Hughes JF, Skaletsky H, Pyntikova T, Koutseva N, Raudsepp T, Brown LG, Bellott DW, Cho T-J, Dugan-Rocha S, et al. 2020. Sequence analysis in *Bos taurus* reveals pervasiveness of X-Y arms races in mammalian lineages. *Genome Res* **30**: 1716–1726.

Jackson EK, Bellott DW, Cho T-J, Skaletsky H, Hughes JF, Pyntikova T, Page DC. 2020 Large palindromes on the primate X Chromosome are preserved by natural selection. bioRxiv doi: https://doi.org/10.1101/2020.12.29.424738.

Jónsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, Hardarson MT, Hjorleifsson KE, Eggertsson HP, Gudjonsson SA, et al. 2017. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549:** 519–522.

Kiktev DA, Sheng Z, Lobachev KS, Petes TD. 2018. GC content elevates mutation and recombination rates in the yeast *Saccharomyces cerevisiae*. *PNAS* **115:** E7109–E7118.

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A, et al. 2012. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488:** 471–475.

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: A resource for timelines, timetrees, and divergence times. *Mol Biol Evol* **34:** 1812–1819.

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.

Kuroda-Kawaguchi T, Skaletsky H, Brown LG, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Silber S, Oates R, Rozen S, et al. 2001. The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. *Nature Genet* **29**: 279–186.

Langergraber KE, Prüfer K, Rowney C, Boesch C, Crockford C, Fawcett K, Inoue E, Inoue-Muruyama M, Mitani JC, Muller MN, et al. 2012. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *PNAS* **109:** 15716–15721.

Li F-W, Kuo L-Y, Pryer KM, Rothfels CJ. 2016. Genes translocated into the plastid inverted repeat show decelerated substitution rates and elevated GC content. *Genome Biol Evol* **8**: 2452–2458.

Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz M. 2008. High-resolution mapping of meiotic crossovers and noncrossovers in yeast. *Nature* **454**: 479–485.

Marais G. 2003. Biased gene conversion: Implications for genome and sex evolution. *Trends Genet* **19:** 330–338.

Matsumura S, Forster P. 2008. Generation time and effective population size in Polar Eskimos. *Proc R Soc B* **275:** 1501–1508.

Mueller JL, Skaletsky H, Brown LG, Zaghlul S, Rock S, Graves T, Auger K, Warren WC, Wilson RK, Page DC. 2013. Independent specialization of the human and mouse X chromosomes for the male germline. *Nature Genet* **45**: 1083–1087.

Montoya-Burgos JI, Boursot P, Galtier N. 2003. Recombination explains isochores in mammalian genomes. *Trends Genet* **19:** 128–130.

Muyle A, Serres-Giardi L, Ressayre A, Escobar J, Glémin S. 2011. GC-biased gene conversion and selection affect GC content in the *Oryza* genus (rice). *Mol Bio Evol* **28:** 2695–2706.

Odenthal-Hesse L, Berg IL, Veselis A, Jeffreys AJ, May CA. 2014. Transmission distortion affecting human noncrossover but not crossover recombination: A hidden source of meiotic drive. *PLoS Genet* **10**: e1004106.

Petes TD, Merker JD. 2002. Context dependence of meiotic recombination hotspots in yeast: The relationship between recombination activity of a reporter construct and base composition. *Genetics* **162:** 2049–2052.

Petrov DA, Hartl DL. 1999. Patterns of nucleotide substitution in *Drosophila* and mammalian genomes. *PNAS* **96:** 1475–1479.

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

Roach JC, Glusman G, Smit AFA, Huff CD, Hubley R, Shannon RT, Rowen L, Pant KP, Goodman N, Bamshad M, et al. 2010. Analysis of genetic inheritance in a family quartet by whole genome sequencing. *Science* **328:** 636–639.

Rozen S, Skaletsky H, Marszalek 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.

Scally A, Durbin R. 2012. Revising the human mutation rate: Implications for understanding human evolution. *Nat Rev Genet* **13**: 745–753.

Scally A, Dutheil JY, Hillier LW, Jordon GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, et al., 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* **483:** 169–175.

Schaffner SF. 2004. The X chromosome in population genetics. *Nat Rev Genet* **5:** 43–51.

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown 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.

Skov L, The Danish Pan Genome Consortium, Schierup MH. 2017. Analysis of 62 hybrid assembled human Y chromosomes exposes rapid structural changes and high rates of gene conversion. *PLoS Genet*. **13**: e1006834.

Smeds L, Mugal CF, Qvarnström A, Ellegren H. 2016. High-resolution mapping of crossover and non-crossover recombination events by whole-genome re-sequencing of an avian pedigree. *PLoS Genet* **12**: e1006044.

Soh YQS, Alfoldi J, Pyntikova T, Brown LG, Minx PJ, Fulton RS, Kremitzki 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.

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.

Tremblay M, Vézina H. 2000. New estimates of intergenerational time intervals for the calculation of age and origins of mutations. *Am J Hum Genet* **66:** 651–658.

Veidenberg A, Medlar A, Löytynoja A. 2016. Wasabi: An integrated platform for evolutionary sequence analysis and data visualization. *Mol Biol Evol* **33:** 1126–1130.

Warburton PE, Giordano J, Cheung F, Gelfand Y, Benson G. 2004. Inverted repeat structures 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.

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

Williams AL, Genovese G, Dyer T, Altemose N, Truax K, Jun G, Patterson N, Myers SR, Curran JE, Duggirala R, et al. 2015 Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *eLife* **4**: e04637.

Xue C, Raveendran M, Harris RA, Fawcett GL, Liu X, White S, Dahdouli M, Deiros DR, Below JE, Salerno W, et al. 2016. The population genomics of rhesus macaques (*Macaca mulatta*) based on whole-genome sequences. *Genome Res* **26**: 1651–1662.

Zhang Z, Gerstein M. 2003. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Res*. **31**: 5338–5348

SUPPLEMENTAL FIGURES AND TABLES

Supplemental Note 1: Inference of gene conversion from conserved palindromes in human, chimpanzee, and rhesus macaque
Supplemental Note 2: Effects of altering substitution rates and generation times for simulations
Supplemental Note 3: Effects of altering the neutral substitution matrix for simulations

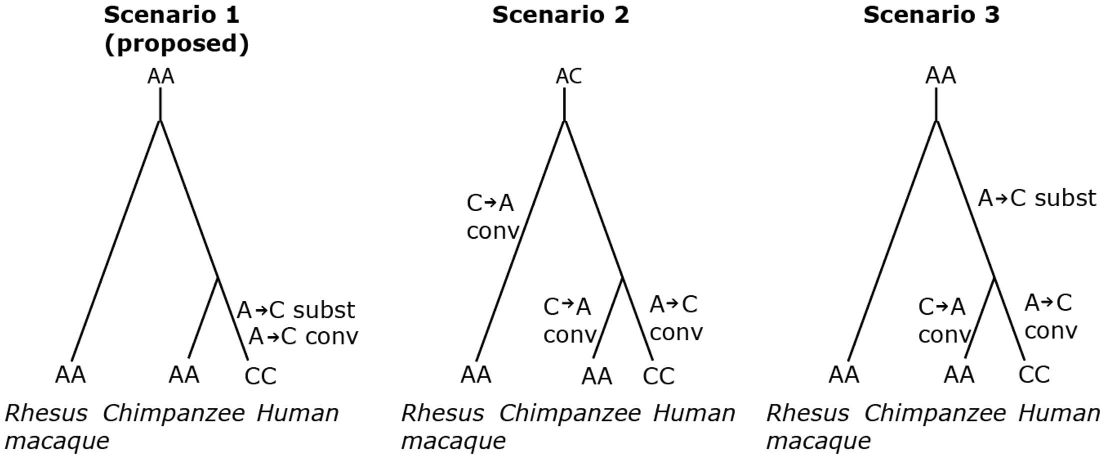
Supplemental Table 1: GC content is elevated in primate X-palindrome arms relative to flanking sequence
Supplemental Table 2: GenBank accession numbers for chimpanzee and rhesus macaque clones analyzed for this project

Supplemental Note 1: Inference of gene conversion from conserved palindromes in human, chimpanzee, and rhesus macaque

Using the logic of Hallast et al. 2013, we inferred gene conversion events from fixed nucleotide replacements in X-palindrome arms that occurred in either human or chimpanzee, using rhesus macaque as an outgroup. For simplicity, the scenarios below describe a fixed replacement in the human lineage. We propose that fixed nucleotide replacements result from a substitution in humans after divergence from chimpanzee, followed by gene conversion that homogenizes the substitution between arms (Scenario 1). In theory, other scenarios could lead to the same result. In one alternative scenario, the ancestral palindrome was heterozygous at the site in question, with gene conversion occurring in one direction in rhesus macaque and chimpanzee, and the opposite direction in human (Scenario 2). We consider this scenario highly unlikely because it requires the initial site to remain heterozygous for 1.1 million generations before undergoing gene conversion in human and chimpanzee (see Figure 3A). Given our inferred intrachromosomal gene conversion rate of 4.5 x 10^{-5} events per nucleotide per generation, the probability of any given site not undergoing gene conversion over 1.1 million generations is (1 - 4.5 x 10^{-5}) ^ 1100000, which is effectively zero (<2.22 x 10^{-308}).

We also considered a scenario in which the initial substitution occurred in the human-chimpanzee common ancestor, then underwent gene conversion in opposite directions in human and chimpanzee (Scenario 3). Given that X palindromes have on average only 1 difference between arms for every 2200 nucleotides, this scenario could explain at most observed nucleotide replacements in 1 out of 2200 positions in X-palindrome arms (0.045%), if all heterozygous sites resolved in opposite directions in each lineage. We observed nucleotide replacements in 2567 out of 409,579 positions in X-palindrome arms (0.65%), suggesting that Scenario 3 can account for no more than 7% (0.045% / 0.65%) of our observations.

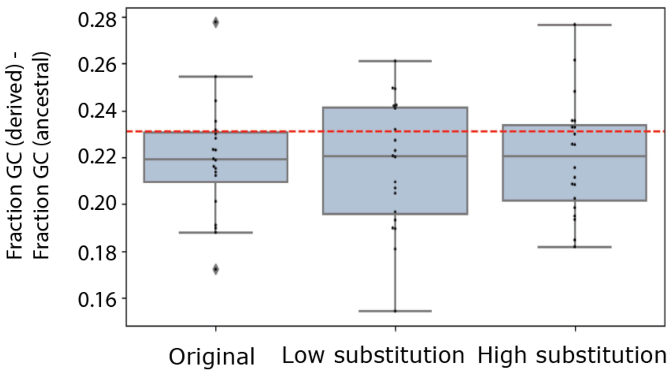
Finally, we used evolutionary simulations with event tracing to estimate what fraction of fixed nucleotide replacements in human, chimpanzee, and rhesus macaque would arise through each scenario under reasonable evolutionary parameters (see Figure 3, Methods). We found that the vast majority of fixed nucleotide replacements (93.3%) arose through Scenario 1, while around 2.5% arose through Scenario 3. As predicted, we never observed fixed replacements arising from Scenario 2. The remaining fixed nucleotide replacements (4.2%) resulted from other scenarios that involved multiple substitution events. Importantly, our conclusions in Figure 3 are agnostic to the method by which each fixed nucleotide replacement arose, and depend only on the ability of a given set of evolutionary parameters to reproduce the replacement patterns seen in Figure 2B.



Supplemental Note 2: Effects of altering substitution rates and generation times for simulations

Although our calculations of generation numbers and substitution rates for each branch of the evolutionary tree were based on estimates of divergence times and divergence times from recent literature, these values nevertheless are subject to uncertainty. For example, while we estimated a substitution rate of 1.20 x 10^{-8} substitutions per nucleotide per generation for Branches 3 and 4 based on observed divergence of 0.84% between human and chimpanzee and assuming 700,000 generations, the observed divergence between human and chimpanzee could instead have resulted from a higher substitution rate combined with a lower number of generations, or vice versa.

To test the effect of this uncertainty on our simulations, we considered two fairly extreme cases: 1) The true substitution rates in all lineages were twice as high as we estimated, while the true number of generations was halved; 2) The true substitution rates in all lineages were half as high as we estimated, while the true number of generations was doubled. We then repeated our simulations for both of these cases (”high substitution” and “low substitution”) and compared our results to the results obtained using our original calculations. We find that these alterations make no difference to our inference of the magnitude of GC bias at 0.70 (difference between observed and simulated results not significant, p>0.05).

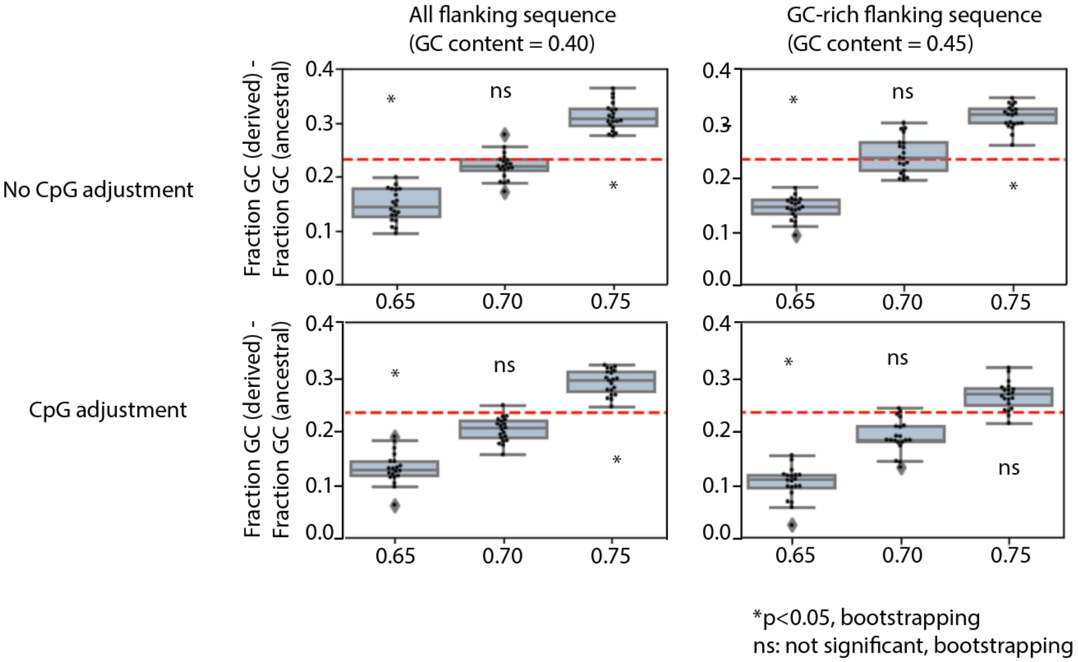


Branch 1:	Rate (subst/nt/gen)	1.86 x 10^{-8}	0.94 x 10^{-8}	3.72 x 10^{-8}
	# generations	1,450,000	2,900,000	725,000
Branch 2:	Rate (subst/nt/gen)	2.07 x 10^{-8}	1.04 x 10^{-8}	4.14 x 10^{-8}
	# generations	1,100,000	2,200,000	550,000
Branch 3/4:	Rate (subst/nt/gen)	1.20 x 10^{-8}	0.60 x 10^{-8}	2.40 x 10^{-8}
	# generations	350,000	700,000	175,000

Supplemental Note 3: Effects of altering the neutral substitution matrix for simulations

We considered the effect that altering the neutral substitution matrix would have on our simulations. In particular, we considered two possible limitations of our inferred neutral substitution matrix. First, our matrix was derived from flanking X chromosome sequence with a median GC content around 40%, while palindrome arms have a median GC content around 46%. Previous work has shown that substitution patterns can differ based on regional GC content, with regions with high GC content showing a lower rate of strong (GC) to weak (AT) mutations (Duret and Arndt 2008). Using a matrix derived from sequence with a lower GC content could in theory lead to over-estimation of AT mutation bias, and subsequent over-estimation of the GC conversion bias required to balance it. We therefore re-calculated our neutral substitution matrix using a subset of flanking sequence (1.3 Mb) with a total GC content of 45% and repeated our simulations (figure below, top panel). Our inference of GC bias remained unchanged: Using 20 simulations of 12 palindromes each, our observed results were still most consistent with a GC bias magnitude of 0.70 (difference between observed and simulated results not significant, p>0.05).

The second limitation we considered was our use of parsimony to infer substitution events for our neutral substitution matrix. While this is appropriate for most types of substitutions on the time scale of human-chimpanzee evolution, it has been shown to under-estimate the frequency of CpG substitutions, which occur more frequently than other substitutions and can thus occur twice at the same site (Duret 2006). Under-estimation of CpG substitutions could lead us to under-estimate the AT mutation bias, and therefore under-estimate the magnitude of GC conversion bias. To determine the impact on our simulations, we re-estimated our rate of CpG substitutions to align with the values found by Duret and Arndt 2008, who found that CpG substitutions (CG→AT at CpG sites) are 14 times more common than the same substitutions at non-CpG sites using a maximum-likelihood method. While adjusting the frequency of CpG adjustments shifted our simulated differences in GC content between derived and ancestral bases slightly downwards, our results were still consistent with a magnitude of GC bias if 0.70 (difference between observed and simulated results not significant, p>0.05) We conclude that our results are robust to reasonable shifts in the neutral substitution matrix.



Supplemental Table 1: GC content is elevated in primate X-palindrome arms relative to flanking sequence. P-values are from chi-square test with Yates correction.

Human

Palindrome	Arm 1 length	Flanking length	Arm 1 GC	Flanking GC	p-value
P2	24849	83517	44.07	39.22	1.80E-266
P3	36398	83392	47.08	37.52	0
P6	37939	203740	48.05	43.91	0
P7	26581	140783	43.69	38.27	0
P8	57336	56225	45.86	47.62	3.50E-07
P9	119125	92765	43.19	40.21	9.95E-92
P10	9193	101002	50.36	42.21	1.34E-142
P11	140582	176247	41.72	39.47	0
P21	46811	88685	46.83	42.73	8.19E-160
P24	10050	92241	61.11	50.83	3.14E-138
P25	35448	288582	53.01	48.46	3.76E-310
P26	50037	250270	41.54	37.61	7.62E-155

Chimpanzee

Palindrome	Arm 1 length	Flanking length	Arm 1 GC	Flanking GC	p-value
P2	25099	82455	43.95	39.36	2.25E-257
P3	36247	83212	47.1	37.53	0
P6	34932	196227	48.89	44.14	0
P7	28529	140226	44.2	38.26	0
P8	53145	65431	45.94	47.86	1.05E-14
P9	119105	88148	43.08	39.74	1.07E-116
P10	7591	100457	46.62	42.05	2.27E-163
P11	160191	185640	41.63	39.19	0
P21	36887	94054	46.6	42.53	6.08E-227
P24	10712	90734	59.62	50.86	2.49E-63
P25	34937	289807	52.74	47.95	2.79E-303
P26	46893	240223	41.2	37.91	2.17E-138

Rhesus macaque

Palindrome	Arm 1 length	Flanking length	Arm 1 GC	Flanking GC	p-value
P2	42623	91615	44.32	39.06	0
P3	34782	96783	47.07	37.85	0
P6	15564	221983	48.34	44.79	0
P7	24152	187629	43.93	38.3	0
P8	38447	89961	47.77	45.91	0
P9	81966	152376	43.61	40.54	0
P10	6561	117616	46.14	41.55	1.65E-172
P11	105994	188318	42.46	39.55	0
P21	21963	89617	44.35	42.31	8.97E-267
P24	7673	90082	61.05	50.28	0
P25	101313	273906	51.23	47.45	0
P26	45037	198305	41.74	37.11	0

Supplemental Table 2: GenBank accession numbers for chimpanzee and rhesus macaque clones analyzed for this project

Clone	Accession	Palindrome
CH250-106M20	AC280444	P7
CH250-114J18	AC280531	P9/P10
CH250-136N6	AC280430	P3
CH250-137I15	AC280580	P26
CH250-168E3	AC280520	P21
CH250-174F12	AC280455	P21
CH250-191K20	AC280440	P2
CH250-197O3	AC280457	P6
CH250-228D11	AC280538	P11
CH250-236O7	AC280541	P8
CH250-371L16	AC280564	P9
CH250-398K19	AC280504	P21
CH250-412K19	AC280483	P6
CH250-417G7	AC280442	P26
CH250-424H13	AC280467	P25
CH250-462M8	AC280473	P21
CH250-487N16	AC280453	P26
CH250-491H11	AC280464	P24
CH250-493M11	AC280503	P25
CH250-498I16	AC280468	P11
CH250-503C21	AC280489	P25
CH250-503N19	AC280524	P25
CH250-540J3	AC280456	P11
CH250-541H5	AC280425	P3
CH250-563M7	AC280492	P3
CH250-80G22	AC280568	P9
CH250-87B7	AC280549	P7
CH250-94G2	AC280518	P3
CH251-161L14	AC280465	P7
CH251-177B21	AC280525	P2
CH251-183G21	AC280578	P8
CH251-239P10	AC280533	P26
CH251-240O17	AC280544	P6
CH251-277H18	AC280556	P7
CH251-285D14	AC280499	P26
CH251-346A10	AC280446	P25

CH251-397P16	AC280507	P3
CH251-4M24	AC280567	P11
CH251-504H5	AC280540	P10
CH251-506D4	AC280415	P6
CH251-514B7	AC280488	P8
CH251-542A6	AC280462	P10
CH251-565G15	AC280459	P21
CH251-635P13	AC280558	P11
CH251-651H9	AC280522	P11
CH251-657L4	AC280546	P9/P10
CH251-658J15	AC280445	P8
CH251-65E21	AC280530	P24
CH251-671I19	AC280576	P3
CH251-677L24	AC280565	P24/P25
CH251-737G9	AC280491	P9/P10
CH251-73C22	AC280423	P2

CHAPTER 4:

Conclusions

Massive palindromes are a common feature of mammalian X and Y chromosomes, yet their biology remains poorly understood. Prior work indicated that palindromes arose convergently on the mouse and human X chromosomes (Mueller et al. 2008, Mueller et al. 2013), as well as several mammalian Y chromosomes (Skaletsky et al. 2003, Hughes et al. 2010, Hughes et al. 2012, Soh et al. 2014, Hughes et al. 2020), yet insights into palindrome evolution have been limited by the dearth of palindromes conserved between species. Through the generation of high-quality reference sequence for twelve X-chromosome palindromes conserved between human, chimpanzee and rhesus macaque, we have demonstrated that a subset of primate X palindromes have persisted over tens of millions of years, shaped by natural selection that preserves palindrome gene families as well as the neutral effects of GC-biased gene conversion. I will conclude by briefly reviewing major conclusions from Chapters 2 and 3, as well as new avenues for future work.

CONCLUSIONS

Prior to the work described in this thesis, sex-chromosome palindromes were viewed primarily as sites of evolutionary innovation and rapid turnover between species (Skaletsky et al. 2003, Hughes et al. 2010, Mueller et al. 2013, Soh et al. 2014, Hughes et al. 2020), as well as sources of pathogenic human structural rearrangements (Lakich et al. 1993, Small et al.1997, Lange et al. 2009, Scott et al. 2010). In Chapter 2, we have challenged this view through the discovery that twelve primate X-chromosome palindromes have existed for at least 25 million years, in some cases with remarkably little structural change between species. This difference in palindrome stability may be partly explained by genomic context: First, the X chromosome as a whole has more highly conserved gene content and gene order between species than the Y chromosome (Ohno 1967), which has been the site of most previous palindrome studies (Rozen et al. 2003, Skaletsky et al. 2003, Hughes et al. 2010, Hughes et al. 2012, Soh et al. 2014, Hughes et al. 2020); second, all primate X palindromes described to date are present in a single copy, which limits opportunities for non-allelic recombination that are abundant among duplicated palindromes on the mouse X chromosome (Mueller et al. 2008) and on primate Y chromosomes

(Skaletsky et al. 2003, Hughes et al. 2010, Hughes et al. 2012). In contrast to the expectation that short lifespans and frequent remodeling might be characteristics of all sex-chromosome palindromes, our work illustrates a highly distinct evolutionary trajectory for a subset of primate X-chromosome palindromes.

The evolutionary persistence of primate X-chromosome palindromes results in part from natural selection acting on protein-coding gene families. Previous literature has emphasized the rapid protein evolution of testis-biased gene families on the primate X chromosome (Stevenson et al. 2007) as well as therapeutic opportunities to exploit the unique expression patterns of many human X-palindrome gene families in the testis and in cancerous tumors (Simpson et al. 2005, Sahin et al. 2020). Our identification of structural and molecular signatures of purifying selection suggests that X-palindrome gene families have conserved, as yet undiscovered functions across primates, revealing a new and more stable dimension to these gene families. While it is tempting to speculate that the functions of different X-palindrome gene families may be similar, we find that the expression patterns of human X-palindrome genes are more diverse than those of human Y-palindrome genes, with a subset of X-palindrome genes showing broad expression across the human body (e.g. *MAGED4* and *FAM156*). Even among testis-biased gene families, expression patterns vary across different stages of spermatogenesis, suggesting that not all X-palindrome genes function during the same stage of germ cell development (Chapter 2). Determining the individual functions of human X-palindrome gene families will likely require association studies using large datasets that link genotype with phenotype, an additional challenge given the frequent exclusion of the X chromosome from such analyses (Wise et al. 2013, see Future Directions).

How might evolution proceed differently within conserved X-chromosome palindromes than in nearby single-copy sequence? The work presented in this thesis supports two unique evolutionary patterns within palindromes: Localized structural instability around the center of palindrome symmetry (Chapter 2), and nucleotide replacements between species that are strongly skewed towards GC bases over AT bases, which likely results from ongoing GC-biased gene conversion between palindrome arms (Chapter 3). Our results in Chapter 2 suggest that localized structural instability is counterbalanced by natural selection that preserves gene families within palindrome spacers and inner arms, preventing degradation of the palindrome structure that might otherwise result from repeated rounds of internal deletions and rearrangements. To date, however, the interaction between natural selection and GC-biased gene conversion is less clear. Previous studies have shown that GC-biased gene conversion can promote the fixation of mildly deleterious GC alleles in humans (Necşulea et al. 2011, Lachance and Tishkoff 2014) and other primates (Galtier et al. 2009), yet with high-quality sequence from only three species, we lacked power to test for this phenomenon among conserved X palindromes and their associated gene families. Future studies that take advantage of genetic diversity within species, as well as X palindromes sequenced among additional primates, will be better equipped to determine the effects of GC-biased gene conversion on the nucleotide evolution of X-palindrome arm gene families.

FUTURE DIRECTIONS

Increased representation of palindromes in mammalian sex-chromosome reference sequences

To our knowledge, the study presented in Chapter 2 contains the largest set of conserved sex-chromosome palindromes published to date. However, our results suggest that palindrome evolution is shaped by a complex mixture of factors that differ between individual palindromes, including the presence or absence of protein-coding gene families, the degree of selective constraint on those gene families, the copy number of the palindrome, and the size of the palindrome spacer. Discerning additional principles of palindrome evolution will therefore require a much larger pool of conserved palindromes, in which the effects of each of these factors can be studied more robustly.

I project that the availability of long-read sequencing technologies will lead to a rapid increase in the representation of palindromes in mammalian reference genomes over the next decade. Reference genome contiguity has increased substantially for many species in recent years; one striking example is the rhesus macaque X chromosome, where the number of assembly gaps decreased from 839 (Zimin et al. 2014) to only 18 (Warren et al. 2020) during the course of the research presented in this thesis. The first gapless assembly of a vertebrate chromosome—a telomere-to-telomere assembly of the human X chromosome, including the centromere, which is notoriously challenging to sequence due to its long arrays of short tandem repeats—was recently produced using a mixture of nanopore, PacBio and Illumina reads (Miga et al. 2020) Palindromes may still be incorrectly assembled even in highly contiguous genomes that were generated using a whole-genome shotgun approach, necessitating a targeted clone-based approach like SHIMS 3.0 for finishing (see Chapter 2 and Appendix). However, improvements in palindrome representation for long-read WGS assemblies compared to short-read WGS assemblies (Chapter 2) hint at a possible future in which clones are no longer needed. Indeed, nanopore reads can now reach lengths > 800 kb (Jain et al. 2018); reads of this length could resolve every palindrome on the human X chromosome, and all but one palindrome on the human Y chromosome. Given the high frequency at which mammalian long-read WGS assemblies are being produced (see Gordon et al. 2016, Bickhart et al. 2017, Jain et al. 2018, Low et al. 2019, and others), the development of long-read WGS protocols capable of consistently resolving palindromes and other complex genomic structures could rapidly expand the presence of palindromes in mammalian reference genomes.

Generation of accurate reference sequence for additional orthologs of the twelve conserved primate X-chromosome palindromes described in Chapters 2 and 3 could shed light on several unresolved questions. We found that twelve palindromes are at least 25 million years old based on their conservation between human and rhesus macaque. However, they could be much older, as some palindromes show little structural change between species, making it plausible that palindromes could persist over longer timespans (Chapter 2), and GC content in palindrome arms is higher than would be expected from 25 million years of GC-biased gene conversion, which also hints at earlier origins (Chapter 3). High-quality X-chromosome reference sequence from more distantly diverged mammalian species, including marmoset (43 million years), bull (96 million years), and even opossum (159 million years), would help to reveal the true depth of palindrome conservation (Kumar et al. 2017). In addition, sequencing of X palindromes from more closely related species, including gorilla and orangutan, could help to determine the frequency of structural changes such as inversions and spacer deletions, while also increasing statistical power to detect purifying or positive selection on X-palindrome gene families. Such studies will need to carefully consider the effects of GC-biased gene conversion on signals of molecular evolution in palindrome arms (Chapter 3), which can sometimes resemble positive selection due to the directional fixation of GC alleles (Berglund et al. 2009, Ratnakumar et al. 2010).

The previous paragraph describes opportunities to further illuminate the evolution of the twelve primate X-chromosome palindromes described in this thesis. However, sequencing palindromes from additional mammalian species may also reveal novel examples of conserved palindromes, providing opportunities to ask how well the principles discerned from primate X palindromes generalize to independent experiments of nature. One key question is whether rearrangements around the center of palindrome symmetry are common in other systems of sex-chromosome palindromes, and in particular, whether they are specific to X-chromosome palindromes (suggesting errors that occur during female meiotic recombination) or occur in both X- and Y-chromosome palindromes (suggesting errors that occur during palindrome arm-to-arm recombination, or during mitosis). Other hypotheses that could be tested in new systems include whether single-copy palindromes are indeed more highly conserved than partially or completely duplicated palindromes, and whether the density of protein-coding genes positively correlates with palindrome survival. Finally, some follow-up questions will become tractable only with the generation of high-quality reference sequence for complete X chromosomes, as opposed to the targeted sequencing approach used for this project. For example, our results in Chapter 3 raise the question of to what extent the high GC content of conserved X palindromes was present in ancestral sequence prior to duplication, versus acquired through GC-biased gene conversion. Answering this question will require X chromosomes with high-quality sequence for both single-copy and palindromic regions, so that the GC content of single-copy regions can be used for evolutionary comparisons. In addition, there are likely to be many mammalian X-chromosome palindromes that lack orthologs in human, and which will therefore only be identified through high-quality assemblies of entire X chromosomes.

Functions of human X-palindrome gene families

One of the most obvious and compelling directions for future inquiry is deciphering the functions of human X-palindrome gene families. As mentioned above, the expression patterns of human X-palindrome gene families suggest somewhat diverse functions: In contrast to mouse X-amplicon genes, which are expressed predominantly in post-meiotic germ cells (Mueller et al. 2008), different human X-palindrome gene families are expressed across varying stages of human spermatogenesis, and around 1/3 are expressed more broadly across the human body (Chapter 2). Only three human X-palindrome gene families are currently associated with phenotypes: Two broadly expressed spacer genes associated with non-reproductive Mendelian phenotypes (Bione et al. 1994, Fox et al. 1998, Clapham et al. 2012), and one testis-biased arm gene family, *SSX2*, involved in a somatic translocation that is a driver of synovial sarcoma (Clark et al. 1994). While knockouts of mouse X-amplicon genes will continue to fuel useful hypotheses about human X-palindrome gene functions, including hypotheses that they could act as modulators of spermatogenesis under stress conditions (Hou et al. 2016, Fon Tacer et al. 2019) or drivers of X versus Y genomic conflict (Cocquet et al. 2009, Cocquet et al. 2012, Kruger et al. 2019), the fact that most human X-palindrome genes lack orthologs in mice means that such studies cannot directly illuminate human gene functions (Mueller et al. 2013).

Several indirect lines of evidence suggest that deletions of human X-palindrome genes may have only mild phenotypes, including the lack of Mendelian disease associations (Mueller et al. 2013, Chapter 2), the high frequencies and lack of male fertility phenotypes for deletions of two testis-biased human X-palindrome gene families (Chapter 2), and relaxed purifying selection on conserved primate X-palindrome gene families (Chapter 2, discussed more below). Based on these findings, I suggest that future studies using quantitative trait associations will be required to elucidate the functions of human X-palindrome gene families. This task is made more challenging by the exclusion of the X chromosome from many GWAS publications to date (Wise et al. 2013), yet recent years have seen an increase in the availability of large datasets that link genotype to phenotypes, including the UK Biobank (Sudlow et al. 2015), the Million Veterans Program (Gaziano et al. 2016), and the All of Us Project (The All of Us Research Program Investigators 2019), which could enable new studies that incorporate the X chromosome. For gene families with testis-biased gene expression, future research could also use targeted approaches similar to the one we presented in Chapter 2, which seek enrichments for mutations or deletions in X-palindrome gene families in oligozoospermic or azoospermic men versus healthy controls.

Revisiting the relationship between amplicons and testis-biased gene families

Ampliconic gene families exhibit predominantly testis-biased expression in nearly all systems of mammalian sex-chromosome palindromes studied to date (Skaletsky et al. 2003, Mueller et al. 2008, Hughes et al. 2010, Hughes et al. 2012, Mueller et al. 2013, Soh et al. 2014, Hughes et al. 2020), yet the reason for this association is poorly understood. Previous work suggested that amplicons might promote the survival of testis-biased gene families, based on the finding that intact testis-biased gene families on the human Y chromosome are found within amplicons, while pseudogenes from the same gene families are scattered randomly among both ampliconic and single-copy sequence (Skaletsky et al. 2003). Our results from Chapter 2 suggest that the converse is also true: Protein-coding genes, including both testis-biased and broadly expressed genes, may promote the long-term survival of palindromes. Out of twelve palindromes conserved among human, chimpanzee, and rhesus macaque, all contain at least one protein-coding gene in their arms or spacer, and protein-coding genes are significantly better conserved than non-coding sequence (Chapter 2). Yet while our results provide an explanation for why protein-coding genes might be enriched within palindromes, they do not address the original question: Why should testis-biased gene families in particular benefit from this arrangement?

Based on results from this thesis as well as previous literature, I propose that this association results from weak purifying selection acting on testis-biased primate X-palindrome gene families. As noted above, several lines of evidence support the hypothesis that X-palindrome gene families evolve under weak purifying selection. First, while we found that dN/dS values for nearly all X-palindrome gene families were less than one, consistent with purifying selection, we also noted that these dN/dS values were higher than the genome-wide average, suggesting that purifying selection may be relatively weak (Chapter 2). Second, we did not detect any association between deletions that remove one copy of the testis-biased gene families *CXorf51* or *CXorf49* with azoospermia or oligozoospermia (Chapter 2), although we cannot rule out that such associations exist below our threshold of detection. This result is reminiscent of X-palindrome deletion phenotypes in mouse, where *Magea* deletions result in mild and context-dependent effects on male fertility (Hou et al. 2016, Fon Tacer et al. 2019). Finally, it is well-established that testis-biased genes in general tend to be younger and evolving under weaker purifying selection than older and more broadly expressed genes (Cai and Petrov 2010, Kryuchkova and Robinson-Rechavi 2015).

If testis-biased gene families on the primate X chromosome are evolving under weak purifying selection, then I propose that they in particular may benefit from the ongoing recombination within palindromes and other amplicons. Recombination increases the efficiency of both positive and negative selection (Felsenstein 1974), i.e. it tends to increase the probability that beneficial mutations will become fixed in a population and that harmful mutations will be eliminated, by expanding the pool of genetic variation upon which natural selection can act. In principle, this should benefit any gene. However, genes evolving under strong purifying selection will tend to survive with or without the benefits of recombination that result from palindrome formation; one striking example is the handful of broadly expressed, dosage sensitive, single-copy genes on the human Y chromosome that have survived without recombination for as much as 200-300 million years (Bellott et al. 2014). In contrast, an increase in the efficiency of natural selection resulting from palindrome formation may significantly increase the survival prospects for a gene under weak purifying selection. Indeed, simulations of Y-chromosome palindrome evolution found that ongoing gene conversion provided large evolutionary benefits under conditions of weak purifying selection, as measured by a dramatic increase in the number of 'mutation-free' Y chromosomes measured at the end of each simulation, compared to only modest benefits under conditions of strong purifying selection (Connallon and Clark 2010). This model predicts that weakly selected testis-biased gene families would be preferentially found within palindromes or other amplicons, consistent with observations reported here and elsewhere (Skaletsky et al. 2003, Hughes et al. 2010, Hughes et al. 2012, Soh et al. 2014, Hughes et al. 2020). Future tests of this model could include empirical work, such as examining the strength of natural selection and expression patterns among conserved palindrome genes across additional species, as well as new simulations of palindrome evolution that investigate the survival of X- and Y-palindrome genes under different strengths of purifying selection.

REFERENCES

The All of Us Research Program Investigators. 2019. The "All of Us" research program. *N Engl J Med* **381:** 668–676.

Berglund J, Pollard KS, Webster MT. 2009. Hotspots of biased nucleotide substitutions in human genes. *PLoS Biol* **71:** e1000026.

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**: 643–650.

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.

Cai JJ, Petrov DA. 2010. Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. *Genome Biol Evol* **2**: 393–409.

Clapham KR, Yu TW, Ganesh VS, Barry B, ChanY, Mei D, Parrini E, Funalot B, Dupuis L, Nezarati MM, et al. 2012. *FLNA* genomic rearrangements cause periventricular nodular heterotopia. *Neurology* **78**: 269–278.

Clark J, Rocques PJ, Crew AJ, Gill S, Shipley J, Chan AML, Gusterson BA, Cooper CS. 1994. Identification of novel genes, *SYT* and *SSX*, involved in the t(X;18)(p11.2;q11.2) translocation found in human synovial sarcoma. *Nat Genet* **7**: 502–508.

Cocquet J, Ellis PJI, Mahadevaiah SK, Affara NA, Vaiman D, Burgoyne PS. 2012. A genetic basis for a postmeiotic X versus Y chromosome intragenomic conflict in the mouse. *PLoS Genet* **8**: e1002900.

Cocquet J, Ellis PJI, Yamauchi Y, Mahadevaiah SK, Affara NA, Ward MA, Burgoyne PS. 2009. The multicopy gene *Sly* represses the sex chromosomes in the male mouse germline after meiosis. *PLoS Biol* **7**: e1000244.

Connallon T, Clark AG. 2010. Gene duplication, gene conversion and the evolution of the Y chromosome. *Genetics* **186:** 277–286.

Felsenstein J. 1974. The evolutionary advantage of recombination. *Genetics* **78**: 737–756.

Fon Tacer K, Montoya MC, Oatley MJ, Lord T, Oatley JM, Klein J, Ravichandran R, Tillman H, Kim MS, Connelly JP, et al. 2019. MAGE cancer-testis antigens protect the mammalian germline under environmental stress. *Sci Adv* **5:** eaav4832.

Fox JW, Lamperti ED, Ekşioğlu YZ, Hong SE, Feng Y, Graham DA, Scheffer IE, Dobyns WB, Hirsch BA, Radtke RA. 1998. Mutations in Filamin 1 prevent migration of cerebral cortical neurons in human periventricular heterotopia. *Neuron* **21**: 1315–1325.

Galtier N, Duret L, Glémin S, Ranwez V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet* **25**: 1–5.

Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, Breeling J, Whitbourne S, Deen J, Shannon C, Humphries D, et al. 2016. Million Veteran Program: a mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol* **70:** 214–223.

Gordon D, Huddleston J, Chaisson MJP, Hill CM, Kronenberg ZN, Munson KM, Malig M, Raja A, Fiddes I, Hillier LDW, et al. 2016. Long-read sequence assembly of the gorilla genome. *Science* **352**: aae0344.

Hou S, Xian L, Shi P, Li C, Lin Z, Gao X. 2016. The *Magea* gene cluster regulates male germ cell apoptosis without affecting the fertility in mice. *Sci Adv* **6:** 26735.

Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Graves T, Fulton RS, Dugan S, Ding Y, Buhay CJ, Kremitzki C, et al. 2012. Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* **483**: 82–86.

Hughes JF, Skaletsky H, Pyntikova T, Graves TA, van Daalen SKM, Minx PJ, Fulton RS, McGrath SD, Locke DP, Friedman C, et al. 2010. Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* **463**: 536–539.

Hughes JF, Skaletsky H, Pyntikova T, Koutseva N, Raudsepp T, Brown LG, Bellott DW, Cho T-J, Dugan-Rocha S, et al. 2020. Sequence analysis in *Bos taurus* reveals pervasiveness of X-Y arms races in mammalian lineages. *Genome Res* **30**: 1716–1726.

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, et al. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* **36**: 338–345.

Kruger AN, Brogley MA, Huizinga JL, Kidd JM, de Rooij DG, Hu Y-C, Mueller JL. 2019. A neofunctionalized X-linked ampliconic gene family is essential for male fertility and equal sex ratio in mice. *Curr Biol* **29**: 3699–3706.

Kryuchkova N, Robinson-Rechavi MA. 2015. Tissue-specific evolution of protein coding genes in human and mouse. *PLoS One* **10:** e0131673

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol Biol Evol* **34**: 1812–1819.

Lachance J, Tishkoff SA. 2014. Biased gene conversion skews allele frequencies in human populations, increasing the disease burden of recessive alleles. *Am J Hum Genet* **954**: 408–420.

Lakich D, Kazazian HH, Antonarakis SE, Gitschier J. 1993. Inversions disrupting the factor VIII gene are a common cause of several haemophilia A. *Nature Genet* **5**: 236–241.

Lange J, Skaletsky H, van Daalen SKM, Embry SL, Cindy M, Brown LG, Oates RD, Silber S, Repping S, Page DC. 2009. Isodicentric Y chromosomes and sex disorders as byproducts of homologous recombination that maintains palindromes. *Cell* **138**: 855–869.

Low WY, Tearle R, Bickhart DM, Rosen BD, Kingan SB, Swale T, Thibaud-Nissen F, Murphy TD, Young R, Lefevre L, et al. 2019. Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity. *Nat Commun* **10**: 1–11.

Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA, et al. 2020. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585:** 79–84.

Mueller JL, Mahadevaiah SK, Park PJ, Warburton PE, Page DC, Turner JMA. 2008. The mouse X chromosome is enriched for multicopy testis genes showing post-meiotic expression. *Nature Genet* **40**: 794–799.

Mueller JL, Skaletsky H, Brown LG, Zaghlul S, Rock S, Graves T, Auger K, Warren WC, Wilson RK, Page DC. 2013. Independent specialization of the human and mouse X chromosomes for the male germline. *Nat Genet* **45**: 1083–1087.

Necşulea A, Popa A, Cooper DN, Stenson PD, Mouchiroud D, Gautier C, Duret L. 2011. Meiotic recombination favors the spreading of deleterious mutations in human populations. *Hum Mutat* **322**:198–206.

Ohno S. 1967. *Sex chromosomes and sex-linked genes*. Springer, Berlin.

Ratnakumar A, Mousset S, Glémin S, Berglund J, Galtier N, Duret L, Webster MT. 2010. Detecting positive selection within genomes: the problem of biased gene conversion. *Philos Trans R Soc B Biol Sci* **365:** 2571–2580.

Rozen S, Skaletsky H, Marszalek 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.

Sahin U, Oehm P, Derhovanessian E, Jabulowsky RA, Vormehr M, Gold M, Maurus D, Schwarck-Kokarakis D, Kuhn AN, Omokoko T, et al. 2020. An RNA vaccine drives immunity in checkpoint-inhibitor-treated melanoma. *Nature* **585:** 107–112.

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.

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

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown 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.

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

Soh YQS, Alfoldi J, Pyntikova T, Brown LG, Minx PJ, Fulton RS, Kremitzki 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.

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 human X chromosome. *BMC Genomics* **8**:129.

Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, et al. 2015. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**: e1001779.

Warren WC, Harris RA, Haukness M, Fiddes 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.

Wise AL, Gyi L, Manolio TA. 2013. eXclusion: Toward integrating the X chromosome in genome-wide association analyses. *Am J Hum Genet* **92:** 643–647.

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.

APPENDIX:

SHIMS 3.0: Highly efficient single-haplotype iterative mapping and sequencing using ultra-long nanopore reads

Daniel W. Bellott, Ting-Jan Cho, Emily K. Jackson, Helen Skaletsky, Jennifer F. Hughes, David C. Page

Acknowledgements
This work was supported by the Howard Hughes Medical Institute, and generous gifts from Brit and Alexander d'Arbeloff and Arthur W. and Carol Tobin Brill

Author Contributions
D.W.B., H.S., J.F.H., and D.C.P. designed the study. D.W.B., T.-J.C., and E.K.J. developed the experimental methods. D.W.B. wrote the scripts for computational analysis. T.J.C. carried out the sequencing of the *TSPY* region. H.S. assembled the *TSPY* region. D.W.B., T.-J.C., and D.C.P. wrote the manuscript.

Adapted from: Bellott DW, Cho T-J, Jackson E, Skaletsky H, Hughes JF, Page DC. 2020. SHIMS 3.0: Highly efficient single-haplotype iterative mapping and sequencing using ultra-long nanopore reads. *bioRxiv* doi: https://doi.org/10.1101/2020.09.18.303735.

This manuscript was prepared for submission to *Nature Protocols*. Formatting and section headings, including detailed description of protocol steps, are in accordance with their guidelines.

ABSTRACT

The reference sequence of structurally complex regions can only be obtained through a highly accurate clone-based approach that we call Single-Haplotype Iterative Mapping and Sequencing (SHIMS). In recent years, improvements to SHIMS have reduced the cost and time required by two orders of magnitude, but internally repetitive clones still require extensive manual effort to transform draft assemblies into reference-quality finished sequences. Here we introduce SHIMS 3.0, using ultra-long nanopore reads to resolve internally repetitive structures and minimize the need for manual finishing of Illumina-based draft assemblies. This protocol proceeds from clone-picking to finished assemblies in 2 weeks for about 80 dollars per clone. We have used SHIMS 3.0 to finish the structurally complex *TSPY* array on the human Y chromosome, which could not be resolved by previous sequencing methods. Our protocol provides access to structurally complex regions that would otherwise be inaccessible from whole-genome shotgun data or require an impractical amount of manual effort to generate an accurate assembly.

INTRODUCTION

Background and Applications

Reference genome sequence quality is of central importance to modern biological research. Experiments based on aligning cheap and abundant short reads to existing reference sequences have become commonplace, permitting studies of variation by genome and exome resequencing, transcription by RNA sequencing, and epigenetic modifications by chromatin immunoprecipitation–sequencing. However, these experiments are limited by the quality and completeness of the underlying reference sequence, so that new insights may emerge from reanalyzing short-read datasets in the light of an improved reference sequence. The foremost obstacles to accurate reference genome assembly are repeated sequences within the genome. The most structurally complex repeats are ampliconic sequences – euchromatic repeats with greater than 99% identity over more than 10 kb (Mueller et al. 2013). The complex repetitive structures in amplicons mediate deletions, duplications, and inversions associated with human disease (Lupski 1998). Amplicons pose special challenges for genome assembly, requiring extremely long and accurate reads to discriminate between amplicon copies and produce a correct reference sequence.

We developed our Single Haplotype Iterative Mapping and Sequencing (SHIMS) approach to cope with the ampliconic sequences of the human Y chromosome (Kuroda-Kawaguchi et al. 2001). Because paralogous ampliconic repeats are more similar than alleles, we sequenced large-insert clones from a single haplotype, allowing us to confidently identify the rare sequence family variants (SFVs) that distinguish paralogous repeats in highly accurate (less than 1 error per megabase) synthetic long reads (Kuroda-Kawaguchi et al. 2001). Mapping and sequencing were coupled; newly sequenced clones provide novel SFVs that refine the clone map and serve as markers to select new clones. SHIMS has been instrumental to producing reference sequences of structurally complex sex chromosomes from several species (Skaletsky et al. 2003, Hughes et al. 2010, Bellott et al. 2010, Hughes et al. 2012, Mueller et al. 2013), as well as the human immunoglobulin gene cluster (Watson et al. 2013), and other structurally complex regions on human autosomes (Dennis et al. 2012). SHIMS remains the only sequencing approach that can reliably disentangle ampliconic repeats. Whole genome shotgun (WGS) strategies are constrained by a tradeoff between read length and accuracy among existing sequencing technologies. Sanger or Illumina reads are accurate, but are not long enough to traverse interspersed repeats, much less ampliconic sequence (She et al. 2004). Single-molecule sequencing technologies like PacBio or nanopore sequencing offer reads long enough to span interspersed repeats and smaller ampliconic sequences, but lack the accuracy to disentangle nearly identical ampliconic repeats (Gordon et al. 2016). As originally implemented, SHIMS 1.0 required the resources of a fully-staffed genome center to generate Sanger reads, assemble draft sequences, and manually ‘finish’ each clone. We developed SHIMS 2.0 to combine the advantages of a hierarchical clone-based strategy with high-throughput sequencing technologies, allowing a small team to generate sequence, while reducing time and cost by two orders of magnitude, while maintaining high accuracy (Bellott et al. 2018). However, SHIMS 2.0 still required intensive manual review to resolve internally repetitive clones, and in some cases – particularly short, nearly perfect, tandem repeats – complete resolution remained impossible.

Here we describe SHIMS 3.0, an extension of our SHIMS sequencing strategy that uses a combination of nanopore and Illumina sequencing technologies to resolve repetitive structures within individual large-insert clones. We describe a protocol for generating full-length nanopore reads for pools of clones, and combining the structural information from these full-length reads with highly accurate short-read data to automatically produce assemblies of internally repetitive clones (Fig. 1). This protocol proceeds from clone-picking to finished assemblies in 2 weeks for about 80 dollars per clone, an improvement of 2 orders of magnitude compared with 24 months and 4000 dollars under SHIMS 1.0. As a proof of principle, we apply SHIMS 3.0 to resolve the *TSPY* array on the human Y chromosome. The *TSPY* array is one of the largest and most homogeneous protein-coding tandem arrays in the human genome (Warburton et al. 2008), and it could not be completely resolved in the SHIMS 1.0 reference sequence of the human Y chromosome (Skaletsky et al. 2003).

Methodology

Large-insert clone libraries derived from a single haplotype are essential to the SHIMS strategy, and are discussed in detail in our description of SHIMS 2.0 (Bellott et al. 2018). In brief, any library derived from an individual of the heterogametic sex will provide a single haplotype source for sequencing sex chromosomes, albeit at half the coverage of the autosomes. Libraries created from inbred strains can provide a single-haplotype source for autosomes. When inbreeding is not possible, special measures may be necessary to obtain a single-haplotype source of DNA (Dennis et al. 2012). The ideal library will have greater than 10x coverage of the chromosome of interest to minimize the number of gaps in library coverage. In SHIMS 1.0 and 2.0 it was important to match the average library insert size to the expected amplicon unit size, such that it was rare for two units to be present in the same clone (Bellott et al. 2018). SHIMS 3.0 uses ultra-long nanopore reads to span entire BAC clones (Quick 2018); therefore, we now

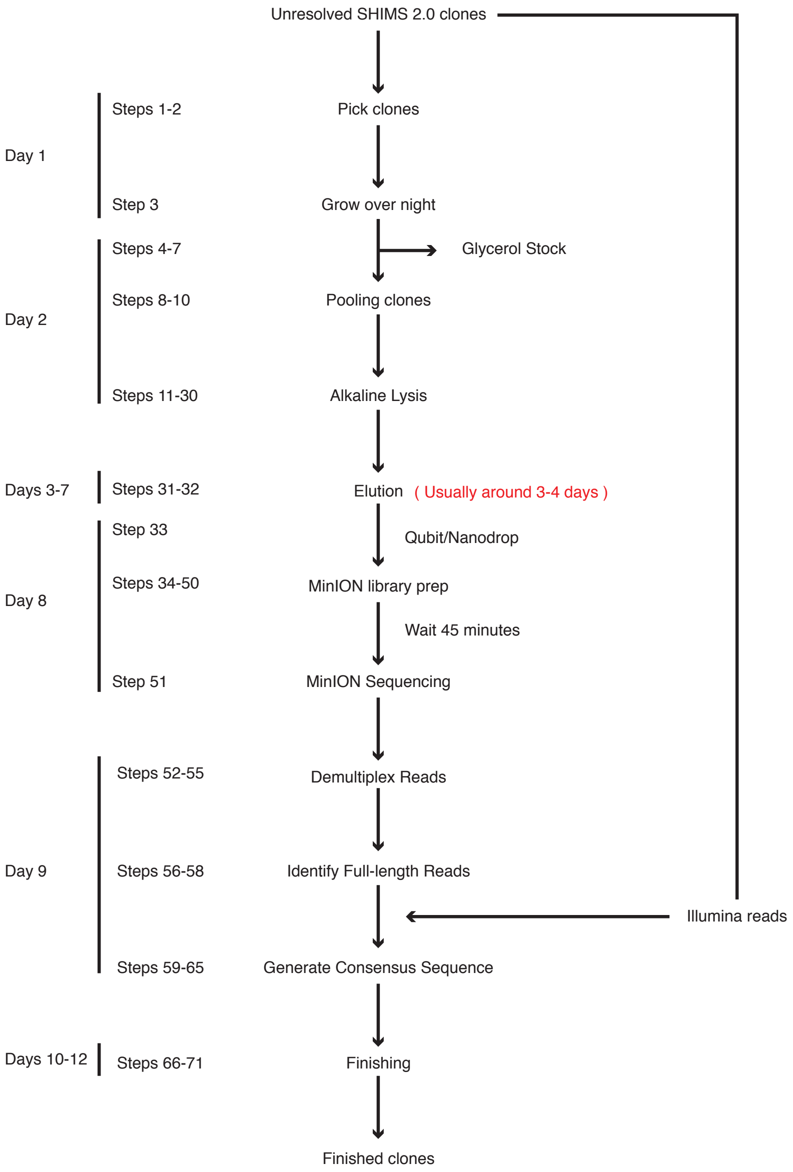


Figure 1. Overview of the SHIMS 3.0 protocol. A timeline of a single iteration of the SHIMS 3.0 protocol, showing the major protocol steps, with key quality controls on the right. During a two-week iteration, 24 clones are processed in parallel to rapidly generate finished sequence from structurally complex clones. A single technician can proceed from a list of clones to full-length nanopore libraries in 8 days. After a brief MinION run overnight, a bioinformatics specialist can demultiplex fastq sequences, identify full-length reads, then polish and edit the consensus of these reads to generate finished clone sequence.

recommend striving for the largest possible insert size, to minimize library construction, screening, and sequencing costs, as fewer clones will be necessary to achieve the required level of coverage.

All SHIMS strategies begin by selecting an initial tiling path of large-insert clones for sequencing and iterative refinement. Depending on the resources available for each library, it may be possible to identify clones of interest electronically, using fingerprint maps or end sequences, screening high-density filters by hybridization with labeled oligos, or high-dimensional pools for sequence-tagged-site content by PCR. It is most cost-effective to confirm the identity of each clone by generating draft sequence with the SHIMS 2.0 protocol, rather than designing specific assays for each clone (Bellott et al. 2018). In brief, this highly parallel method involves shearing BAC DNA to generate large (~1 kb) fragments for individually indexed Illumina TruSeq–compatible libraries to sequence and assemble pools of 192 clones in a single week (Bellott et al. 2018). In our experience, the structure of ampliconic regions is often unclear until a nearly complete tiling path is assembled, as the sequence map gradually unfolds as new variants are identified by sequencing. It is therefore preferable to seed the first iteration with as many clones as possible to identify sequence family variants early, and minimize the total number of iterations.

Draft clone assemblies generated from Sanger or Illumina reads are accurate enough to identify sequence family variants, and identify a minimum tiling path of clones. In previous iterations of SHIMS, each clone in this path would be painstakingly ‘finished’ to produce as correct and contiguous a sequence as possible. Highly skilled technicians would inspect draft assemblies for errors and anomalies, order and orient all draft sequence contigs, close all gaps, and resolve or annotate all sequence ambiguities (e.g. SSRs). SHIMS 3.0 departs from this approach, instead relying on the ability of nanopore-based sequencing technology to generate full-length reads to scaffold short-read assemblies and eliminate the need for laborious and time-consuming experiments, such as subcloning, PCR reactions, restriction digests, and transposon bombing, that were used to correct draft assemblies in the past.

We adapted existing methods for generating ultra-long reads (Quick 2018) for use with pools of large-insert clones on the Oxford Nanopore Technologies (ONT) MinION platform. Successful generation of full-length reads requires intact DNA of high concentration and purity. We optimized our protocol to avoid unnecessary manipulations that could damage DNA or introduce contamination. We culture 24 clones separately, and then pool all cultures for DNA isolation, library preparation, and sequencing. In contrast to conventional plasmids, BACs and fosmids are present in only a single copy per host cell, and common reagents for increasing the efficiency of DNA precipitation, such as glycogen or SPRI beads, are incompatible with nanopore sequencing. We compensate for this by starting with large volume (~15 ml) BAC and fosmid cultures to ensure that we harvest a sufficient amount of intact DNA. To preserve the integrity of high-molecular-weight (HMW) DNA, we handle it as little as possible, pipetting very slowly, using only wide-bore tips. We allow precipitated DNA to resuspend in water slowly over several days, rather than mixing by vortexing, or pipetting up-and-down. We have had the best results generating libraries from 7.5 to 15 µg of HMW DNA in 15 µl using the transposase-based RAD-004 library preparation kit from ONT. At these concentrations, solutions of HMW DNA will be extremely viscous, and it is difficult to measure the concentration precisely; some trial-and-error may be required to get the correct ratio of transposase to DNA, but we find that 0.5 µl of FRA for 15 ug is generally a good starting point to ensure that most BAC clones are cut only once. It is important to wait 45 minutes between loading the nanopore flow cell and starting the run, to allow time for full-length molecules to diffuse to the pores, otherwise the run will be dominated by shorter molecules. In contrast to our approach for generating Illumina reads, indexing or barcoding individual clones is not necessary, as full-length nanopore reads can be uniquely assigned to clones, even within the same amplicon.

Full-length nanopore reads transform clone finishing into a purely computational exercise. The tool chain for handling ultra-long nanopore reads is not yet fully mature, but it is developing rapidly. We rely on *Minimap2* for alignments involving full-length reads (Li 2018). This includes assigning nanopore reads to clones based on SHIMS 2.0 draft sequences, identifying full-length reads that start and end in vector sequence, and aligning a mix of long and short reads to generate a consensus. We use *Racon* for polishing the consensus sequence (Vaser et al. 2017), and *SAMtools* and custom scripts to manipulate read alignments (Li et al. 2009). We use *Gap5* and *Consed* for visualizing discrepant bases and manually editing the consensus (Bonfield and Whitwham 2010, Gordon and Green 2013). While full-length reads guarantee the correct overall sequence structure, a variety of alignment artifacts may occur in clones with highly identical internal repeats. In this case, we find it is best to electronically split the clone sequence into individual repeat units, and correct each unit separately, before merging them together to create a finished clone sequence.

Performance

The *TSPY* array on the Y chromosome is the largest and most homogeneous protein-coding tandem array in the human genome (Warburton et al. 2018), consisting of a 20.4-kb unit present in a highly variable number of copies, ranging from 11 to 72 per individual (Tyler-Smith et al. 1998). *TSPY* encodes a testis-specific protein, implicated in gonadoblastoma (Tsuchiya et al. 1995), that regulates cell proliferation (Oram et al. 2006); *TSPY* copy number is positively correlated with sperm count and sperm concentation (Giachini et al. 2009). As a demonstration of the expected performance of SHIMS 3.0, we fully resolved this array for the first time, using clones from the RP11 BAC library that we previously employed for our SHIMS 1.0 assembly of the male-specific region of the human Y chromosome (Skaletsky et al. 2003) (Fig. 2). This array spans 600 kb and contains 29 repeat units (Fig. 2a & b). We sequenced a redundant path of 19 clones and identified 94 sequence family variants that allowed us to select a non-redundant tiling path of 9 clones for finishing (Fig. 2c). On average, each unit differs from the others by 1 in 100 bases. The array encodes 14 distinct *TSPY* transcripts, encoding 10 different proteins, and it includes one pseudogene (**Fig. 2d**). We observed that at least 4 transcript variants were expressed in published testis RNA-seq datasets from other males (Lin et al. 2014), indicating that multiple copies are expressed.

Using only Sanger or Illumina reads, the presence of multiple amplicon copies within a single insert causes the clone assembly to collapse. The median BAC clone in the *TSPY* array contains 9 repeat units, making it impossible to accurately assemble even a single clone from this region using SHIMS 1.0 without months of manual finishing efforts at a genome center. By incorporating full-length nanopore reads, SHIMS 3.0 permits a small team – a technician and bioinformatics specialist – to finish these

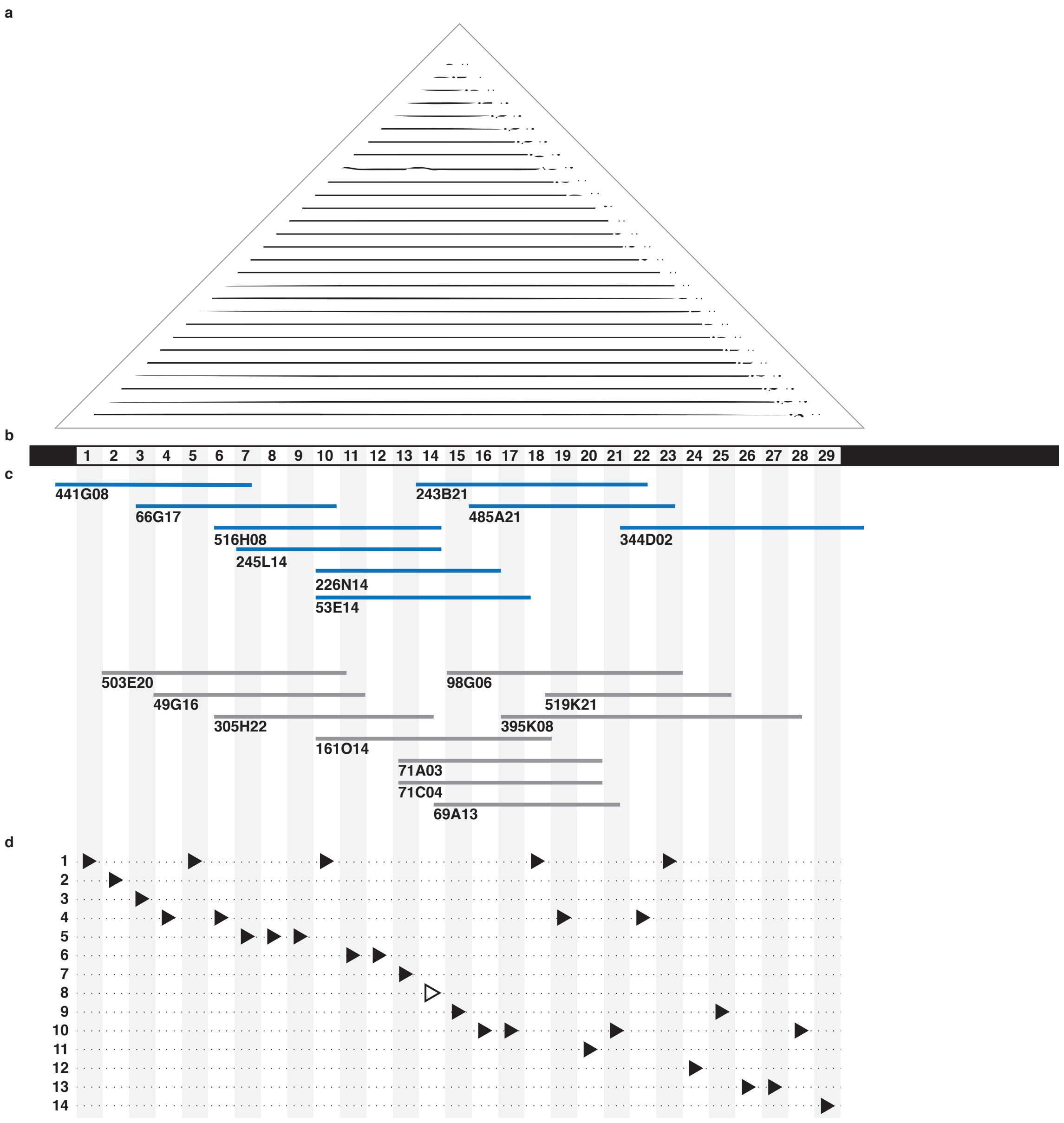


Figure 2. Structure of the Human *TSPY* array. a) Triangular dot plot of *TSPY* array region (AC279304) assembled using SHIMS3.0; each dot represents a 100 nucleotide perfect match between sequences within the array. b) Schematic representation of the 29 repeat units of the *TSPY* array. c) Clones from the RP11 BAC library sequenced with SHIMS 2.0 and SHIMS 3.0 (blue) to obtain finished sequence, or SHIMS 2.0 alone (grey) to identify sequence family variants used to map the array. d) There are 14 distinct *TSPY* transcripts (triangles), including 1 pseudogene (open triangle).

challenging sequencing targets. Our SHIMS 3.0 protocol can improve 24 SHIMS 2.0 draft sequences to finished quality in 2 weeks at a cost of $80 per clone. Despite the enormous reduction in staffing, cost, and time, the quality of finished sequence remains extremely high. We observe less than 1 error per megabase in overlaps between clones, on par with previous versions of SHIMS.

Comparison with other methods

SHIMS produces de novo sequence assemblies with higher accuracy than any other technique, making it possible to produce accurate reference sequence of the most extreme repetitive regions, from ampliconic sequences like the nearly-perfect multi-megabase duplications on the mouse Y chromosome (Soh et al. 2014), to the thousands of centromeric satellite repeats that form the centromere of the human Y chromosome (Jain et al. 2018a). This extremely high accuracy is due to the clone-based nature of SHIMS. Each clone represents a single long molecule that can be sequenced repeatedly, with complimentary technologies, to generate an assembly that is accurate at the level of overall structure as well as the identity of individual nucleotides. This property makes it possible to identify and repeatedly verify the rare sequence family variants that distinguish ampliconic repeats, and build a high-confidence map from individual clones.

The impressive advances in single-molecule sequencing technologies that enabled SHIMS 3.0 have also increased the capabilities of whole genome shotgun approaches (Jain et al. 2018b). It is now routine to generate nanopore sequencing runs where half of the bases are in reads longer than 100 kb, so that interspersed repeats and smaller ampliconic structures can be spanned by a single long read. Whole genome shotgun with nanopore reads enabled the complete assembly of the human X chromosome from a single haplotype source, the CHM13hTERT cell line (Miga et al. 2020). This effort required deep coverage from nanopore reads as well as from a broad array of complementary sequencing and mapping technologies, combined with manual review of structurally complex regions (Miga et al. 2020). Error rates were still orders of magnitude higher than clone based strategies – 1 error in 10 kb in single-copy sequence, and 7 errors per kb in sequences present in more than one copy (Miga et al. 2020). This elevated error rate in multi-copy sequence is due to the inherent difficulties of uniquely mapping short reads to long paralogous repeats; instead of reconstructing the true sequence, error correction with short reads blurs all paralogs together into an erroneous consensus. A second, orthogonal quality control measure indicates that shotgun sequencing with nanopore reads still lags behind clone-based approaches; 18% of CHM13 BAC sequences from segmental duplications and other difficult-to-assemble regions were missing from the whole genome shotgun assembly of CHM13hTERT (Miga et al. 2020). While sequencing technologies continue to improve in read length and accuracy, clone-based approaches will continue to be relevant for generating highly accurate reference sequence, particularly in otherwise inaccessible ampliconic regions.

Relative to previous versions of SHIMS, SHIMS 3.0 greatly reduces the resources required to successfully generate finished sequence. Under SHIMS 1.0, the cost to produce draft sequence averaged about $5000 per clone, while finishing averaged around $4000. Weeks of bench experiments and many days of expert review were required to transform each low coverage (5-8x) Sanger draft sequence with frequent gaps into a complete and contiguous assembly. In SHIMS 2.0, we reduced the need for costly finishing activities by opting for much higher coverage (50-80x) in much cheaper Illumina reads. We encountered fewer coverage gaps at this higher depth, and also fewer library gaps because of differences in the library preparation protocol. We relied on sonication to provide random shearing, and amplify library fragments by PCR, as opposed to cloning fragments in *E. coli*. Although this deep and relatively even coverage ensured that wet-bench experiments were rarely required for finishing, structurally complex clones still required several days of expert review using an assembly editor like Consed. In the most complex cases, involving many paralogous repeats within a single clone, such as the *TSPY* array or centromeric satellite repeats, it was still impossible to completely resolve the correct structure.

By incorporating full-length nanopore reads from each clone, SHIMS 3.0 now makes it possible to assemble even the most internally repetitive clones. Full-length nanopore reads provide complete certainty about the overall clone structure; there is no doubt about the order and orientation of sequences, and no question about the copy number of complex repeats. This limits finishing activity to the simple matter of resolving the few remaining discrepancies between nanopore and Illumina reads. In most clones, this requires less than an hour of effort for even inexperienced finishers, and results in highly accurate sequences, with less than 1 error per megabase. Highly repetitive clones require more attention, but they can be resolved by a simple divide-and-conquer strategy, where each paralogous repeat is finished separately, with special attention to SFV sites, and then merged to create the full finished sequence. Correctly mapping short reads to repeated sequences becomes more difficult as the number of paralogs increases, increasing the chances that each paralogous repeat unit is blurred toward the consensus. In contrast to WGS strategies, in SHIMS 3.0, this blurring is confined to the boundaries of a single clone, and comparisons with neighboring clones can be used to resolve the position of paralogous SFVs. SHIMS 3.0 dramatically decreases the time, cost, and effort required to obtain finished sequence; using an optimized protocol for preparing HMW DNA in parallel from pools of BAC clones, a small team can finish 24 Illumina draft assemblies in 2 weeks for $80 per clone.

Limitations of SHIMS 3.0

SHIMS 3.0 exceeds the capabilities of previous iterations of the SHIMS technique, providing access to the longest, most highly identical ampliconic sequences, as well as arrays of repeated sequences shorter than a single clone. However, SHIMS 3.0 shares two of the same limitations as previous versions of SHIMS and other clone-based approaches. First, the maximum size of BAC inserts limits SHIMS to resolving duplications with <99.999% identity. This limitation will remain until long-read technologies are able to surpass BAC sequencing in both read length and accuracy, or a reliable cloning technology emerges that exceeds the insert size of BACs. Second, SHIMS is limited to sequences that can be cloned into *E. coli*. Sequences that are toxic to *E. coli* are underrepresented in BAC and fosmid libraries. These library gaps can be resolved by directed efforts that avoid cloning in *E. coli*, like sequencing long-range PCR products (Skaletsky et al. 2003), or using the emerging selective sequencing (“ReadUntil”) capability of nanopore-based sequencers to enrich for reads flanking the gap.

Practitioners of SHIMS 3.0 also face new challenges due to their reliance on nanopore reads spanning 100-300 kb BAC inserts. Bioinformatics tools for aligning, visualizing, and editing reads of this length are not fully mature. SAM and BAM files both encode alignment details in the CIGAR format, however, the BAM format is limited to 65535 operations, which is frequently too few to encode the many transitions between matches, mismatches, insertions and deletions encountered in alignments of ultra-long nanopore reads (Li 2020). Moreover, Consed does not reliably display alignments of reads longer than 1 kb. Our workaround has been to split SAM formatted alignments of nanopore reads into uniquely named sub-alignments every 1000 match operations, and convert the resulting SAM files to BAM format, which is accepted by Consed. This permits full visualization of nanopore reads alongside Illumina reads during finishing.

Expertise

As with our previous protocol for SHIMS 2.0, we have designed the SHIMS 3.0 protocol to be carried out by a small team. A single technician can process 24 BAC clones from frozen stocks to nanopore sequencing libraries in 5 days with common molecular biology lab equipment. A bioinformatics specialist can set up a pipeline to identify full-length reads for each clone, generate a consensus sequence, automatically correct most errors using alignments with short reads, and manually review the resulting assembly for errors and identify SFVs. It is important to keep abreast of new developments in software for processing nanopore data, as all aspects from base-calling to alignment and error correction are continuously being improved.

MATERIALS AND METHODS

MATERIALS

- Fisherbrand Low-Retention Microcentrifuge Tubes (Fisher Scientific, cat. no. 02-681-320)
- ART Barrier Specialty Pipette Tips, 1000, wide bore (Fisher Scientific, cat. no. 2069GPK)
- 50 mL Falcon Tube (Fisher Scientific, cat. no. 14-959-49A)
- Nalgene™ PPCO Centrifuge Bottles with Sealing Closure (Fisher Scientific, cat. no 3141-0250)
- Costar Assay Plate 96-well (Corning, cat. no. 3797)

REAGENTS

- ZymoPURE II Plasmid Maxiprep Kit (Zymo Research, cat. no. D4203)
- Rapid Sequencing Kit (Oxford Nanopore Technologies, cat. no. SQK-RAD004)

EQUIPMENT

- EZ-Vac Vacuum Manifold (Zymo Research, cat. no. S7000)
- MinION (Oxford Nanopore Technologies)
- NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific, cat. no. ND 1000)
- Centrifuge 5810 R (Eppendorf, cat. no. 00267023)
- Microcentrifuge 5425 (Eppendorf, cat. no. 2231010059)
- Vortex-Genie 2 (Scientific Industries, cat. no. SI-0236)
- Portable Pipet-Aid XP2 Pipette (Drummond, cat. no. 4-000-501)
- Eppendorf ThermoMixer (Eppendorf, cat. no. 5380000028)
- Beckman Coulter Avanti J-E centrifuge (Beckman Coulter, cat, no. A20698)

SOFTWARE

- minimap2 (https://github.com/lh3/minimap2)

- racon (https://github.com/lbcb-sci/racon)
- samtools (http://www.htslib.org)
- tg_index (http://staden.sourceforge.net)
- gap5 (http://staden.sourceforge.net)
- (optional) consed (http://www.phrap.org/consed/consed.html)

REAGENT SETUP

70% (vol/vol) Ethanol Mix 30 ml of 100% (vol/vol) ethanol with 70 ml of ddH$_2$O. ▲**CRITICAL** 70% (vol/vol) ethanol should be prepared on the day of the experiment.

1M Tris-Cl, pH 8.5 Dissolve 121 g of Tris base in 800 ml of ddH$_2$O. Adjust pH to 8.5 with concentrated HCl, then adjust volume with ddH$_2$O to 1 L. 1M Tris-Cl can be prepared in advance and stored at room temperature for up to a year.

10 mM Tris-Cl, pH 8.5 Mix 0.5 ml of 1M Tris-Cl with 49.5 ml of ddH$_2$O. This solution can be prepared in advance and stored at room temperature for up to a year.

18% PEG/NaCl, 18% PEG/1M NaCl Solution Add 135 g of PEG-8000 powder into 1 L bottle. Add 150 ml of 5M NaCl, 7.5 ml of Tris-HCl, 1.5 ml of 0.5M EDTA and 450 ml of ddH$_2$O to make PEG-buffer.

PROCEDURE

Pick Clones and Grow Cultures ●TIMING 18 h

1| Fill each well of a Nunc 96 DeepWell plate with 1.9 ml of 2X LB containing 34 µg/ml chloramphenicol. ▲CRITICAL STEP Rich media (2X LB) is appropriate for single-copy plasmids like BACs or fosmids, which use chloramphenicol resistance as a selectable marker.

2| Use a clean pipette tip to scrape the surface of a frozen glycerol stock and drop the tip directly into the DeepWell plate to inoculate a well. Inoculate each sample 8 times for a total of 15.2 mL/sample. 24 samples in total for each library prep.

3| Seal plates with AirPore Tape Sheets and place at 37 °C for 16-17 h, shaking at 220 RPM. ▲CRITICAL STEP Overgrowth of cultures (cell density > 3-4 x 10^9 cells per ml) will decrease yield of BAC DNA.

Glycerol Stock Plate ●TIMING 30 min

4| Dispense 150 µl of 80% (vol/vol) glycerol solution into two rows of a Costar Assay Plate.

5| Transfer 150 µl of each sample culture from Step 3 to a corresponding well of the assay plate and mix by pipetting up and down 20 times.

6| Seal the glycerol stock plate with aluminum adhesive foil.

7| Store the glycerol stock plate at -80 °C.

Pooling Clones ●TIMING 1-2h

8| Pour overnight cultures from Step 3 into a large beaker to combine pool.

9| Divide pooled culture into two 250 mL Nalgene bottles and spin down culture at 6000 x g for 30 minutes at 4 °C.

10| Remove media by pouring into a waste-collecting container. Be careful not to disturb the pellets.

■ PAUSE POINT Store at -20 °C for up to a week

Alkaline Lysis ●TIMING 1-2 h

11| Add 7 mL of ZymoPURE P1 (Red) to each bacterial cell pellet and resuspend completely by pipetting. Combine into one bottle when cells are completely resuspended.

12| Add 14 mL of ZymoPURE P2 (Green) and immediately mix by gently inverting the tube 6 times. ▲CRITICAL STEP Do not vortex! Let sit at room temperature for 3 minutes. Cells are completely lysed when the solution appears clear, purple, and viscous.

13| Add 14 mL of ZymoPURE P3 (Yellow) and mix gently but thoroughly by inversion. ▲CRITICAL STEP Do not vortex! The sample will turn yellow when the neutralization is complete, and a yellowish precipitate will form.

14| Ensure the plug is attached to the Luer Lock at the bottom of the ZymoPURE Syringe Filter. Place the syringe filter upright in a tube rack and load the lysate into the ZymoPURE Syringe Filter and wait 8 minutes for the precipitate to float to the top.

15| Remove the Luer Lock plug from the bottom of the syringe and place it into a clean 50 mL conical tube. Place the plunger in the syringe and push the solution through the ZymoPURE Syringe Filter in one continuous motion until approximately 33-35 mL of cleared lysate is recovered. Save the cleared lysate!

16| Add 14ml ZymoPURE Binding Buffer to the cleared lysate from step 5 and mix thoroughly by inverting the capped tube 10 times.

17| Ensure the connections of the Zymo-Spin V-P Column Assembly are finger-tight and place onto a vacuum manifold.

18| With the vacuum off, add the entire mixture from step 6 into the Zymo-Spin V-P Column Assembly, and then turn on the vacuum until all the liquid has passed completely through the column.

19| Remove and discard the 50 mL reservoir from the top of the Zymo-Spin V-P Column Assembly.

20| With the vacuum off, add 5 mL of ZymoPURE Wash 1 to the 15 mL Conical Reservoir. Turn on the vacuum until all the liquid has passed completely through the column.

21| With the vacuum off, add 5 mL of ZymoPURE Wash 2 to the 15ml Conical Reservoir. Turn on the vacuum until all the liquid has passed completely through the column. Repeat this wash step.

22| Remove and discard the 15 mL Conical Reservoir and place the Zymo-Spin V-P Column in a Collection Tube. Centrifuge at ≥10,000 x g for 1 minute, in a microcentrifuge, to remove any residual wash buffer.

23| Transfer the column into a clean 1.5ml microcentrifuge tube and add 450 µl of 10 mM Tris-Cl (pre-warm at 50 °C) directly to the column matrix. Wait 10 minutes, and then centrifuge at ≥ 10,000 x g for 1 minute in a microcentrifuge.

24| Add 450 µl of 18% PEG/NaCl to the tube containing sample. Mix by flicking and rotating the 1.5 mL Eppendorf tube.

25| Centrifuge at ≥10,000 x g for 30 minutes at 4 °C, in a microcentrifuge.

26| Remove supernatant from the tube without disturbing the pellet.

27| Add 1 mL of 70% EtOH and spin for 10 minutes at 4 °C.

28| Repeat step 27 and 28.

29| Remove supernatant and any left over 70% EtOH from Eppendorf tube.

30| Air dry for 10 minutes or until no visible liquid is left in the tube. ▲**CRITICAL STEP** Do not over dry the pellet.

31| Dissolve DNA pellet in 18 µl 10 mM Tris-Cl.

32| Store DNA at 4 °C for several days until pellet completely dissolves into solution.

33| Check DNA concentration and quality with Qubit or NanoDrop.

MinION Library prep ●TIMING 30 Minutes

34| Adjust sample concentration from step 33 to 1 µg/µl with 10 mM Tris-Cl.

35| Using a wide bore pipet tip, slowly aspirate 15 µl into a low retention Eppendorf tube.

36| In a separate Eppendorf tube, add 0.5 µl FRA to 4.5 µl 10 mM Tris-Cl. Flick the tube to mix well.

37| Add the diluted FRA solution from step 36 into sample tube from step 35.

38| Gently flick the tubes a few times to mix.

39| Incubate sample on 30 °C heat block for 35 seconds, then move the tube to 80 °C heat block. Incubate for 2 minutes at 80C.

40| Remove the tube from heat block and incubate on ice for 1 minute, then move the tube off the ice. Equilibrate to room temperature, about 1 minute.

41| While the sample is equilibrating to room temperature, add 4.5 µl 10 mM Tris-Cl to 0.5 µl of RAP. Flick to mix well.

42| Add RAP dilution from step 41 into sample tube. Slowly flick the tube a few times to mix. Keep the sample at room temperature before loading.

MinION Library loading ●TIMING 30 Minutes

43| Add 30 µl of FLT to tube of FLB, vortex to mix the solution, follow by a quick spin.

44| Perform QC on a new MinION flow cell to check available pores and ensure that enough pores are present

45| Use a P1000 pipet to remove about 20-30 µl of storage buffer from priming pore. Load 800 µl flush buffer via the pore slowly. Wait 5 minutes

46| Lift SpotON cover and load 200 µl flush buffer slowly. Try to dispense at a speed where each bead of liquid is siphoned into the SpotON port as soon as it is visible.

47| Add 34 µl SQB and 15 µl DEPC Water to sample tube from step 42.

48| Flick the tube gently to mix, follow by a quick spin down to collect library to the bottom of the tube.

49| Slowly aspirate 75 µl of library with a wide bore tip. Very slowly, load the library into SpotON pore drop by drop.

50| Close both priming pores and put the SpotON cover back onto the pore.

51| After loading the library, leave the flow cell on bench for 45 minutes before starting the run.

Demultiplex Reads ●TIMING 30 Minutes

52| Prepare file of draft clone sequences in fasta format: *draft_clones.fa* ▲**CRITICAL STEP** When concatenating draft sequence assemblies, ensure that each sequence has a unique name

53| Prepare file of vector sequence in fasta format: *vector.fa*

54| Download fastq formatted reads from the device running MinION control software: *nanopore.fq*

55| Align nanopore reads to file of draft sequences to assign nanopore reads to clones by best match:

```
minimap2 -x map-ont draft_clones.fa nanopore.fq |  sort -r -n -k 10 | awk '!seen[$1]++' >
best_clone_match.paf
```

```
grep clone_name best_clone_match.paf | cut -f 1 >clone_name.txt
```

```
grep -A 3 -f clone_name.txt nanopore.fq | grep -v "^--$" > clone_name.nanopore.fq
```

Identify Full-length Reads ●TIMING 30 Minutes

▲CRITICAL We have automated Steps **56-65** with a custom Perl script (available at https://github.com/dwbellott/shims3_assembly_pipeline/), but the workflow is described below to allow for direct use of the individual software tools or substitution of alternative tools.

56| For each clone, align nanopore reads to file of vector sequence:

```
minimap2 -x map-ont vector.fa clone_name.nanopore.fq -o clone_name.vector.paf
```

57| Search for reads that begin and end with high-quality matches to vector sequence on the same strand – these are full-length reads

```
cut -f 1,5,6 clone_name.vector.paf | sort | uniq -c | sed ‘s/^//’ | grep “^2” | cut -f 2 -d ’ ’
>clone_name.2x.txt
```

```
awk ‘$2 - $3 < $7 && $12 == 60’ clone_name.vector.paf | cut -f 1,5,6 | grep -f clone_name.2x.txt
>clone_name.2x.right.txt
```

```
awk ‘$4 < $7 && $12 == 60’ clone_name.vector.paf | cut -f 1,5,6 | grep -f clone_name.2x.right.txt | cut -f
1 | sort | uniq >clone_name.fl.txt
```

58| For each clone, generate a fastq file of full-length reads, as well as a fasta file of the longest full-length read to use as a scaffold for final assembly.

```
grep -A 3 -f fl.txt clone_name.nanopore.fq | grep -v "^--$" > clone_name.fl.fq
```

```
grep -A 1 `head -n 1 clone_name.fl.txt` nanopore.fq | sed 's/^\@.*/\>clone_name/'
>clone_name.longest.fl.fa
```

Generate Consensus Sequence ●TIMING 30 Minutes

59| Polish the longest read twice, using the other full-length nanopore reads

```
minimap2 -x map-ont clone_name.longest.fl.fa clone_name.fl.fq > clone_name.longest.fl.paf
```

```
racon clone_name.fl.fq clone_name.longest.fl.paf clone_name.longest.fl.fa >
clone_name.longest.fl.racon.fa
```

```
minimap2 -x map-ont clone_name.longest.fl.racon.1.fa clone_name.fl.fq >
clone_name.longest.fl.racon.paf
```

```
racon clone_name.fl.fq clone_name.longest.fl.racon.1.paf clone_name.longest.fl.racon.1.fa
>clone_name.fl.consensus.fa
```

60| Gather up Illumina, nanopore, and (if available) PacBio reads for each clone.

```
cat clone_name.illumina.forward.fq clone_name.illumina.reverse.fq clone_name.illumina.single.fq
clone_name.nanopore.fq clone_name.pacbio.fq >> clone_name.allreads.fq
```

61| Polish the nanopore consensus sequence, using both long and short reads.

```
minimap2 -x asm20 clone_name.nanopore.consensus.fa clone_name.illumina.single.fq >>
clone_name.polish.1.paf
```

```
minimap2 -x sr clone_name.nanopore.consensus.fa clone_name.illumina.forward.fq
clone_name.illumina.reverse.fq >> clone_name.polish.1.paf
```

```
minimap2 -x map-ont clone_name.nanopore.consensus.fa clone_name.nanopore.fq >>
clone_name.polish.1.paf
```

```
minimap2 -x map-pb clone_name.nanopore.consensus.fa clone_name.pacbio.fq >>
clone_name.polish.1.paf
```

```
racon clone_name.allreads.fq clone_name.polish.1.paf clone_name.nanopore.consensus.fa
>clone_name.polish.1.fa
```

62| Repeat Step 61 four more times, for a total of 5 rounds of polishing

63| Align reads one last time to generate SAM format alignments suitable for assembly editors. ▲CRITICAL The BAM file format cannot accommodate CIGAR strings with greater than 65535 operations. Alignments involving nanopore reads spanning the full length of a BAC clone will exceed this limit. We strongly recommend storing alignments in SAM or CRAM format to avoid the loss of detailed alignment information.

```
minimap2 -x asm20 -a -L --sam-hit-only -R '@RG\tID:S\tSM:S\tPL:ILLUMINA' clone_name.polish.5.fa
clone_name.illumina.single.fq | samtools sort -O SAM - >clone_name.single.sorted.sam
```

```
minimap2 -x sr -a -L --sam-hit-only -R '@RG\tID:FR\tSM:FR\tPL:ILLUMINA' clone_name.polish.5.fa
clone_name.illumina.forward.fq clone_name.illumina.reverse.fq | samtools sort -O SAM -
>clone_name.paired.sorted.sam

minimap2 -x map-pb -a -L --sam-hit-only -R  '@RG\tID:P\tSM:P\tPL:PACBIO' clone_name.polish.5.fa
clone_name.pacbio.fq | samtools sort -O SAM - >clone_name.pacbio.sorted.sam

minimap2 -x map-ont -a -L --sam-hit-only -R '@RG\tID:N\tSM:N\tPL:PACBIO' clone_name.polish.5.fa
clone_name.nanopore.fq | samtools sort -O SAM -  >clone_name.nanopore.sorted.sam
```

64| *Combine alignments*

```
samtools merge -f clone_name.allreads.sorted.sam clone_name.single.sorted.sam
clone_name.paired.sorted.sam clone_name.pacbio.sorted.sam clone_name.nanopore.sorted.sam
```

65| Generate database for Gap5 ▲CRITICAL We now recommend Gap5 over Consed, because Gap5 natively supports loading data directly from SAM files and displaying full-length nanopore reads. It is possible to split SAM alignments of full-length nanopore reads into smaller fragments that can be encoded in a BAM file and displayed by Consed without loss of information. For those who wish to use Consed, we implement this work-around in a custom Perl script (available at https://github.com/dwbellott/shims3_assembly_pipeline/).

```
tg_index -o clone_name.g5d -p -9 -s clone_name.allreads.sorted.sam
```

Finishing ●TIMING 0-8 h

66| Open the assembly in Gap5:

```
gap5 clone_name.g5d
```

67| Select ‘Edit Contig’ from the ‘Edit’ menu.

68| Resolve discrepancies between Illumina reads and full-length nanopore reads (**Fig. 3**).

▲CRITICAL In Gap5, it is not possible to directly edit the consensus sequence. Instead, indicate which readings are authoritative by marking bases as high quality with the ‘]’ key, and the consensus will automatically update.

▲CRITICAL We usually resolve the consensus in favor of the Illumina reads. The vast majority of discrepancies between these technologies occur at homopolymer repeats, where nanopore reads are especially prone to insertion and deletion errors (**Fig. 3a**). More rarely, we encounter systematic errors in nanopore base calling that generate a consensus that is not supported by any Illumina read.

▲CRITICAL We resolve disagreements among Illumina reads in favor of the consensus of full-length nanopore reads. In clones that contain duplicated sequences, short Illumina reads can be mapped to the wrong repeat unit, but full-length nanopore reads are not subject to this artifact, and will usually have the correct base at each SFV.

69| Resolve SSRs by realigning reads around the SSR region. Select reads by clicking on their names on the left hand side of the edit window, and choose 'Realign Selection' from the 'Command' menu. ▲CRITICAL STEP Stutter noise from replication slippage in SSRs causes divergent reads and low-quality base calls. In some cases, unambiguous resolution of these repeats may not be possible, and they should be annotated as unresolved in Step 71.

70| Remove any vector-sequence contamination at the ends of the clone. In the Gap5 edit window, use the 'Search' button to search the consensus sequence for the sequences at the cloning site of your vector. Trim away the vector sequence outside of the restriction sites used to generate your clone library (usually EcoRI, BamHI, or MboI).

71| Annotate any remaining ambiguities in the clone sequence (e.g., unresolved simple sequence repeats, where neither Illumina or nanopore reads are completely accurate) by compiling a feature table (National Center for Biotechnology Information 2017), which will be useful when finished clone sequences are submitted to GenBank.

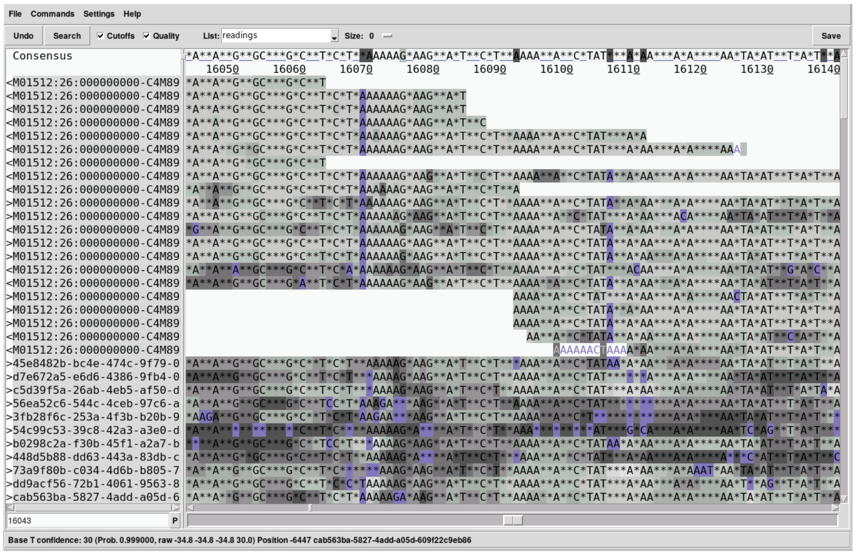
●TIMING

Steps 1-3, pick clones and grow cultures: 18 h
Steps 4-7, glycerol stock plate: 30 min
Steps 8-10, pooling clones: 1-2 h
Steps 11-33, alkaline lysis: 1-2 h
Steps 34-42, MinION library prep: 30 min
Steps 43-51, MinION library loading: 30 min
Steps 52-55, demultiplex reads: 30 min
Steps 56-58, Identify full-length reads: 30 min
Steps 59-65, generate consensus sequence: 30 min
Steps 66-71, finishing: 0-8 h

TROUBLESHOOTING

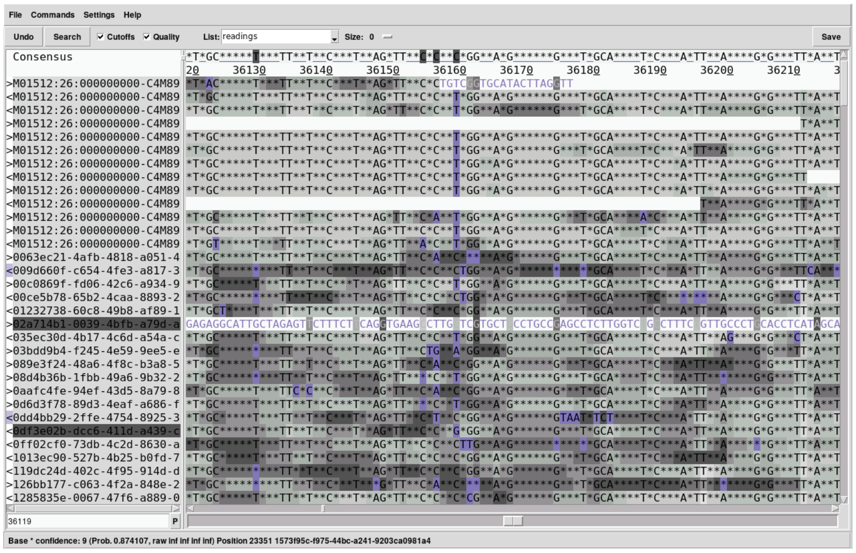
Troubleshooting advice can be found in **Table 1**.

a



```
 Consensus                 *A**A**G**GC***G*C**T*C*T**AAAAAG*AAG**A*T**C*T**AAAA**A**C*TAT***A*AA***A*A****AA*TA*AT**T*A*T**A
                              16050     16060     16070     16080     16090     16100     16110     16120     16130     16140
<M01512:26:000000000-C4M89 *A**A**G**GC***G*C**T
<M01512:26:000000000-C4M89 *A**A**G**GC***G*C**T*C*T*AAAAAAG*AAG**A*T
<M01512:26:000000000-C4M89 *A**A**G**GC***G*C**T*C*T*AAAAAAG*AAG**A*T
<M01512:26:000000000-C4M89 *A**A**G**GC***G*C**T*C*T*AAAAAAG*AAG**A*T**C
<M01512:26:000000000-C4M89 *A**A**G**GC***G*C**T*C*T*AAAAAAG*AAG**A*T**C*T**AAAA**A**C*TAT***A*A
<M01512:26:000000000-C4M89 *A**A**G**GC***G*C**T*C*T*AAAAAAG*AAG**A*T**C*T**AAAA**A**C*TAT***A*AA***A*A****AAA
<M01512:26:000000000-C4M89 *A**A**G**GC***G*C**T
<M01512:26:000000000-C4M89 *A**A**G**GC***G*C**T*C*T*AAAAAAG*AAG**A*T**C*T**AAAA**A**C*TATA**A*AA***A*A****AA*TA*AT**T*A*T**A
<M01512:26:000000000-C4M89 *A**A**G**GC***G*C**T*C*T*AAAAAAG*AAG**A*T**C*T**A
>M01512:26:000000000-C4M89 *A**A**G**GC***G*C**T*C*T*AAAAAAG*AAG**A*T**C*T**AAAA**A**C*TATA**A*AA***A*A****AA*TA*AT**T*A*T**A
>M01512:26:000000000-C4M89 *A**A**G**GC***G*C**T*C*T*AAAAAAG*AAG**A*T**C*T**AAAA**A**C*TAT***A*AA***ACA****AA*TA*AT**T*A*T**A
<M01512:26:000000000-C4M89 *G**A**G**GC***G*C**T*C*T*AAAAAAG*AAG**A*T**C*T**AAAA**A**C*TATA**A*AA***A*A****AA*TA*AT**T*A*T**A
>M01512:26:000000000-C4M89 *A**A**G**GC***G*C**T*C*T*AAAAAAG*AAG**A*T**C*T**AAAA**A**C*TATA**A*AA***A*A****AA*TA*AT**T*A*T**A
>M01512:26:000000000-C4M89 *A**A**G**GC***G*C**T*C*T*AAAAAAG*AAG**A*T**C*T**AAAA**A**C*TATA**A*AA***A*A****AA*TA*AT**T*A*T**A
<M01512:26:000000000-C4M89 *A**A**A**GC***G*C**T*C*A*AAAAAAG*AAG**A*T**C*T**AAAA**A**C*TAT***ACAA***A*A****AA*TA*AT**G*A*C**A
<M01512:26:000000000-C4M89 *A**A**G**GC***G*A**T*C*T*AAAAAAG*AAG**A*T**C*T**AAAA**A**C*TATA**A*AA***A*A****AA*TA*AT**T*A*T**A
>M01512:26:000000000-C4M89                                                  AAAA**A**C*TAT***A*AA***A*A****AACTA*AT**T*A*T**A
>M01512:26:000000000-C4M89                                                  AAAA**A**C*TATA**A*AA***A*A****AA*TA*AT**T*A*T**A
>M01512:26:000000000-C4M89                                                  AAAA**A**C*TATA**A*AA***A*A****AA*TA*AT**T*A*T**A
<M01512:26:000000000-C4M89                                                    AA**A**C*TATA**A*AA***A*A****AA*TA*AT**C*A*T**A
<M01512:26:000000000-C4M89                                                        AAAAAACTAAAA*AA***A*A****AA*TA*AT**T*A*T**A
>45e8482b-bc4e-474c-9f79-0 *A**A**G**GC***G*C**T*C*T**AAAAAG*AAG**A*T**C*T***AAA**A**C*TATAA*A*AA***A*A****AA*TA*AT**T*A*T**A
>d7e672a5-e6d6-4386-9fb4-0 *A**A**G**GC***G*C**T*C*T***AAAAG*AAG**A*T**C*T**AAAA**A**C*TAT******A***A*A****AA*TA*AT**T*A*T**A
>c5d39f5a-26ab-4eb5-af50-d *A**A**G**GC***G*C**T*C*T***AAAAG*AAG**A*T**C*T**AAAA**A**C*TAT***A*AA***A*A****AA*TA*AT**T*A*TA*A
>56ea52c6-544c-4ceb-97c6-a *A**A**G**GC***G*C**TCC*T*AAAGA***AAG**A*T**C*T**AAAA**A**C*TAT*********A*A****AA*TA*AT**T*A*T**A
>3fb28f6c-253a-4f3b-b20b-9 *AAGA**G**GC***G*C**T*C*T*AAGAA***AAG**A*T**C*T**AAAA**A**C*T************A*A****AA*TA*AT**T*A*T***
>54c99c53-39c8-42a3-a3e0-d *A**A************C**T*C****AAAAAG*A****A*T**C*T**AAA**************G*CA***A*A****AA*TC*AG**T*A*T***
>b0298c2a-f30b-45f1-a2a7-b ****A**G**GC***G*C**TCC*T***AAAAG*AAG**A*T**C*T**AAAA**A**C*TATAA*A*AA***A*A****AA*TA*AT**T*A*T**A
>448d5b88-dd63-443a-83db-c *A**A**G**GC***G*C**T*C*T**AAAAAG*A****A*T**C*T***AAA**A**C*TAT***A*AA***A*A****A***C*AT**T*A*T**C
>73a9f80b-c034-4d6b-b805-7 *A**A**G**GC***G*C**T*C******AAAG*AAG**A*T**C*T**AAAA**A**C*TAT***A*AA***A*AAAT*AA*TA*AT**T*A*T**A
>dd9acf56-72b1-4061-9563-8 *A**A**G**GC***G*C**T*C*C*TAAAAAG*AAG**A*T**C*T**AAAA**A**C*TAT***A*AA***A*A****AA*T**AG**T*A*T**A
>cab563ba-5827-4add-a05d-6 *A**A**G**GC***G*C**T*C*T*AAAAAGA*AAG**A*T**C*T**AAAA**A**C*TAT***A*AA***A*A****AA*TA*AT**T*A*T**A
16043
Base T confidence: 30 (Prob. 0.999000, raw -34.8 -34.8 -34.8 30.0) Position -6447 cab563ba-5827-4add-a05d-609f22c9eb86
```

b



```
 Consensus                 *T*GC*****T***TT**T**C***T**AG*TT**C*C**C*GG**A*G******G***T*GCA****T*C***A*TT**A****G*G***TT*A**T
                           20      36130     36140     36150     36160     36170     36180     36190     36200     36210     3
>M01512:26:000000000-C4M89 *T*AC*****T***TT**T**C***T**AG*TT**C*CTGTCGGTGCATACTTAGGTT
<M01512:26:000000000-C4M89 *T*GC*****T***TT**T**C***T**AG*TT**C*C**T*GG**A*G******G***T*GCA****T*C***A*TT**A****G*G***TT*A**T
<M01512:26:000000000-C4M89 *T*GC*****T***TT**T**C***T**AG*TT**C*C**T*GG**A*G******G***T*GCA****T*C***A*TT**A****G*G***TT*A**T
>M01512:26:000000000-C4M89                                                                                             T*A**T
>M01512:26:000000000-C4M89 *T*GC*****T***TT**T**C***T**AG*TT**C*C**T*GG**A*G******G***T*GCA****T*C***A*TT**A****G*G***TT*A**T
>M01512:26:000000000-C4M89 *T*GC*****T***TT**T**C***T**AG*TT**C*C**T*GG**A*G******G***T*GCA****T*C***A*TT**A****G*G***TT*A**T
>M01512:26:000000000-C4M89 *T*GC*****T***TT**T**C***T**AG*TT**C*C**T*GG**A*G******G***T*GCA****T*C***A*TT**A****G*G***TT*A**T
<M01512:26:000000000-C4M89 *T*GC*****T***TT**T**C***T**AG*TT**C*C**T*GG**A*G******G***T*GCA****T*C***A*TT**A****G*G***TT
>M01512:26:000000000-C4M89 *T*GC*****T***TT**T**C***T**AG*TT**C*C**T*GG**A*G******G***T*GCA****T*C***A*TT**A****G*G***TT*A**T
<M01512:26:000000000-C4M89                                                                              T**A****G*G***TT*A**T
>M01512:26:000000000-C4M89 *T*GC*****T***TT**T**C***T**AG*TT**C*A**T*GG**A*G******G***T*GCA****A*C***A*TT**A****G*G***TT*A**T
>M01512:26:000000000-C4M89 *T*GC*****T***TT**T**C***T**AG*TT**C*C**T*GG**A*G******G***T*GCA****T*C***A*TT**A****G*G***TT*A**T
<M01512:26:000000000-C4M89 *T*GT*****T***TT**T**C***T**AG*TT**A*C**T*GG**A*G******G***T*GCA****T*C***A*TT**A****G*G***TT*A**T
>0063ec21-4afb-4818-a051-4 *T*GC*****T***TT**T**C***T**AG*TT**C*A**C*****A*G******G***T*GCA****T*C***A*TT**A****G*G***TT*A**T
<009d660f-c654-4fe3-a817-3 *T*GC*****T***TT**T**C***T**AG*TT**C*C**CTGG**A*G************GCA****T*C***A*TT**A****G*G***TTCA***
>00c0869f-fd06-42c6-a934-9 *T*GC*****T***TT**T**C***T**AG*TT**C*C**T*GG**A*G******G***T*GCA****T*C***A*TT**A****G*G***TT*A**T
<00ce5b78-65b2-4caa-8893-2 *T*GC*********TT**T**C***T**AG*TT**C*C**CTGG**A*G******G***T*GCA****T*C*********A****G*G***CT*A**T
<01232738-60c8-49b8-af89-1 *T*GCT****T***TT**T**C***T**AG*TT**C*C**C*GG**A*G******G***T*GCA****T*C***A*TT**A****G*G***TT*A**T
>02a714b1-0039-4bfb-a79d-a GAGAGGCATTGCTAGAGTTCTTTCT CAGGTGAAG CTTG TCGTGCT CCTGCCGAGCCTCTTGGTC G CTTTC GTTGCCCT CACCTCATAGCA
<035ec30d-4b17-4c6d-a54a-c *T*GC*****T***TT**T**C***T**AG*TT**C*C**T*GG**A*G******G***T*GCA****T*C***A*TT**AG***G*G***CT*A**T
>03bdd9b4-f245-4e59-9ee5-e *T*GC*********TT**T**C***T**AG*TT**CTG**A*GG**A*G******G***T*GCA****T*C***A*TT**A****G*G***TT*A**T
>089e3f24-48a6-4f8c-b3a8-5 *T*GC*****T***TT**T**C***T**AG*TT**C*A**C*GG**A*G******G***T*GCA****T*C***A*TT**A****G*G***TT*A**T
>08d4b36b-1fbb-49a6-9b32-2 *T*GC*********TT**T**C***T**AG*TT****C**T*GG**A*G******G***T*GCA****T*C***A*TT*******G*G***TT*A**T
>0aafc4fe-94ef-43d5-8a79-8 *T*GC*****T***TTC*C**C***T**AG*TT**C*A**C*GG**A*G******G***T*GCA****T*C***A*TT**A****G*G***TT*A**T
>0d6d3f78-89d3-4eaf-a686-f *T*GC*****T***TT**T**C***T**AG*TT****C**T*GG**A*G******G***T*GCA****T*C***A*TT**A****G*G***TT*A**T
<0dd4bb29-2ffe-4754-8925-3 *T*GC*****T***TT**T**C***T**AG*TT**C*T**C*GG**A*G******GTAAT*TCT****T*C***A*TT**A****G*G***TT*A**T
<0df3e02b-dcc6-411d-a439-c *T*GC*****T***TT**T**C***T**AG*TT**C*C**G*GG**A*G******G***T*GCA****T*C***A*TT**A****G*G***TT*A**T
<0ff02cf0-73db-4c2d-8630-a *T*GC*****T***TT**T**C***T**AG*TT**C*C**CTTG**A*G************GCA****T*C***A*TT**A******G***TT*A**T
<1013ec90-527b-4b25-b0fd-7 *T*GC*****T***TT**T**C***T**AG*TT**C*C**C*GG**A*G******G***T*GCA****T*C***A*TT**A****G*G***TT*A**T
<119dc24d-402c-4f95-914d-d *T*GC*********TT**T**C***T**AG*TT**C*C**C*GG**A*G******G***T*GCA****T*C***A*TT**A****G*G***TT*A**T
>126bb177-c063-4f2a-848e-2 *T*GC*********TT**T**C***T***G*TT**C*A**C*****A*G******G***T*GCA****T*C***A*TT*******G*G***TT*A**T
<1285835e-0067-47f6-a889-0 *T*GC*****T***TT**T**C***T**AG*TT**C*C**C*CG**A*G******G***T*GCA****T*C***A*TT**A****G*G***TT*A**T
36119
Base * confidence: 9 (Prob. 0.874107, raw inf inf inf inf) Position 23351 1573f95c-f975-44bc-a241-9203ca0981a4
```

Figure 3. Editing clone assemblies in Gap5**.** Screenshots from Gap5 with reads sorted by technology (Illumina on top; nanopore on bottom), showing two instances where errors in the consensus can be resolved by correcting to the consensus of the Illumina reads: a) frequent insertion and deletion errors at homopolymer runs, and b) more rare substitution errors.

Table 1 | Troubleshooting Table.

Step	Problem	Possible Reason	Solution
33	Low DNA concentration	Culture undergrowth or overgrowth	Check culture OD600 is between 0.2-0.35
		Incomplete Lysis	Make sure to thoroughly mix the solution until the color is uniform
		Incomplete neutralization	Solution from step 13 should not appear viscous and precipitate should float to the surface
		Incomplete DNA elution	Pre-warm elution buffer to 50°C
34	Concentration varies when checking with NanoDrop or Qubit	DNA is not completely mixed	After adjusting concentration from step 33, leave DNA solution on a heated shaker at the gentlest setting at 50°C until DNA is completely mixed
51	Pores decrease rapidly	Impure DNA sample	Re-check DNA concentration. Extract DNA again if NanoDrop and Qubit results are discordant, 260/280 < 1.7, 260/280 > 2.0, 260/230 < 2.0, or 260/230 > 2.2
		Bubbles introduced during loading	Pipet very slowly and take care not to introduce bubbles during flow cell priming and library loading
55	No reads for one or more clones	Clone culture failed	Regrow and add to the next run, or replace the clone with another Regrow the clone for an additional round of sequencing
		Bookkeeping error; some common bookkeeping errors result from transposing digits, rotating a plate by 180°, or contamination from a clone in an adjacent well	Resolve bookkeeping error, and rerun a new clone or replace with another clone
58	Low fraction of long reads	FRA treatment time too long	Promptly heat-inactivate FRA at 35 seconds Adjust the FRA incubation time below 35 seconds
		Shearing during library prep	Use wide-bore tips for all mixing and loading steps

71	Clone sequence is shorter than expected or missing known sequence	Deletion during culture	Regrow the clone from the original culture or another library copy, and replace with the alternate clone
		Sequence toxic to *E. coli*	Close the gap by long-range PCR or region-specific extraction

Anticipated Results

We typically pool 24 clones for a single MinION run, generating about 300,000 reads with a read n50 of 20 kb, and a total of about 1.5 Gb of sequence data. Each clone typically receives 1-5% of the total reads. Occasionally some clones will have no reads; this usually indicates that the culture of the clone (Steps 1-3) has failed (see troubleshooting information for Step 55). Expect to obtain 3-10 full-length reads per clone. Because of the high rate of insertions and deletions in individual nanopore reads, full-length reads may differ in length by 10 kb or more. Occasionally, a clone will have no reads that start and end in vector sequence, but the clone length will be apparent from a peak in the tail of the distribution of read lengths. It may still be possible to reconstruct a full-length consensus sequence by rotating one of these putative full-length reads to place the vector sequence at the beginning. However, we do not recommend this procedure for internally repetitive clones, particularly tandem arrays. Instead, sequence the clone again, and use these ambiguous reads to help polish the consensus.

REFERENCES

Bellott DW, Cho TJ, Hughes JF, Skaletsky H, Page DC. 2018. Cost-effective high-throughput single-haplotype iterative mapping and sequencing for complex genomic structures. *Nat Protoc* **13**: 787–809.

Bellott DW, Skaletsky H, Pyntikova T, Mardis ER, Graves T, Kremitzki C, Brown LG, Rozen S, Warren WC, Wilson RK, et al. 2010. Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition. *Nature* **466**: 612–616.

Bonfield JK, Whitwham A. 2010. Gap5—Editing the billion fragment sequence assembly. *Bioinformatics* **26**:1699–1703.

Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Graves TA, Nefedov M, Rosenfelt JA, Sajjadian S, Malig M, Kotkiewicz H, et al. 2010. Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell* **149**: 912–922.

Giachini C, Nuti F, Turner DJ, Laface I, Xue Y, Daguin F, Forti G, Tyler-Smith C, Krausz C. 2009. TSPY1 copy number variation influences spermatogenesis and shows differences among Y lineages. *J Clin Endocrinol Metab* **94**: 4016–4022.

Gordon D, Huddleston J, Chaisson MJP, Hill CM, Kronenberg ZN, Munson KM, Malig M, Raja A, Fiddes I, Hillier LDW, et al. 2016. Long-read sequence assembly of the gorilla genome. *Science* **352**: aae0344.

Gordon D, Green P. 2013. Consed: A graphical editor for next-generation sequencing. *Bioinformatics* **29**: 2936–2937.

Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Graves T, Fulton RS, Dugan S, Ding Y, Buhay CJ, Kremitzki C, et al. 2012. Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* **483**: 82–86.

Hughes JF, Skaletsky H, Pyntikova T, Graves TA, van Daalen SKM, Minx PJ, Fulton RS, McGrath SD, Locke DP, Friedman C, et al. 2010. Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* **463**: 536–539.

Jain M, Olsen HE, Turner DJ, Stoddart D, Bulazel KV, Paten B, Haussler D, Willard HF, Akeson M, Miga KH. 2018a. Linear assembly of a human centromere on the Y chromosome. *Nat Biotechnol* **36**: 321–323.

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, et al. 2018b. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* **36**: 338–345.

Kuroda-Kawaguchi T, Skaletsky H, Brown LG, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Silber S, Oates R, Rozen S, et al. 2001. The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. *Nature Genet* **29**: 279–186.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Li H. 2020. A versatile pairwise aligner for genomic and spliced nucleotide sequences. *minimap2* https://github.com/lh3/minimap2.

Li H. 2018. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100.

Lin S, Lin Y, Nery JR, Urich MA, Breschi A, Davis CA, Dobin A, Zaleski C, Beer MA, Chapman WC, et al. 2014. Comparison of the transcriptional landscapes between human and mouse tissues. *PNAS* **111**: 17224–17229.

Lupski JR. 1998. Genomic disorders: Structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet* **14**: 417–422.

Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA, et al. 2020. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585:** 79–84.

Mueller JL, Skaletsky H, Brown LG, Zaghlul S, Rock S, Graves T, Auger K, Warren WC, Wilson RK, Page DC. 2013. Independent specialization of the human and mouse X chromosomes for the male germline. *Nature Genet* **45**: 1083–1087.

Oram SW, Liu XX, Lee TL, Chan WY,Lau Y-FC. 2006. TSPY potentiates cell proliferation and tumorigenesis by promoting cell cycle progression in HeLa and NIH3T3 cells. *BMC Cancer* **6**: 154.

Quick J. 2018. Ultra-long read sequencing protocol for RAD004 v3 (protocols.io.mrxc57n). doi:10.17504/protocols.io.mrxc57n.

National Center for Biotechnology Information. 2017. What is tbl2asn? https://www.ncbi.nlm.nih.gov/genbank/tbl2asn2/.

She X, Jiang Z, Clark RA, Liu G, Cheng Z, Tuzun E, Church DM, Sutton G, Halpern AL, Eichler EE. 2004. Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431**: 927–930.

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown 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.

Soh YQS, Alfoldi J, Pyntikova T, Brown LG, Minx PJ, Fulton RS, Kremitzki 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.

Tsuchiya K, Reijo R, Page DC, Disteche CM. 1995. Gonadoblastoma: Molecular definition of the susceptibility region on the Y chromosome. *Am J Hum Genet* **57**: 1400–1407.

Tyler-Smith C, Taylor L, Müller U. 1988. Structure of a hypervariable tandemly repeated DNA sequence on the short arm of the human Y chromosome. *J Mol Biol* **203**: 837–848.

Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* **27**: 737–746.

Warburton PE, Hasson D, Guillem F, Lescale C, Jin X, Abrusan G. 2008. Analysis of the largest tandemly repeated DNA families in the human genome. *BMC Genomics* **9**: 533.

Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein J, Willsey AJ, Joy JB, Scott JK, Graves TA, et al. 2013. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *The Am J Hum Genet* **92**: 530–546.