Inferring chimpanzee Y chromosome history and amplicon diversity from whole genome sequencing

Matthew T. Oetjens[1], Feichen Shen[1], Zhengting Zou[2], Jeffrey M. Kidd*[1,2]

[1]Department of Human Genetics, University of Michigan Medical School, Ann Arbor, MI, 48109, USA
[2] Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI, 48109, USA

* To whom correspondence should be addressed.
Tel: (734) 763-7083; Email: jmkidd@med.umich.edu

Key words:
Chimpanzee, bonobo, Y chromosome, amplicon

Abstract

Due to the lack of recombination, the male-specific region of the Y chromosome (MSY) is a unique resource for tracking the genetic history of populations. The MSY is also enriched for large, nearly identical repetitive regions known as amplicons, which harbor many of the genes essential for spermatogenesis. In humans, sequence diversity on the unique segment of the MSY is greatly reduced compared to the autosomes, an observation consistent with the action of strong selection. Here, we analyze 9 chimpanzee (representing three subspecies: *Pan troglodytes schweinfurthii*, *Pan troglodytes ellioti*, and *Pan troglodytes verus*) and two *Pan paniscus* male whole-genome sequences to assess Y chromosome nucleotide and ampliconic copy-number diversity across the *Pan* genus. In total, we identified 23,946 *Pan spp.* SNVs across 4.2 million callable sites. Comparisons with autosomal, X chromosome, and mitochondrial sequences from the same samples indicate that nucleotide diversity on the chimpanzee MSY is reduced relative to neutral expectations with an equal sex ratio. Additionally, the estimated common chimpanzee Y chromosome T_{MRCA} (0.44 mya [0.31-0.56]) is half the age of the mitochondria T_{MRCA} (0.97 mya [0.65-1.35]), indicating an unequal sex ratio or Y chromosome selection in the common chimpanzee ancestral population. We observe that the copy-number of Y chromosome amplicons is variable amongst chimpanzees and bonobos, and identify several lineage-specific patterns, including variable copy-number of the testis-expressed genes *RBMY* and *DAZ.* We detect recurrent switchpoints of copy-number change along the ampliconic tracts across chimpanzee populations, which may be the result of localized genome instability or selective forces.

Introduction

In mammals, males paternally inherit the Y chromosome, which contains the genetic content for canalizing embryos toward masculine development. With few exceptions (Warren, et al. 2008), the male specific region of the Y chromosome (MSY) is in a haploid state and does not undergo recombination during prophase I of meiosis. A nearly identical MSY sequence is transmitted from father to son with some changes as the result of random mutation. Consequently, the MSY has been exploited for population studies since relationships between haplotypes can be directly inferred with simple models that rely only on the mutation rate or calibrations from the fossil record (Jobling and Tyler-Smith 2003). Similarly, the mitochondrion is a maternally inherited non-recombinant locus. For these reasons, the mitochondria and the MSY are powerful markers for inferring sex-specific demographies. Although these loci each represent single lineages, differences between estimates produced by each can reveal differences in ancestral sex ratios, variances in reproductive success, sex-biased migration events, and selection acting on these loci (Karmin, et al. 2015). Recently, next-generation sequencing (NGS) has been applied to diverse panels of human Y chromosomes and mitochondria to reconstruct the phylogenetic tree with greater resolution (Francalacci, et al. 2013; Karmin, et al. 2015; Poznik, et al. 2013; Wei, et al. 2013).

Currently, the sequences of four mammalian MSYs have been assembled: the human (*Homo sapiens*), chimpanzee (*Pan troglodyes*), rhesus macaque (*Macaca mulatta*) and mouse (*Mus musculus*) (Hughes, et al. 2012; Hughes, et al. 2010; Skaletsky, et al. 2003; Soh, et al. 2014). Two major sequences classes, ampliconic and X-dengenerate, dominate the landscape of the human and chimpanzee MSY. The ampliconic region consists of interspersed palindromes, tandem and inverted repeats, and has undergone exceptional structural rearrangements since the chimpanzee-human divergence ~6 million years ago (mya) (Hughes, et al. 2010). The repetitive elements termed amplicons display a high degree of self-identity, which renders their sequences unsuitable for reliable genotype calling (Poznik, et al. 2013). Amplicon regions are under constant change in the human lineage with structural polymorphisms accruing at ~10,000 times the rate of nucleotide polymorphism (Repping, et al. 2006). In contrast, the X-degenerate sequence is composed of long stretches of sequence containing single-copy genes and amenable to short read mapping. The X-degenerate region also appears to have remained relatively stable in the primate lineage, and conservation of its genic content is evidence that the Y chromosome is not degenerating to the point of extinction (Graves 2006; Hughes, et al. 2012). Additionally, there are short stretches of sequence on the Y chromosome near the telomeres that exhibit XY homology and undergo recombination, termed the psuedoautosomal regions (Helena Mangs and Morris 2007). NGS analyses of the MSY for phylogenetic inference rely on quality control filters based on read depth and mapping quality to identify callable regions of sequence (Poznik, et al. 2013). In general, this filtering strategy reduces the data to sites within the X-degenerate region.

The principal role of the genic content on the Y chromosome is in conferring sexual dimorphism through the expression of the Y-linked sex-determining region *(SRY)* (Lahn, et al. 2001; Sinclair, et al. 1990). The remaining genes exhibit a striking consistency in function as either homologues of housekeeping genes

found on the X chromosome or genes highly expressed in the testes that lack an active X linked homologue (Lahn and Page 1997; Lahn, et al. 2001). The abundance of genes with testis-specific expression in the MSY suggests it is important in spermatogenesis (Lahn, et al. 2001). In fact, deletions in the AZFc region of the human MSY are associated with increased risk of spermatogenic failure (Kuroda-Kawaguchi, et al. 2001; Rozen, et al. 2012).

Recent studies have reported that the human MSY is under significant selective constraints that reduce its sequence diversity relative to autosomes (Malaspina, et al. 1990; Wilson Sayres, et al. 2014). All sites are linked on the MSY, therefore purifying selection that acts to remove harmful variants will lower diversity across the entire MSY, in a process termed background selection (Bachtrog 2013; Muller 1964). Human population studies have revealed that genes in the X-degenerate region display a strong bias towards synonymous over non-synonymous substitutions (Rozen, et al. 2009). Additionally, forward simulations modeling diversity of the MSY in human populations have demonstrated that the number of sites that would have to be under purifying or positive selection to explain the low levels of diversity observed exceed the amount of coding sites in the X-degenerate region (Wilson Sayres, et al. 2014). An alternative hypothesis is that additional selection on the MSY is driven by traits related to male fertility affected by the testis-specific genes located in the ampliconic regions. Intriguingly, the X-linked ampliconic genes expressed predominately in the testis have been implicated in hard selective sweeps across the Great Apes (Nam, et al. 2015). However, it is unknown whether selection on the MSY is human-specific or a phenomenon inherent to primate MSYs.

Copy number variation (CNV) of ampliconic genes has been observed in chimpanzees and humans by florescence *in situ* hybridization (FISH) experiments (Repping, et al. 2006; Schaller, et al. 2010). *Pan* species exhibit polyandrous mating behavior, which can drive sperm competition among males (Good, et al. 2013). It is hypothesized that CNV of the spematogenic genes in the ampliconic region confers male-specific reproductive advantages, thereby subjecting this region to selection for structural changes. Interestingly, in the monandrous gorillas, where sperm competition is expected to be low, CNV of Y-linked spermatogenic genes among individuals has not been observed (Greve, et al. 2011). However, limited data on Great Ape MSY diversity is available as the chromosome is absent from the gorilla (gorGor3.1) and the orangutan (ponAbe2) assemblies (Locke, et al. 2011; Scally, et al. 2012).

The two extant species within the *Pan* genus are the common chimpanzee (*Pan troglodytes*) and the bonobo (*Pan paniscus*), which diverged from a common ancestor in Central Africa sometime between one and two mya. Since the chimpanzee-bonobo divergence, founder events and/or migrations of chimpanzees to geographically isolated regions of Africa have resulted in the formation of at least four genetically distinct subspecies. The Central chimpanzee, *(Pan troglodytes troglodytes*) is believed to be the oldest subspecies, and shows genetic evidence of a long-standing large population size (Wegmann and Excoffier 2010). The Eastern chimpanzee *(Pan troglodytes schweinfurthii*) emerged from the Central chimpanzees ancestors by a founder event and is the most recent sub-speciation event amongst chimpanzees (Hey 2010). Moving westward in Africa from the Central-Eastern chimpanzee distribution, two related subspecies are recognized: Nigerian-Cameroon

chimpanzee (*Pan troglodytes ellioti*) and Western chimpanzee (*Pan troglodytes verus*). The Sanaga River serves as a natural barrier between Nigerian-Cameroon and Central chimpanzees, however the geography that maintains the isolation between the Nigerian-Cameroon and Western subspecies is not fully understood.

Here, using 11 chimpanzee and bonobo samples sequenced by the Great Ape Genome Project, we conducted a comprehensive analysis of MSY diversity in the *Pan* lineage. Our analysis of nucleotide and copy number diversity is focused on the X-degenerate and ampliconic regions, respectively. To assess nucleotide diversity within the X-degenerate region, we applied a two-part filtering scheme (regional and site level) and identified 23,946 SNVs (without human) across 4.2 million callable sites. From these data, we reconstructed a phylogenetic tree of the *Pan* lineage and provide T_{MRCA} estimates based on the MSY. We then compared *Pan* MSY diversity to autosomes, the X, and mitochondria to reveal a similar reduction of MSY nucleotide diversity in *Pan* as in human populations. To assess *Pan* amplicon diversity, we applied a novel k-mer counting pipeline to estimate the CNV of amplicons across samples, and we observed amplicon copy-number to be highly variable in the *Pan* lineage, enabling the identification of several lineage-specific patterns.

Results

Callset Construction

We performed a series of analyses to identify segments of the X-degenerate region for variant calling using Illumina sequence data. Following the procedure developed by Poznik et al. (Poznik, et al. 2013) for the analysis of the human MSY, we applied a two stage filtering procedure based on read depth, mapping quality, and the presence of heterozygous genotype calls in male samples. Since masks produced separately from the two bonobos showed minimal differences in comparison to masks produced from the 9 chimpanzees (Supplemental Figure 1), we analyzed the 11 chimpanzee samples together (Figure 1A). To serve as an out-group, a callable genome mask was also constructed based on mapping of human sample HGDP00222 to the chimpanzee genome reference (Figure 1B). This resulted in a final set of 4,233,540 bp of sequence that passed the human and *Pan* filters (Supplemental Table 1). As expected, the bounds of callable sequence fall within the previously mapped boundaries of the X-degenerate region. We also performed a similar analysis to identify 75,176,291 callable bp on the X chromosome, and 106,767,230 on chromosome 7, and we then utilized constructed mitochondria assemblies to enable comparisons among genomic compartments with alternative inheritance patterns.

Chimpanzee MSY Sequence Diversity

We identified 7,752 variants including 1,672 singletons, in the *Pan troglodyte* only dataset, 23,946 (5,871 singleton) variant sites in a *Pan* only dataset, and 78,022 (59,446 singleton) variant sites when including the human outgroup. The Western chimpanzee Clint is a major component of the panTro4 assembly (Hughes, et al. 2010), and we identified only six positions on the MSY where the inferred Clint allele differs from the PanTro4 assembly. Among the remaining samples, the Western chimpanzee Bosco and the hybrid (Nigerian-

Cameroon/Western) chimpanzee Donald show the lowest amount of raw nucleotide differences with Clint, 103 and 73, respectively, suggesting that Donald's MSY is of Western origin. However, given the complex ancestry on the rest of Donald's genome (Prado-Martinez, et al. 2013), he was excluded from subsequent analysis unless otherwise noted. The average number of differences between Clint and the Eastern and Nigerian-Cameroon samples was 4,488 and 3,952.5, respectively. As expected, the two bonobos had the largest number of differences with Clint (mean 16,435). We also counted the number of nucleotide differences between the human sample and each chimpanzee and observed modest variation between samples: the maximum number of differences was 62,931 (bonobo) and the minimum 62,654 (Eastern).

To measure genetic diversity, we calculated π, the average nucleotide differences per site, on the MSY within sample groups (Table 1). Averaged across the three common chimpanzee subspecies, π equaled 7.76 x 10^{-4} substitutions per base pair (bp). Previous studies have indicated that the autosomes of bonobos and Western chimpanzees have the lowest levels of heterozygosity among the Great Apes, with the exception of Non-African humans (Prado-Martinez, et al. 2013). Among our limited samples, we find that the Western chimpanzee has the lowest MSY π (2.43 x 10^{-5} substitutions per bp) and the bonobos have the highest (9.85 x 10^{-4} substitutions per bp). Strikingly, the number of nucleotide differences between the two bonobo samples suggests a T_{MCRA} as old as the T_{MRCA} of the common chimpanzee clade (Figure 2). To identify regional differences in diversity, we calculated average π and "d" (divergence) in contiguous 1kb windows across the Y chromosome (Supplemental Figure 2). In general, we observe that π and d values are mostly sustained across the positions within our callable sites. Intriguingly, a peak of high diversity at chrY: 22,759,00-22,791,000 is observed between the two bonobos. However, no genes were annotated in this region and a BLAST search across the NCBI nucleotide collection did not reveal homologous sequence of functional significance.

Phylogenetic Tree Reconstruction and Estimates of the Time to Most Recent Common Ancestor

To explore the relationships among the samples, we constructed a neighbor-joining phylogenetic tree based on MSY sequence data (Figure 2 and Supplemental Figure 3A). Our tree reveals a deep chimpanzee-bonobo divergence relative to branches extending from common chimpanzee subspecies nodes. Consistent with autosomal and mitochondria trees (Bjork, et al. 2011; Prado-Martinez, et al. 2013), the monophyletic nature of three chimpanzee subspecies clades are supported as well as a deeper Nigerian-Cameroon and Western chimpanzee node, reaffirming the relationship between the two groups.

To recover sex-specific histories in *Pan*, we estimated the MSY and mitochondria T_{MRCA}s at nodes of interest using BEAST (Drummond, et al. 2012). We focused our analysis on the divergence between common chimpanzee and bonobos, the chimpanzee subspecies, and the Nigerian-Cameroon and Western subspecies. BEAST can model phylogenetic inference under a relaxed-clock model (Drummond, et al. 2006), which allows for rate heterogeneity among branches. Evolutionary rates can vary across a tree because of differences in generation time, metabolic rate, efficiency of DNA repair, population size and selective pressure(s) (Pybus

2006). Since our dataset includes both inter- and intra-specific data, we applied a two-part Yule-bootstrap/Coalescent method that was developed for an analysis of chimpanzee mitochondrial samples (Bjork, et al. 2011) (see methods). We calibrated the rate of evolution of our mitochondrion and MSY datasets by applying a *Pan-Homo (*T_{MRCA} = 6 mya) prior sampled from a lognormal distribution. At nodes of interest and using a comparable calibration, we were able to closely replicate T_{MRCA} estimates previously reported using mitochondria sequence data by Bjork et al. 2011 (Table 2), suggesting that our dataset is sufficient for producing analogous estimates with the MSY data. As expected, the Yule-Bootstrap and chimpanzee-only coalescent analyses produce similar T_{MRCA} estimates at the Nigerian-Cameroon and *Pan troglodytes* ancestral nodes (Table 2, Supplemental Figure 4 and Supplemental Figure 5). To investigate the degree of rate heterogeneity (σ_r) amongst branches in the two coalescent trees, we examined the posterior distribution of the two coefficients of variation (Drummond, et al. 2006). A distribution of σ_r abutting zero is evidence for a strict molecular clock, whereas larger mean σ_r values indicate greater levels of heterogeneity. Interestingly, based on the coalescent trees, we find that we can reject the strict-molecular clock for the MSY (σ_r = 0.40, 95% HPD interval: 0.22-0.63) but not for the mitochondria (σ_r = 0.15, 95% HPD interval: 5.94×10^{-7} - 0.38).

In general, we find that the T_{MRCA}'s estimated from the MSY are younger than the mitochondria T_{MRCA}'s (Table 2). This is observation is most striking for the *Pan troglodytes* T_{MRCA}: the mean estimate of the MSY (0.44 [0.34- 0.56]) is less than half that of the mitochondria 0.97 [0.65 – 1.35]. The *Pan* T_{MRCA} estimate for the MSY is also younger, however, confidence intervals between the two estimates overlap (MSY: 1.57 [1.24 - 2.00]; mitochondria: 2.07 [1.46- 2.86]. In contrast, the Nigerian-Cameroon and Western T_{MRCA} show considerable overlap across the loci (MSY 0.37 [0.21 - 0.52] and the mitochondria 0.46 [0.21 – 0.52]).

The calculated *Pan troglodytes* T_{MRCA} values include coalescence in the ancestral population as well as divergence following species formation. Assuming a population split time of ~428 kya, (Prado-Martinez, et al. 2013) the ancestors of the surviving MSY and mitochondria lineages coalesced 10 kya and 539 kya prior to the Eastern/Central and Nigerian-Cameroon/Western separation (Supplementary Figure 6A). These values reflect the coalescence in the ancestral population of the two lineages that ultimately gave rise to the extant subspecies. The MSY and mitochondria T_{MRCA}s need not agree, since they each represent single loci and the coalescence process is stochastic. A larger female-to-male sex ratio in the ancestral population would reduce the expected MSY coalescence time relative to mitochondria, and might account for the observed differences in T_{MRCA} values. Since the expected coalescence time for two lineages has a high variance, we performed simulations to assess how unusual the observed difference in T_{MRCA} is. When there are equal male and female effective populations sizes the observed T_{MRCA} difference is highly unexpected, appearing in less than 0.0015% of simulations. In contrast, when the male effective population size is greatly reduced, the observed T_{MRCA} differences frequently occur (Supplementary Figure 6B).

The T_{MRCA}s produced in this study are the first produced by NGS data of a diverse panel of *Pan* MSY and represents a unique estimate based on male transmission of a haploid non-recombinant chromosome. In the

absence of gene flow, the T_{MRCA} of any locus sampled in multiple populations, including the MSY and the mitochondria, should be older than the population split times. Our estimates of the MSY and mitochondria T_{MRCA} broadly agree with previous estimates of chimpanzee population split times, although some differences are observed (Supplemental Table 2) (Bataillon, et al. 2015; Becquet and Przeworski 2007; Caswell, et al. 2008; Gonder, et al. 2011; Hey 2010; Prado-Martinez, et al. 2013; Stone, et al. 2010; Wegmann and Excoffier 2010).

These estimates, however, may be misleading since the MSY represents only a single locus and chimpanzee species have experienced gene flow following their separation (Becquet and Przeworski 2007; Hey 2010; Prado-Martinez, et al. 2013; Wegmann and Excoffier 2010; Won and Hey 2005). To gain perspective on Y chromosome divergence times relative to population separation times, we assessed split times using cross-population Pairwise sequentially Markovian coalescent (PSMC) analysis based on pseudo-diploid X chromosomes obtained from the males in our study (Li and Durbin 2011) (Supplemental Figure 7). Under a model of clean population separation, the time of population split is approximated by the time when there is no longer any cross-population coalescence, thus resulting in an extremely large N_E as inferred by PSMC; gene flow during population separation results in a step-wise increase in the PSMC curve. Comparisons between Eastern and Nigerian-Cameroon species show the most gradual increase in inferred N_E as well as the highest variation across comparisons, consistent with the high degree of gene flow inferred between these two species (Prado-Martinez, et al. 2013). Calibrating all mutation rates using a 6 mya divergence, we find that Y chromosome and mitochondria T_{MRCA}s predate the estimated population separation times, as indicated by the point at which the pseudo-diploid PSMC curves begin to increase.

Comparisons with X, Autosomal, and Mitochondria Diversity

The observed nucleotide diversity, π, is a function of the effective population sizes (N_E) and the mutation rate. Under neutral conditions and adjusting for chromosome-specific mutation rates, π on the X, Y, and mitochondria are decreased relative to the autosomes proportional to the chromosome-specific N_E. The MSY and mitochondria are haploid and transmitted by only one gender, and consequently their N_E is reduced to one-fourth that of the autosomes when the male and female population sizes are equal. The X chromosome is haploid in males and diploid in females, which reduces the N_E to three-fourth that of the autosomes under an equal sex ratio. However, nucleotide diversity on the MSY in human populations is far lower than neutral expectations would predict (Wilson Sayres, et al. 2014). We calculated autosomal corrected diversity in our data to determine if the MSY is similarly reduced in our *Pan* samples. To correct for chromosome-specific mutation rates we estimated *d* across the genomic regions based on the average pairwise differences between the *Pan* and the human samples. Assuming a 25-year generation time and a 6 million year *Pan-Homo* divergence, the inferred mutation rates per base pair per generation of the genomic compartments under study are 2.49×10^{-8} for the autosomes, 1.78×10^{-8} for the X chromosome, 3.09×10^{-8} for the Y chromosome, and 1.96×10^{-7} for the mitochondria. We note that these phylogenetic rates are larger than recent estimates based on direct sequencing

of pedigrees, which imply a more ancient *Pan-Homo* divergence (Langergraber, et al. 2012; Scally and Durbin 2012; Venn, et al. 2014).

Similar to human populations, we observe a reduction of diversity on the MSY in chimpanzee subpopulations relative to expectations under an equal sex ratio (Figure 3). The autosome corrected MSY diversity in all three chimpanzee subspecies were below 0.25 (Nigerian-Cameroon chimpanzees: 0.14, Eastern chimpanzees: 0.12, and Western chimpanzees: 0.06). However, the bonobo MSY diversity strongly contradicts this trend of low MSY diversity as we find Y/A (chromosome Y / autosome) diversity above 1 (1.02). This is consistent with the deep divergence we see between the bonobo MSYs (Figure 2). Notably, across all three genomic compartments the Western chimpanzees have the lowest and the bonobo's have the highest autosome corrected π/d. It is surprising the two groups are disparate, given that autosomal π of the bonobos and the Western chimpanzees are low amongst the Great Apes. On the other hand, mitochondria diversity conforms to neutral expectation in Nigerian-Cameroon chimpanzees (0.27) and in Eastern chimpanzees (0.26). Similar to the MSY, we see a strong depression of mt/A diversity in Western chimpanzees (0.15) and inflation of bonobo diversity (1.29). We compared our chimpanzee Y/A values to those reported in a human study and find higher values in *Pan* than *Homo*, which is consistent with previous observations (Stone, et al. 2002; Wilson Sayres, et al. 2014). We do not observe a 3/4 ratio of X/A diversity in any of the chimpanzee subspecies and instead report significant inflation of diversity across the *Pan* lineage. However, X chromosome diversity deviating from expectations under an equal sex ratio has also been observed in human populations (Hammer, et al. 2008; Keinan, et al. 2009). The X chromosome also appears to have been under stronger evolutionary pressures than the autosomes, which complicates measurements of π and d (Dutheil, et al. 2015; Hammer, et al. 2010; Hvilsom, et al. 2012) .

A comparative study of mitochondrial diversity revealed that chimpanzees have a significantly higher female to male effective population size ratio than humans (Gaggiotti and Excoffier 2000). Additionally, the effective sizes of chimpanzee populations have undergone radical changes over time, which can alter the strength of genetic drift across different genomic compartments (Pool and Nielsen 2007). Therefore, we performed simulations of autosome corrected diversity modeled with historical demographic scenarios informed by PSMC analyses of the chimpanzee populations under a variable male-to-female gender bias (Prado-Martinez, et al. 2013). Across all chimpanzee subspecies, we find lower Y/A values than predicted by simulations under an equal sex ratio (Supplemental Figure 8). However, we caution that our estimates of autosome corrected diversity may be sensitive to the limited sample size that we have available and larger studies will be required to more fully characterize these differences in diversity.

Selection of X-degenerate genes in the MSY

Comparative genomics studies have revealed that X-degenerate genes are under purifying selection (Rozen, et al. 2009). To observe how selection may have shaped X-degenerate diversity in chimpanzees, we characterized the function of variants identified in our study population. Of the 11 genes with callable sites, we

did not observe any samples with X-degenerate genes harboring non-sense, frame-shift, or splice site mutations. We calculated the dN/dS ratio of all genes in pairwise comparisons between each sample and a reference sample (Supplemental Figure 9). Compared with Clint, we observe a bias towards synonymous substitutions (*Pan*: $p = 1.60 \times 10^{-5}$; *P. trog.* $p = 8.40 \times 10^{-5}$). In fact, few non-synonymous variants are observed. However, after a Bonferroni correction for testing multiple genes and samples ($p < (0.05 / (11 \text{ genes} \times 12 - 1 \text{ samples})) = 4.13 \times 10^{-4}$), no single gene reaches statistical significance for dN/dS. Compared with the human sample, the bias towards synonymous substitution is noticeably amplified, likely the result of the larger number of observed substitutions accumulated as a result of the longer divergence time. In contrast to the bulk of the X-degenerate genes, *SRY* displays a consistent bias towards non-synonymous substitutions relative to the human in all chimpanzee subspecies and bonobos (Pamilo and O'Neill 1997). Overall, these results suggest that the X-degenerate genes remain under purifying selection in the chimpanzee lineage, with the notable exception of *SRY*.

Amplicon Diversity

Since ampliconic segments make up > 50% (14.7/ 25.8Mb) of the chimpanzee MSY and harbor many genes important for sperm development, we sought to characterize ampliconic CNV across our samples. First, we mapped the positions of amplicons on a whole chromosome dot-plot self-alignment and annotated stretches of self identify (~50-800kb) as amplicons (Supplemental File): in total we count 51 total units across 10 amplicon families (annotated as in Hughes et al. 2010). As expected, our amplicon definitions are nearly identical to those previously reported by Hughes et al., though one minor difference is that we mapped amplicon boundaries where homology is shared across all members of the family (i.e. all members within a family are close to equal in amplicon length and the copy number of an amplicon should be consistent by position based on the panTro4 assembly). Additionally, we mapped the unique "spacer" sequence between palindrome arms and identified ampliconic-gene positions (based on refSeq and ensembl) from the UCSC genome browser (Karolchik, et al. 2003). Using these segment definitions, copy-numbers were estimated for each amplicon family, spacer, and gene using the paralog sensitive method Quick-mer, a k-mer counting pipeline (see Methods). Donald was also included as an additional Western chimpanzee data point in this analysis given that his MSY is of Western chimpanzee origin.

As a control, we compared copy number estimates obtained from Clint's read data to copy number based our annotation of the reference sequence (Table 3A, Supplemental Table 3A). After rounding median k-mer values to the nearest integer, we find strong agreement between the values, suggesting that Quick-mer provides accurate estimates of amplicon copy number. However, a few disagreements between our estimate of Clint's copy number and the reference sequence are evident. Quick-mer estimates Clint's copy number of "violet arms" based on the median depth to be 11.54 (S.D. = 2.53) compared with the 13 represented in the reference assembly. Our copy number estimates are also different from the reference in the *CDY* (Clint: 3.98 vs. panTro4: 5), *VCY* (Clint: 0.77 vs. panTro4: 2), and *TSPY* (Clint: 5.46 vs. panTro4: 6) genes. Based on our data, we are

unable to determine if these differences arise from noise in our analysis or assembly errors in the panTro4 reference. We also measured copy number in the human sample (HGDP00222) based on the same 30-mers used to estimate chimpanzee copy-number. Many of our copy number estimates of HGDP00222 match the structures present in the human reference sequence as reported by Hughes et al. (Hughes, et al. 2010). Observed differences between the copy number present in the human assembly (hg19) and HGDP00222 reflect a loss of chimpanzee-human homology for some 30-mers or amplicon polymorphism differences between HGDP00222 and the hg19 reference.

We see substantial variance in the copy number of amplicons, spacers, and ampliconic-genes across the chimpanzee subspecies and bonobos (Figure 2, Table 3A and Table 3B). However, within individual *Pan* lineages we observe an overall conservation of copy number in the ampliconic regions and find that samples within lineages cluster together (Supplemental Figure 10). The teal and violet palindromes, found within the 2.2Mb palindrome array near the centromere in the reference sequence, appear to be diminished in copy number in the Western lineage, while the *RMBY* gene nested within the palindrome array is highly variable across the *Pan* lineages (median k-mer depth; Western: 9.16, Eastern: 14.34, Nigerian-Cameroon: 24.12, bonobos: 32.28) (Figure 4B). The Western chimpanzees are also comparably lacking in pink arms and the adjacent *TSPY* gene: the median copy number of the pink amplicon (3.81) and the *TSPY* gene (10.27) is one-fourth and one-half of the estimates in Nigerian-Cameroon chimpanzees (pink amplicon: 11.87, *TSPY*: 26.35) and bonobos (pink amplicon: 13.47, TSPY: 32.10), respectively. However, the *TSPY* gene shows uneven copy number by position. Copy number increases as ones moves along the TSPY gene across all samples, therefore it is possible our estimates of *TSPY* copy number (based on k-mer counts) may also include truncated of copies of the gene (Supplemental Figure 11T). The *DAZ* gene nested within the red palindrome is a candidate fertility factor whose deletions are linked to azoospermia in humans (Reijo, et al. 1995). We find that this gene and amplicon ranges in copy number between two (Eastern and bonobos) and four (Nigerian-Cameroon and Westerns) copies. Interestingly, within *DAZ* we observe two exon-spanning deletions in the human and bonobo samples that share a common break point located in intron 15 (ENSPTRG00000022523) (Figure 4A). A similar event is observed in the purple arms (Figure 4D), where there is amplification in the Eastern chimpanzees and deletion in the bonobos on the 5' of a break point and the reverse of the pattern, amplification of Bonobos and deletion of Eastern chimpanzees, on the 3'.

Discussion

In this report, we present evidence that the chimpanzee MSY displays the hallmarks of selective forces reducing diversity beyond neutral expectations. Studies of the MSY in human populations have found evidence for selection acting on the X-degenerate coding sites and the ampliconic region reducing diversity (Rozen, et al. 2009; Wilson Sayres, et al. 2014). Similarly, a comparative study of Great Ape X chromosomes identified hard selective sweeps that markedly reduced diversity in regions that overlap testes-expressed ampliconic genes (Nam, et al. 2015). Our data support both targets as contributing to the reduction of MSY diversity in *Pan*. First,

for the three chimpanzee subspecies sampled here we find that MSY diversity is below one-fourth of the autosomes and consistently less than what is observed for the mitochondria. We find a striking difference in the mitochondria and MSY T_{MRCA} for the entire chimpanzee lineage. This difference is extremely unexpected under models of equal male and female effective population sizes in the chimpanzee ancestral population, but can be explained by a greatly reduced Y chromosome N_E. This could be the result of a highly skewed sex-ratio, strong selection on the Y, or a combination of factors.

We also present evidence for a significant skew of non-synonymous over synonymous mutations in X-degenerate genes (*Pan*: $p = 1.60 \times 10^{-5}$; *P. trog.* $p = 8.40 \times 10^{-5}$). While selection of structural variation on the MSY is difficult to measure, we report population-specific ampliconic CNVs containing testis-expressed genes. Differences in the environment, social structure, and mating behavior of the subspecies may have differentially selected for specific ampliconic content. Our MSY sequence and structural diversity observations coupled together, suggest that Y-linked male fertility traits are under selection in *Pan*. However, all sites on the MSY are linked. Therefore selection acting on any site will contribute to a drop in diversity across the entire sequence, which can render the untangling of target(s) difficult.

Notable divergence was detected between two bonobo MSY sequences. In fact, the π of these two sequences (9.9×10^{-4}) is approximately equal to the divergence of Eastern and Nigerian-Cameroon chimpanzees (1.08×10^{-3}). This result suggests there may be deeply rooted bonobo MSY haplotypes that if discovered could give insight to the evolution of the *Pan MSY.* This may reflect strong male-population structure in bonobos, as previously reported (Eriksson, et al. 2006), where lack of male migration may give rise to local regions where deeply diverged lineages are maintained. Furthermore, if *Pan* CNV diversity in the ampliconic region were only the result of spontaneous rearrangement, one would expect an increase of the X-degenerate T_{MRCA} to be highly predictive of structural diversity between two MSYs. In contrast, however, the structural diversity of the ampliconic region of the bonobo MSYs analyzed is limited: the two only appreciably differ in respect to the green amplicon, which harbors *BPY2.* In light of this, we suggest that although species have unique amplicon structures, selection has reduced amplicon copy-number diversity *within species*, even among highly divergent sequences such as found in the two bonobos.

Despite this general conservation, copy number of *RBMY* displayed conspicuous variability across our samples (Supplemental Table 4B). While copy number of the *RBMY* k-mers in the common chimpanzees is relatively constant by position, we observe abrupt changes in bonobo copy number at switch points within the gene (Supplemental Figure 11R). In chimpanzees, multiple copies of *RBMY* are nested within the "palindrome array" of violet and teal amplicons, which display amplification in *Pan* populations relative to Western chimpanzees and the panTro4 reference (Table 3A). These results suggest that the palindrome array is subject to structural rearrangement on the *Pan* MSY. Furthermore, retainment of the *RBMY* on the Y chromosome since marsupials (divergence from placental mammals was ~160 mya (Luo, et al. 2011) suggests that this locus has an important male-specific function (Delbridge, et al. 1997). This stands in contrast to another spermatogenetic

candidate, *DAZ*, which moved from the autosomes to the MSY exclusively in the primate lineage (Gromoll, et al. 1999). Additionally, the testis exhibit high levels of alternative splicing and *RBMY* is one of the few genes identified to activate testis-specific splicing events (Liu, et al. 2009; Yeo, et al. 2004). The *RBMY* CNVs may result in variable expression of testis-specific isoforms across *Pan* and could contribute to population specific male-fertility traits.

In contrast to *Pan*, the copy number of ampliconic genes in humans appears to be more variable as revealed by a FISH analysis using 47 Y chromosomes across a robust genealogical tree (Repping, et al. 2006). In humans, *TSPY* and *DAZ* range between 23-64 units and 2-8 units, respectively. Comparatively, we find less variability in our *Pan* dataset: the median copy number of *DAZ* and *TSPY* range 1.76 - 4.11 and 5.46 - 38.85, respectively. This trend is also evident in the *BPY2* and *CDY* genes. While our *Pan* dataset has fewer samples, it surprising to observe less copy number diversity across all ampliconic genes given the much older divergence times within the *Pan* genealogical tree compared with the human tree. *Pan's* polyandrous mating system may drive higher levels of sperm competition relative to humans, resulting in stronger pressure that repeatedly fixes advantageous alleles while maintaining a specific physiology of the testis.

The major limitations of this study are the small size of our sample population and the lack of Central chimpanzee samples. While the *Pan* and common chimpanzee T_{MRCA} estimates reported here are not affected by the absence of Central chimpanzees because the Eastern - Central node post-dates them, assaying their structural variation is vital to a complete picture of chimpanzee MSY diversity. Since they are the oldest chimpanzee subspecies, a compelling future study could determine if structural variation of the Eastern, Nigerian-Cameroon, and Western subspecies falls within the range of the Centrals or if the MSY has undergone independent rearrangements in each population since their divergence. We identified unique Western chimpanzee sequence within the non-repetitive sequence in the ampliconic region termed "other" (Supplemental Figure 4E). We also speculate that there may be subspecies-specific amplicons not assayed here. Future studies of the *Pan* MSY would benefit from high-quality MSY assemblies specific for chimpanzee subspecies and bonobos to identify patterns of gain and loss of unique sequence. Additionally, our method of assaying CNVs of ampliconic genes does not distinguish between intact and pseudogenized copies. While the relative quantity of annotated amplicons generated from short read data is reported here, assemblies would provide the spatial diversity of amplicons. With assemblies in hand, one could recreate the non-homologous recombination events that gave rise to the structural diversity within *Pan* and further clarify the contribution of ampliconic sequence to subspecies evolution. The chimpanzee subspecies appear to be in an early stage of speciation and it is likely that selection of the MSY will promote further sequence and structural diversity in the *Pan* genus as it continues down along its evolutionary path (Hey 2010).

Material and Methods

Sample selection and data processing

We analyzed 11 individuals previously sequenced by the Great Ape Genome Project with evidence of minimal sequence contamination (Prado-Martinez, et al. 2013). Selected samples included two bonobos (*Pan paniscus*: A919_Desmond, A925_Bono), and nine chimpanzees from three subspecies including four Nigerian-Cameroon chimpanzees (*Pan troglodytes ellioti*: Akwaya_Jean, Basho, Damian, Koto), two Eastern chimpanzees (*Pan troglodytes schweinfurthii*: 100037_Vincent, A910_Bwambale), two Western chimpanzees (*Pan troglodytes verus:* 9668_Bosco, Clint), and one Western/Central hybrid (9730_Donald) (Table S2). Analysis utilized reads mapped to the panTro-2.1.4 (UCSC panTro4) assembly processed as previously described. We included human sample HGDP00222 (Pathan) (Martin, et al. 2014) as an out-group. Human reads were mapped to panTro4 using bwa (Li and Durbin 2009) (version 0.5.9) with options aln –q 15 –n 0.01 and sampe -o 1000 and processed as previously described using Picard (version 1.62) and the Genome Analysis Toolkit (version 1.2-65) (McKenna, et al. 2010).

Variant calling and filtering

To identify a robust set of variants, we imposed a series of regional and site-level filters similar to the procedure utilized for the human MSY (Poznik, et al. 2013). First, we generated a regional callability mask, which defines the regions across the chromosomes that yield reliable genotype calls with short read sequence data. We calculated average filtered read depth and the MQ0 ratio (the number of reads with a mapping quality of zero divided by total read depth) in contiguous 1kb windows based on the output of the GATK Unified Genotyper ran in emit all sites mode. We computed an exponentially-weighted moving average (EWMA) across the windows and removed regions that deviated from a narrow envelope that excluded the tails of the read depth and MQ0 distributions.

We next applied a series of site-level filters to the individual positions that passed the regional masking (Supplemental Table 1, Supplemental Table 4). We excluded positions with MQ0 fraction < 0.10, that contained missing genotypes in at least one sample, or contained a site where the maximum likelihood genotype was heterozygous in at least one sample (X and Y chromosomes only). Of the remaining sequence, we plotted the distribution of site-level depths and removed positions with depths at the tail ends of the distribution. We excluded all tri-allelelic sites and further filtered variants within 5bp of indels. Separate regional callability masks and site-level filters were generated for the human and the combined chimp-bonobo samples and positions that failed either mask were dropped from the analysis. The combined call set consisted of 4,233,540 bp on chrY, 75,176,291 bp on chrX, and 106,767,230 bp on chr7.

For chimpanzee mitochondrial genomes, we used previously assembled sequences (Prado-Martinez, et al. 2013). We obtained the HGDP00222 mitochondrion sequence was from GENBANK (GI:556701605) (Benson, et al. 2015; Lippold, et al. 2014). The assembled mitochondrial genome of Basho, a Nigerian-Cameroon chimpanzee, was not available. To generate his sequence, we used GATK to call mitochondrial variants based on an alignment of his sequencing reads to the mitochondria reference. As a quality control, we also included two other Nigerian-Cameroon chimpanzees (Akwaya Jean and Damian). All three sequences were

formatted to FASTA and aligned to the previously assembled mitochondrion genomes with MAFFT (Katoh and Standley 2013). We confirmed that the genotype calls of Akwaya Jean and Damian were identical to that of the assembled genomes, therefore we included Basho's mitochondrial sequence in our analysis. After alignment, we identified 15,098bp of callable sites. For BEAST phylogenetic analyses, we further reduced the data to the 12 guanine-rich protein-coding genes located on the mitochondria genome's heavy strand. To preserve codon positions, gene sequences were manually aligned across samples to adjust indels and remove gaps. Our gene-only mitochondrial alignment consisted of 10,869bp (Drummond, et al. 2012).

Variant analysis and tree construction

Nucleotide diversity (π) was calculated as the average number of pairwise nucleotide differences per site. Divergence (d) was calculated as the average pairwise genotype differences between all chimpanzee samples and the human (HGDP00222). π and d were computed directly from the genotype calls stored in VCF files. For calculating π and d of the mitochondria, we used the Tamura-Nei substitution model implemented in MEGA 6.06 (Tamura, et al. 2013). A neighbor-joining tree (500 bootstraps) of the MSY callable sites was generated in MEGA 6.06 using the Tamura-Nei substitution model (Tamura, et al. 2013).

Simulations of autosomal diversity under a variable male-to-female gender ratio and various demographic parameters were carried out in ms using a previously described method (Hudson 2002; Wilson Sayres, et al. 2014) . The effective population sizes for each genomic compartments (N_{auto}, N_{chrX}, N_{chrY}, and N_{mito}), for given male and female effective population sizes were calculated as follows (Hartl and Clark 2007):

$$N_{auto} = 4N_mN_f/(N_m+N_f)$$

$$N_{chrX} = 9N_mN_f/(4N_m+2N_f)$$

$$N_{chrY} = N_m/2$$

$$N_{mito} = N_f/2$$

Male and female effective population sizes for a fixed ratio of male-to-females ($R = N_m/N_f$) were calculated from the total effective population size N_{auto} as:

$$Nf = N_{auto}(1+R)/4R$$

$$N_m=N_fR$$

Simplified demographic histories of chimpanzee subspecies were modeled after the split times and N_es of three ancestral populations analyzed with PSMC (Prado-Martinez, et al. 2013): *The common chimpanzee (428 kya), Nigerian-Cameroon - Western (235 kya), and Central - Eastern (175 kya)*. The demographic histories of the chimpanzee subspecies were simulated as follows: Western chimpanzees were modeled with a contemporary N_E of 5,100, a N_E of 9,400 at the time of the Nigerian-Cameroon - Western ancestral node and a N_E of 10,100 at the ancestral node of the common chimpanzee. Nigerian-Cameroon chimpanzees were modeled the identical to the Westerns except the contemporary N_E was 8,800. The Eastern chimpanzees were modeled with a contemporary N_E of 12,400, a N_E of 17,120 at the time of the Eastern - Central ancestral node, and a N_E of 10,100 at the ancestral node of the common chimpanzee. Bonobos were modeled a contemporary N_E of 5,400 and 37,483 at

the chimpanzee-bonobo ancestral node (Prado-Martinez, et al. 2013). Time was scaled for 25-year generation time.

We applied the Bayesian Markov chain Monte Carlo phylogenetics software BEAST v1.8.2 to the MSY and mitochondria sequence data for the construction of phylogenetic trees and estimating T_{MRCA}s of interest (Drummond, et al. 2012). To generate the XML input files for BEAST, we formatted our sequence into the NEXUS format and uploaded it into BEAUTi v1.8.2 (Drummond, et al. 2012). jModelTest identified the general time reversible (GTR) substitution model for the Y chromosome (Darriba, et al. 2012). For the mitochondria we applied the SRD06 model of nucleotide substitution, which partitions the third codon from positions 1 and 2 allowing differences in the Ti/Tv ratio, substitution rate, and the shape of the gamma distribution of rate heterogeneity. We utilized a lognormal relaxed molecular clock, which allows evolutionary rates to account for lineage-specific rates of heterogeneity. Since our sample set includes both intra- and inter- specific sequences, we used a two-step Yule-bootstrap/Coalescent method for a comprehensive analysis of our data (Bjork, et al. 2011). In the first step, we inferred a tree with a Yule speciation prior to a sample set that includes only one sequence per species or subspecies. Data from all samples were included in the analysis by iterating through all possible sample combinations (n = 32) and combining the results with Log Combiner. Each of the 32 bootstrapped alignments was run with a 35 million MCMC chain length (20% burnin). In the second step, we performed chimpanzee-only coalescent models with a constant growth tree prior.

For each chimpanzee-only coalescent model, two runs were performed with a 100 million chain length (10% burnin). For calibration of the Yule species tree, we applied a lognormal prior Homo-Pan T_{MRCA} distribution with a lognormal mean of zero, a standard deviation of 0.56, and offset to 5. The offset ensures that the median and mean values of the sampled distribution were approximately equal to 6. The chimpanzee coalescent trees were calibrated with the *Pan troglodytes* T_{MRCA} estimated from the Yule-prior species tree. Effective sample sizes (ESS) of all parameters in combined runs were >200 (Supplemental Table 5). We generated an estimate of the phylogenetic tree from our BEAST output with Tree Annotator v1.8.2 and visualized it with FigTree.

Selection of X-degenerate genes

We extracted X-degenerate gene sequences based on RefSeq coordinates reported in the UCSC table browser and identified 11 genes will callable sites. To calculate the dN/dS (ω) ratio of each gene, we used the PAML program, codeml (Yang 1998). In our analysis we set codons to equal frequency, no clock, with one ω for all branches. From the results, we extracted pairwise dN/dS values between each sample and the Western chimpanzee Clint and human sample HGDP00222. Clint was selected as a basis for comparison because his genome is a major component of the panTro4 reference genome. To test for a bias in the proportion of coding variants, we counted the total number of synonymous or non-synonymous substitutions and their fold-degeneracy with MEGA 6.06. Non-synonymous sites included all 0-fold generate sites and two-thirds of all 2-

fold degenerate sites. Synonymous sites included all 4-fold degenerate sites and one-third of all 2-fold degenerate sites. From this data we constructed a contingency table and performed a two-sided Fisher's exact test for a significant bias in the proportion of coding vs. non-coding variants within X-degenerate genes.

PSMC Analysis

We analyzed population split times using the pairwise sequential Markovian coalescent (PSMC) model applied to pseudo-diploid individuals constructed from X chromosome sequences obtained from males from different populations(Li and Durbin 2011). Input for PSMC was constructed by tabulating callable and variable sites in 100 bp windows. PSMC was ran with the following options: -N 25 -t 15 -r 5 -p "4+25*2+4+6". Results were plotted assuming a generation time of 25 years (Langergraber, et al. 2012) and an X chromosome mutation rate calibrated to a human-chimpanzee divergence time of 6 mya.

Mitochondria to MSY T_{MRCA} Predata Distribution

Assuming a constant populations size, $TMRCA = \sum_{i=2}^{n} T(i)$ where T_i time between coalescent events. T(i) is modeled as an exponential random variable with a rate parameter of $\lambda = \frac{n(n-1)}{2N}$. We simulated 100,000 draws of T_{MRCA} for n =2 lineages of the mitochondria and MSY and a variable male-to-female sex ratio. Distributions of the mitochondria-to-MSY ratio were generated by subtracting the MSY from the mitochondria data points matched by an arbitrary index. Additionally, T_{MRCA} distributions were scaled by a factor of N_E x g.

Amplicon Analysis

The Y chromosome sequence assembly in panTro4 differs slightly from that originally reported (Hughes, et al. 2010). We therefore determined the boundaries of amplicon and palindrome units defined by Hughes et al. using the dotplot software Gepard(Krumsiek, et al. 2007). We assessed copy-number of each amplicon family using Quick-mer, a novel pipeline for paralog-specific copy-number analysis (Shen et al, in preparation, https://github.com/KiddLab/QuicK-mer). Briefly, QuicK-mer utilizes the Jellyfish-2 (Marcais and Kingsford 2011) program to efficiently tabulate the depth of coverage of a predefined set of k-mer sequences (here, k=30) in a set of sequencing reads. The resulting k-mer counts are normalized to account for effects of local GC percentage on read depth (Alkan, et al. 2009). For this analysis, we utilized two sets of k-mers: 30 mers determined to be unique throughout the genome, and k-mers determined to be specific to each individual amplicon family. For example, since there are four copies of the “red” amplicon in panTro4, we identified 93,200 k-mers present in each copy of the red amplicon and otherwise absent from the reference.

To identify candidate k-mers specific to each amplicon family, the sequences of the amplicon family were extracted from the reference sequence and then blatted to confirm that the sequence was found at the expected copy number in the reference. Next, 30bp k-mers found across the amplicon set were tabulated using Jellyfish-2 and those k-mers present at the expected copy-number were extracted. Genome wide unique k-mer

candidates were identified based on Jellyfish analysis of the genome reference assembly. In all cases, Jellyfish-2 was ran in such a way that a k-mer and its reverse complement were considered to be identical. We applied a series of three quality control filters to remove k-mers that share identity with other sequence in the reference genome. First, we predefined a mask consisting of 15mers that are overrepresented in the genome. Any candidate 30mer intersecting with this mask was eliminated. Second, all matches against the reference genome within 2 substitutions were identified using mrsFAST (Hach, et al. 2010) and any kmer with >100 matches was eliminated. Finally, all mapping locations within an edit distance of 2, including indels, were identified using mrFAST (Alkan, et al. 2009; Xin, et al. 2013) and any kmer with >100 matches was eliminated.

Depth for genome-wide unique and amplicon-specific kmers was then determined using Quick-mer (Alkan, et al. 2009). GC normalization curves were calculated based on autosomal regions not previously identified as copy-number variable. K-mer depths were converted to copy-number estimates by dividing by the average depth of Y chromosome k-mers that passed the callable genome mask filter in the X-degenerate region on the chimpanzee Y. Prior to analysis all Y chromosome k-mers with a depth >0 in the female chimpanzee sample Julie_A959 (SAMN01920535) were removed.

Acknowledgments

This work was supported by the National Institutes of Health (grant number 1DP5OD009154). We thank Brenna Henn and Jacob Mueller for thoughtful comments on the manuscript draft.

References

Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, Sahinalp SC, Gibbs RA, Eichler EE 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet 41: 1061-1067.

Bachtrog D 2013. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. Nat Rev Genet 14: 113-124. doi: 10.1038/nrg3366

Bataillon T, Duan J, Hvilsom C, Jin X, Li Y, Skov L, Glemin S, Munch K, Jiang T, Qian Y, Hobolth A, Wang J, Mailund T, Siegismund HR, Schierup MH 2015. Inference of purifying and positive selection in three subspecies of chimpanzees (Pan troglodytes) from exome sequencing. Genome Biol Evol 7: 1122-1132. doi: 10.1093/gbe/evv058

Becquet C, Przeworski M 2007. A new approach to estimate parameters of speciation models with application to apes. Genome Res 17: 1505-1519. doi: 10.1101/gr.6409707

Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW 2015. GenBank. Nucleic Acids Res 43: D30-35. doi: 10.1093/nar/gku1216

Bjork A, Liu W, Wertheim JO, Hahn BH, Worobey M 2011. Evolutionary history of chimpanzees inferred from complete mitochondrial genomes. Mol Biol Evol 28: 615-623. doi: 10.1093/molbev/msq227

Caswell JL, Mallick S, Richter DJ, Neubauer J, Schirmer C, Gnerre S, Reich D 2008. Analysis of chimpanzee history based on genome sequence alignments. PLoS Genet 4: e1000057. doi: 10.1371/journal.pgen.1000057

Darriba D, Taboada GL, Doallo R, Posada D 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 9: 772. doi: 10.1038/nmeth.2109

Delbridge ML, Harry JL, Toder R, O'Neill RJ, Ma K, Chandley AC, Graves JA 1997. A human candidate spermatogenesis gene, RBM1, is conserved and amplified on the marsupial Y chromosome. Nat Genet 15: 131-136. doi: 10.1038/ng0297-131

Drummond AJ, Ho SY, Phillips MJ, Rambaut A 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol 4: e88. doi: 10.1371/journal.pbio.0040088

Drummond AJ, Suchard MA, Xie D, Rambaut A 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29: 1969-1973. doi: 10.1093/molbev/mss075

Dutheil JY, Munch K, Nam K, Mailund T, Schierup MH 2015. Strong Selective Sweeps on the X Chromosome in the Human-Chimpanzee Ancestor Explain Its Low Divergence. PLoS Genet 11: e1005451. doi: 10.1371/journal.pgen.1005451

Eriksson J, Siedel H, Lukas D, Kayser M, Erler A, Hashimoto C, Hohmann G, Boesch C, Vigilant L 2006. Y-chromosome analysis confirms highly sex-biased dispersal and suggests a low male effective population size in bonobos (Pan paniscus). Mol Ecol 15: 939-949. doi: 10.1111/j.1365-294X.2006.02845.x

Francalacci P, Morelli L, Angius A, Berutti R, Reinier F, Atzeni R, Pilu R, Busonero F, Maschio A, Zara I, Sanna D, Useli A, Urru MF, Marcelli M, Cusano R, Oppo M, Zoledziewska M, Pitzalis M, Deidda F, Porcu E, Poddie F, Kang HM, Lyons R, Tarrier B, Gresham JB, Li B, Tofanelli S, Alonso S, Dei M, Lai S, Mulas A, Whalen MB, Uzzau S, Jones C, Schlessinger D, Abecasis GR, Sanna S, Sidore C, Cucca F 2013. Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. Science 341: 565-569. doi: 10.1126/science.1237947

Gaggiotti OE, Excoffier L 2000. A simple method of removing the effect of a bottleneck and unequal population sizes on pairwise genetic distances. Proc Biol Sci 267: 81-87. doi: 10.1098/rspb.2000.0970

Gonder MK, Locatelli S, Ghobrial L, Mitchell MW, Kujawski JT, Lankester FJ, Stewart CB, Tishkoff SA 2011. Evidence from Cameroon reveals differences in the genetic structure and histories of chimpanzee populations. Proc Natl Acad Sci U S A 108: 4766-4771. doi: 10.1073/pnas.1015422108

Good JM, Wiebe V, Albert FW, Burbano HA, Kircher M, Green RE, Halbwax M, Andre C, Atencia R, Fischer A, Paabo S 2013. Comparative population genomics of the ejaculate in humans and the great apes. Mol Biol Evol 30: 964-976. doi: 10.1093/molbev/mst005

Graves JA 2006. Sex chromosome specialization and degeneration in mammals. Cell 124: 901-914. doi: 10.1016/j.cell.2006.02.024

Greve G, Alechine E, Pasantes JJ, Hodler C, Rietschel W, Robinson TJ, Schempp W 2011. Y-Chromosome variation in hominids: intraspecific variation is limited to the polygamous chimpanzee. PLoS One 6: e29311. doi: 10.1371/journal.pone.0029311

Gromoll J, Weinbauer GF, Skaletsky H, Schlatt S, Rocchietti-March M, Page DC, Nieschlag E 1999. The Old World monkey DAZ (Deleted in AZoospermia) gene yields insights into the evolution of the DAZ gene cluster on the human Y chromosome. Hum Mol Genet 8: 2017-2024.

Hach F, Hormozdiari F, Alkan C, Hormozdiari F, Birol I, Eichler EE, Sahinalp SC 2010. mrsFAST: a cache-oblivious algorithm for short-read mapping. Nat Methods 7: 576-577. doi: 10.1038/nmeth0810-576

Hammer MF, Mendez FL, Cox MP, Woerner AE, Wall JD 2008. Sex-biased evolutionary forces shape genomic patterns of human diversity. PLoS Genet 4: e1000202. doi: 10.1371/journal.pgen.1000202

Hammer MF, Woerner AE, Mendez FL, Watkins JC, Cox MP, Wall JD 2010. The ratio of human X chromosome to autosome diversity is positively correlated with genetic distance from genes. Nat Genet 42: 830-831. doi: 10.1038/ng.651

Hartl DL, Clark AG. 2007. Principles of population genetics. Sunderland, Mass.: Sinauer Associates.

Helena Mangs A, Morris BJ 2007. The Human Pseudoautosomal Region (PAR): Origin, Function and Future. Curr Genomics 8: 129-136.

Hey J 2010. The divergence of chimpanzee species and subspecies as revealed in multipopulation isolation-with-migration analyses. Mol Biol Evol 27: 921-933. doi: 10.1093/molbev/msp298

Hudson RR 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337-338.

Hughes JF, Skaletsky H, Page DC 2012. Sequencing of rhesus macaque Y chromosome clarifies origins and evolution of the DAZ (Deleted in AZoospermia) genes. Bioessays 34: 1035-1044. doi: 10.1002/bies.201200066

Hughes JF, Skaletsky H, Pyntikova T, Graves TA, van Daalen SK, Minx PJ, Fulton RS, McGrath SD, Locke DP, Friedman C, Trask BJ, Mardis ER, Warren WC, Repping S, Rozen S, Wilson RK, Page DC 2010. Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. Nature 463: 536-539. doi: 10.1038/nature08700

Hvilsom C, Qian Y, Bataillon T, Li Y, Mailund T, Salle B, Carlsen F, Li R, Zheng H, Jiang T, Jiang H, Jin X, Munch K, Hobolth A, Siegismund HR, Wang J, Schierup MH 2012. Extensive X-linked adaptive evolution in central chimpanzees. Proc Natl Acad Sci U S A 109: 2054-2059. doi: 10.1073/pnas.1106877109

Jobling MA, Tyler-Smith C 2003. The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet 4: 598-612. doi: 10.1038/nrg1124

Karmin M, Saag L, Vicente M, Wilson Sayres MA, Jarve M, Talas UG, Rootsi S, Ilumae AM, Magi R, Mitt M, Pagani L, Puurand T, Faltyskova Z, Clemente F, Cardona A, Metspalu E, Sahakyan H, Yunusbayev B, Hudjashov G, DeGiorgio M, Loogvali EL, Eichstaedt C, Eelmets M, Chaubey G, Tambets K, Litvinov S, Mormina M, Xue Y, Ayub Q, Zoraqi G, Korneliussen TS, Akhatova F, Lachance J, Tishkoff S, Momynaliev K, Ricaut FX, Kusuma P, Razafindrazaka H, Pierron D, Cox MP, Sultana GN, Willerslev R, Muller C, Westaway M, Lambert D, Skaro V, Kovacevic L, Turdikulova S, Dalimova D, Khusainova R, Trofimova N, Akhmetova V, Khidiyatova I, Lichman DV, Isakova J, Pocheshkhova E, Sabitov Z, Barashkov NA, Nymadawa P, Mihailov E, Seng JW, Evseeva I, Migliano AB, Abdullah S, Andriadze G, Primorac D, Atramentova L, Utevska O, Yepiskoposyan L, Marjanovic D, Kushniarevich A, Behar DM, Gilissen C, Vissers L, Veltman JA, Balanovska E, Derenko M, Malyarchuk B, Metspalu A, Fedorova S, Eriksson A, Manica A, Mendez FL, Karafet TM, Veeramah KR, Bradman N, Hammer MF, Osipova LP, Balanovsky O, Khusnutdinova EK, Johnsen K, Remm M, Thomas MG, Tyler-Smith C, Underhill PA, Willerslev E, Nielsen R, Metspalu M, Villems R, Kivisild T 2015. A recent bottleneck of Y chromosome diversity coincides with a global change in culture. Genome Res 25: 459-466. doi: 10.1101/gr.186684.114

Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, Kent WJ, University of California Santa C 2003. The UCSC Genome Browser Database. Nucleic Acids Res 31: 51-54.

Katoh K, Standley DM 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30: 772-780. doi: 10.1093/molbev/mst010

Keinan A, Mullikin JC, Patterson N, Reich D 2009. Accelerated genetic drift on chromosome X during the human dispersal out of Africa. Nat Genet 41: 66-70. doi: 10.1038/ng.303

Krumsiek J, Arnold R, Rattei T 2007. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics 23: 1026-1028. doi: 10.1093/bioinformatics/btm039

Kuroda-Kawaguchi T, Skaletsky H, Brown LG, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Silber S, Oates R, Rozen S, Page DC 2001. The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. Nat Genet 29: 279-286. doi: 10.1038/ng757

Lahn BT, Page DC 1997. Functional coherence of the human Y chromosome. Science 278: 675-680.

Lahn BT, Pearson NM, Jegalian K 2001. The human Y chromosome, in the light of evolution. Nat Rev Genet 2: 207-216. doi: 10.1038/35056058

Langergraber KE, Prufer K, Rowney C, Boesch C, Crockford C, Fawcett K, Inoue E, Inoue-Muruyama M, Mitani JC, Muller MN, Robbins MM, Schubert G, Stoinski TS, Viola B, Watts D, Wittig RM, Wrangham RW, Zuberbuhler K, Paabo S, Vigilant L 2012. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. Proc Natl Acad Sci U S A 109: 15716-15721. doi: 10.1073/pnas.1211740109

Li H, Durbin R 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760. doi: btp324 [pii]
10.1093/bioinformatics/btp324

Li H, Durbin R 2011. Inference of human population history from individual whole-genome sequences. Nature 475: 493-496. doi: nature10231 [pii]
10.1038/nature10231

Lippold S, Xu H, Ko A, Li M, Renaud G, Butthof A, Schroder R, Stoneking M 2014. Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences. Investig Genet 5: 13. doi: 10.1186/2041-2223-5-13

Liu Y, Bourgeois CF, Pang S, Kudla M, Dreumont N, Kister L, Sun YH, Stevenin J, Elliott DJ 2009. The germ cell nuclear proteins hnRNP G-T and RBMY activate a testis-specific exon. PLoS Genet 5: e1000707. doi: 10.1371/journal.pgen.1000707

Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth LV, Muzny DM, Yang SP, Wang Z, Chinwalla AT, Minx P, Mitreva M, Cook L, Delehaunty KD, Fronick C, Schmidt H, Fulton LA, Fulton RS, Nelson JO, Magrini

V, Pohl C, Graves TA, Markovic C, Cree A, Dinh HH, Hume J, Kovar CL, Fowler GR, Lunter G, Meader S, Heger A, Ponting CP, Marques-Bonet T, Alkan C, Chen L, Cheng Z, Kidd JM, Eichler EE, White S, Searle S, Vilella AJ, Chen Y, Flicek P, Ma J, Raney B, Suh B, Burhans R, Herrero J, Haussler D, Faria R, Fernando O, Darre F, Farre D, Gazave E, Oliva M, Navarro A, Roberto R, Capozzi O, Archidiacono N, Della Valle G, Purgato S, Rocchi M, Konkel MK, Walker JA, Ullmer B, Batzer MA, Smit AF, Hubley R, Casola C, Schrider DR, Hahn MW, Quesada V, Puente XS, Ordonez GR, Lopez-Otin C, Vinar T, Brejova B, Ratan A, Harris RS, Miller W, Kosiol C, Lawson HA, Taliwal V, Martins AL, Siepel A, Roychoudhury A, Ma X, Degenhardt J, Bustamante CD, Gutenkunst RN, Mailund T, Dutheil JY, Hobolth A, Schierup MH, Ryder OA, Yoshinaga Y, de Jong PJ, Weinstock GM, Rogers J, Mardis ER, Gibbs RA, Wilson RK 2011. Comparative and demographic analysis of orang-utan genomes. Nature 469: 529-533. doi: 10.1038/nature09687

Luo ZX, Yuan CX, Meng QJ, Ji Q 2011. A Jurassic eutherian mammal and divergence of marsupials and placentals. Nature 476: 442-445. doi: 10.1038/nature10291

Malaspina P, Persichetti F, Novelletto A, Iodice C, Terrenato L, Wolfe J, Ferraro M, Prantera G 1990. The human Y chromosome shows a low level of DNA polymorphism. Ann Hum Genet 54: 297-305.

Marcais G, Kingsford C 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27: 764-770. doi: 10.1093/bioinformatics/btr011

Martin AR, Costa HA, Lappalainen T, Henn BM, Kidd JM, Yee MC, Grubert F, Cann HM, Snyder M, Montgomery SB, Bustamante CD 2014. Transcriptome sequencing from diverse human populations reveals differentiated regulatory architecture. PLoS Genet 10: e1004549. doi: 10.1371/journal.pgen.1004549 PGENETICS-D-13-02851 [pii]

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297-1303. doi: gr.107524.110 [pii] 10.1101/gr.107524.110

Muller HJ 1964. The Relation of Recombination to Mutational Advance. Mutat Res 106: 2-9.

Nam K, Munch K, Hobolth A, Dutheil JY, Veeramah KR, Woerner AE, Hammer MF, Great Ape Genome Diversity P, Mailund T, Schierup MH 2015. Extreme selective sweeps independently targeted the X chromosomes of the great apes. Proc Natl Acad Sci U S A 112: 6413-6418. doi: 10.1073/pnas.1419306112

Pamilo P, O'Neill RJ 1997. Evolution of the Sry genes. Mol Biol Evol 14: 49-55.

Pool JE, Nielsen R 2007. Population size changes reshape genomic patterns of diversity. Evolution 61: 3001-3006. doi: 10.1111/j.1558-5646.2007.00238.x

Poznik GD, Henn BM, Yee MC, Sliwerska E, Euskirchen GM, Lin AA, Snyder M, Quintana-Murci L, Kidd JM, Underhill PA, Bustamante CD 2013. Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. Science 341: 562-565. doi: 341/6145/562 [pii] 10.1126/science.1237619

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, Cagan A, Theunert C, Casals F, Laayouni H, Munch K, Hobolth A, Halager AE, Malig M, Hernandez-Rodriguez J, Hernando-Herraez I, Prufer K, Pybus M, Johnstone L, Lachmann M, Alkan C, Twigg D, Petit N, Baker C, Hormozdiari F, Fernandez-Callejo M, Dabad M, Wilson ML, Stevison L, Camprubi C, Carvalho T, Ruiz-Herrera A, Vives L, Mele M, Abello T, Kondova I, Bontrop RE, Pusey A, Lankester F, Kiyang JA, Bergl RA, Lonsdorf E, Myers S, Ventura M, Gagneux P, Comas D, Siegismund H, Blanc J, Agueda-Calpena L, Gut M, Fulton L, Tishkoff SA, Mullikin JC, Wilson RK, Gut IG, Gonder MK, Ryder OA, Hahn BH, Navarro A, Akey JM, Bertranpetit J, Reich D, Mailund T, Schierup MH, Hvilsom C, Andres AM, Wall JD, Bustamante CD, Hammer MF, Eichler EE, Marques-Bonet T 2013. Great ape genetic diversity and population history. Nature 499: 471-475. doi: nature12228 [pii] 10.1038/nature12228

Pybus OG 2006. Model selection and the molecular clock. PLoS Biol 4: e151. doi: 10.1371/journal.pbio.0040151

Reijo R, Lee TY, Salo P, Alagappan R, Brown LG, Rosenberg M, Rozen S, Jaffe T, Straus D, Hovatta O, et al. 1995. Diverse spermatogenic defects in humans caused by Y chromosome deletions encompassing a novel RNA-binding protein gene. Nat Genet 10: 383-393. doi: 10.1038/ng0895-383

Repping S, van Daalen SK, Brown LG, Korver CM, Lange J, Marszalek JD, Pyntikova T, van der Veen F, Skaletsky H, Page DC, Rozen S 2006. High mutation rates have driven extensive structural polymorphism among human Y chromosomes. Nat Genet 38: 463-467. doi: 10.1038/ng1754

Rozen S, Marszalek JD, Alagappan RK, Skaletsky H, Page DC 2009. Remarkably little variation in proteins encoded by the Y chromosome's single-copy genes, implying effective purifying selection. Am J Hum Genet 85: 923-928. doi: 10.1016/j.ajhg.2009.11.011

Rozen SG, Marszalek JD, Irenze K, Skaletsky H, Brown LG, Oates RD, Silber SJ, Ardlie K, Page DC 2012. AZFc deletions and spermatogenic failure: a population-based survey of 20,000 Y chromosomes. Am J Hum Genet 91: 890-896. doi: 10.1016/j.ajhg.2012.09.003

Scally A, Durbin R 2012. Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet 13: 745-753. doi: 10.1038/nrg3295

Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, McCarthy S, Montgomery SH, Schwalie PC, Tang YA, Ward MC, Xue Y, Yngvadottir B, Alkan C, Andersen LN, Ayub Q, Ball EV, Beal K, Bradley BJ, Chen Y, Clee CM, Fitzgerald S, Graves TA, Gu Y, Heath P, Heger A, Karakoc E, Kolb-Kokocinski A, Laird GK, Lunter G, Meader S, Mort M, Mullikin JC, Munch K, O'Connor TD, Phillips AD, Prado-Martinez J, Rogers AS, Sajjadian S, Schmidt D, Shaw K, Simpson JT, Stenson PD, Turner DJ, Vigilant L, Vilella AJ, Whitener W, Zhu B, Cooper DN, de Jong P, Dermitzakis ET, Eichler EE, Flicek P, Goldman N, Mundy NI, Ning Z, Odom DT, Ponting CP, Quail MA, Ryder OA, Searle SM, Warren WC, Wilson RK, Schierup MH, Rogers J, Tyler-Smith C, Durbin R 2012. Insights into hominid evolution from the gorilla genome sequence. Nature 483: 169-175. doi: 10.1038/nature10842

Schaller F, Fernandes AM, Hodler C, Munch C, Pasantes JJ, Rietschel W, Schempp W 2010. Y chromosomal variation tracks the evolution of mating systems in chimpanzee and bonobo. PLoS One 5. doi: 10.1371/journal.pone.0012482

Sinclair AH, Berta P, Palmer MS, Hawkins JR, Griffiths BL, Smith MJ, Foster JW, Frischauf AM, Lovell-Badge R, Goodfellow PN 1990. A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. Nature 346: 240-244. doi: 10.1038/346240a0

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, Chinwalla A, Delehaunty A, Delehaunty K, Du H, Fewell G, Fulton L, Fulton R, Graves T, Hou SF, Latrielle P, Leonard S, Mardis E, Maupin R, McPherson J, Miner T, Nash W, Nguyen C, Ozersky P, Pepin K, Rock S, Rohlfing T, Scott K, Schultz B, Strong C, Tin-Wollam A, Yang SP, Waterston RH, Wilson RK, Rozen S, Page DC 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature 423: 825-837. doi: 10.1038/nature01722

Soh YQ, Alfoldi J, Pyntikova T, Brown LG, Graves T, Minx PJ, Fulton RS, Kremitzki C, Koutseva N, Mueller JL, Rozen S, Hughes JF, Owens E, Womack JE, Murphy WJ, Cao Q, de Jong P, Warren WC, Wilson RK, Skaletsky H, Page DC 2014. Sequencing the mouse Y chromosome reveals convergent gene acquisition and amplification on both sex chromosomes. Cell 159: 800-813. doi: 10.1016/j.cell.2014.09.052

Stone AC, Battistuzzi FU, Kubatko LS, Perry GH, Jr., Trudeau E, Lin H, Kumar S 2010. More reliable estimates of divergence times in Pan using complete mtDNA sequences and accounting for population structure. Philos Trans R Soc Lond B Biol Sci 365: 3277-3288. doi: 10.1098/rstb.2010.0096

Stone AC, Griffiths RC, Zegura SL, Hammer MF 2002. High levels of Y-chromosome nucleotide diversity in the genus Pan. Proc Natl Acad Sci U S A 99: 43-48. doi: 10.1073/pnas.012364999

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol 30: 2725-2729. doi: 10.1093/molbev/mst197

Venn O, Turner I, Mathieson I, de Groot N, Bontrop R, McVean G 2014. Nonhuman genetics. Strong male bias drives germline mutation in chimpanzees. Science 344: 1272-1275. doi: 10.1126/science.344.6189.1272

Warren WC, Hillier LW, Marshall Graves JA, Birney E, Ponting CP, Grutzner F, Belov K, Miller W, Clarke L, Chinwalla AT, Yang SP, Heger A, Locke DP, Miethke P, Waters PD, Veyrunes F, Fulton L, Fulton B, Graves T, Wallis J, Puente XS, Lopez-Otin C, Ordonez GR, Eichler EE, Chen L, Cheng Z, Deakin JE, Alsop A, Thompson K, Kirby P, Papenfuss AT, Wakefield MJ, Olender T, Lancet D, Huttley GA, Smit AF, Pask A, Temple-Smith P, Batzer MA, Walker JA, Konkel MK, Harris RS, Whittington CM, Wong ES, Gemmell NJ, Buschiazzo E, Vargas Jentzsch IM, Merkel A, Schmitz J, Zemann A, Churakov G, Kriegs JO, Brosius J, Murchison EP, Sachidanandam R, Smith C, Hannon GJ, Tsend-Ayush E, McMillan D, Attenborough R, Rens W, Ferguson-Smith M, Lefevre CM, Sharp JA, Nicholas KR, Ray DA, Kube M, Reinhardt R, Pringle TH, Taylor J, Jones RC, Nixon B, Dacheux JL, Niwa H, Sekita Y, Huang X, Stark A, Kheradpour P, Kellis M, Flicek P, Chen Y, Webber C, Hardison R, Nelson J, Hallsworth-Pepin K, Delehaunty K, Markovic C, Minx P, Feng Y, Kremitzki C, Mitreva M, Glasscock J, Wylie T, Wohldmann P, Thiru P, Nhan MN, Pohl CS, Smith SM, Hou S, Nefedov M, de Jong PJ, Renfree MB, Mardis ER, Wilson RK 2008. Genome analysis of the platypus reveals unique signatures of evolution. Nature 453: 175-183. doi: 10.1038/nature06936

Wegmann D, Excoffier L 2010. Bayesian inference of the demographic history of chimpanzees. Mol Biol Evol 27: 1425-1435. doi: 10.1093/molbev/msq028

Wei W, Ayub Q, Chen Y, McCarthy S, Hou Y, Carbone I, Xue Y, Tyler-Smith C 2013. A calibrated human Y-chromosomal phylogeny based on resequencing. Genome Res 23: 388-395. doi: 10.1101/gr.143198.112

Wilson Sayres MA, Lohmueller KE, Nielsen R 2014. Natural selection reduced diversity on human y chromosomes. PLoS Genet 10: e1004064. doi: 10.1371/journal.pgen.1004064

Won YJ, Hey J 2005. Divergence population genetics of chimpanzees. Mol Biol Evol 22: 297-307. doi: 10.1093/molbev/msi017

Xin H, Lee D, Hormozdiari F, Yedkar S, Mutlu O, Alkan C 2013. Accelerating read mapping with FastHASH. BMC Genomics 14 Suppl 1: S13. doi: 10.1186/1471-2164-14-S1-S13

Yang Z 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15: 568-573.

Yeo G, Holste D, Kreiman G, Burge CB 2004. Variation in alternative splicing across human tissues. Genome Biol 5: R74. doi: 10.1186/gb-2004-5-10-r74

Figure Legends

Figure 1 Pan and Human Y chromosome Callability Masks

Exponentially weighted moving averages (EWMA) of read depth (blue line) and the mq0/unfiltered depth ratio (pink line) are plotted against position on the Y chromosome (PanTro4). Dashed lines represent maximum and minimum thresholds for filtered depth (green) and a maximum threshold for the mq0 ratio (red). Colored bars below plot indicate regions masked by the depth filter (blue), masked by the mq0 ratio filter (pink), excluded from the analysis (grey), and included (black) for further filtering steps. The two panels A. Chimp and bonobo callability mask. B. Human (HGDP00222) illustrate regional filtering prior to merging the two datasets.

Figure 2 Inferred Y chromosome Tree

The phylogenetic relationship among sequenced *Pan* and human Y chromosomes is illustrated with neighbor joining tree using a Tamura-Nei substitution model. The size of the color bars to the right of the tree indicate relative median copy number of each amplicons estimated per sample. Note: the order of the color bars does not reflect the spatial distribution of amplicons.

Figure 3 Autosome Corrected Diversity by Genomic Region

Y/Autosome, X/Autosome, and Mitochondria/Autosome diversity are by plotted by group. Autosome corrected human diversity values from the literature are included as purple bars (Wilson Sayres, et al. 2014). Expected values based on equal numbers of male and females Y/Autosome (0.25), Mitochondria/Autosome (0.25) and the X Chromosome (0.75) are drawn as solid black lines. To adjust for chromosome-specific mutation rates, diversity values were normalized by divergence from human.

Figure 4 Estimated Amplicon Copy Number

Amplicon and *RMBY* copy number are plotted against k-mer position. Each line represents a single sample colored by species/subspecies and smoothed with lowess function. Copy number of the PanTro4 reference is drawn with a solid black line. *A. Red Amplicon B. RBMY C. Yellow Amplicon D. Purple Amplicon.*

Tables

Table 1 Nucleotide diversity (π) and divergence of Pan populations by chromosome

Chromosome	Group	π	Eastern	Western	Nigerian	Bonobo	Divergence (from human)
ChrY	Eastern	0.00023	-	0.00106	0.00108	0.00385	0.01480
	Western	0.00002		-	0.00093	0.00388	0.01481
	N.C.+	0.00020			-	0.00390	0.01484
	Bonobo	0.00099				-	0.01486
ChrX	Eastern	0.00098	-	0.00112	0.00105	0.00193	0.00853
	Western	0.00042		-	0.00092	0.00187	0.00856
	N.C.	0.00077			-	0.00170	0.00855
	Bonobo	0.00052				-	0.00857
Autosome (Chr7)	Eastern	0.00155	-	0.00176	0.00164	0.00303	0.01194
	Western	0.00069		-	0.00144	0.00288	0.01194
	N.C.	0.00119			-	0.00264	0.01195
	Bonobo	0.00074				-	0.01199
Mitochondria*	Eastern	0.00326	-	0.01881	0.01871	0.03665	0.09313
	Western	0.00086		-	0.00922	0.03630	0.09528
	N.C.	0.00244			-	0.03687	0.09450
	Bonobo	0.00743				-	0.09355

**Mitochondria diversity was calculated with Tamura-Nei substitution model*
+Nigerian-Cameroon

Table 2 Pan and P. troglodytes T_{MRCA} estimates at nodes of interest

	Bjork et al. 2011 (Mitochondria)		**Mitochondria (This Study)**		**MSY (This Study)**	
Node	**Yule**	**Coalescent**	**Yule**	**Coalescent**	**Yule**	**Coalescent**
Pan - Homo	5.758 (5.216-6.367)	-	6.041 (5.2046-7.2123)	-	6.041 (5.207-7.211)	-
Pan	2.187 (1.621-2.663)	-	2.065 (1.4573-2.8565)	-	1.569 (1.243-1.996)	-
P. trog.	1.041 (0.770-1.288)	1.002 (0.734-1.269)	0.967 (0.653-1.348)	0.912 (0.477-1.351)	0.438 (0.338-0.561)	0.423 (0.313-0.542)
N.C.[+] - Western	0.518 (0.340-0.679)	0.508 (0.301-0.715)	0.457 (0.285-0.656)	0.455 (0.195-0.719)	0.368 (0.279-0.474)	0.371 (0.211-0.524)
N.C.[+]	-	0.157 (0.083-0.242)	-	0.1255 (0.05-0.2045)	-	0.150 (0.062-0.249)
Western	-	0.148 (0.076-0.223)	-	0.0317 (0.00773-0.0611)	-	0.012 (0.004-0.021)
Eastern	-	0.116 (0.066-0.171)	-	0.144 (0.053-0.242)	-	0.093 (0.037 - 0.158)

[+]Nigerian-Cameroon

Table 3 Copy-number of amplicons and ampliconic genes by Pan population A. Amplicons

	Hughes et al. 2010 (Human)	**Hughes et al. 2010 (Chimpanzee)**	**Clint**	**Human**	**Eastern Average***	**Western Average***	**Nigerian-Cameroon Average***	**Bonobo Average***
aqua arms	0	6	5.49	0	5.9	5.53	13.43	10.16
aqua spacer	0	3	3.22	0	2.91	3.11	6.93	6.3
blue arms	4	4	3.84	2.66	1.98	3.85	3.86	2.2
green arms	3	2	1.87	3.43	1.04	1.91	1.93	1.72
orange arms	0	3	2.7	0	3.16	2.77	3.73	1.05
other			0.93	0	0.93	0.92	0.92	0.94
pink arms	0	4	3.86	0.95	5.99	3.81	11.87	13.47
purple arms	0	3	2.82	0	0	2.78	4.74	0
red arms	4	4	3.92	4.39	2.01	3.89	3.92	2.26
red spacer	2	2	1.87	1.52	1.06	1.92	1.96	1.09
teal arms	2	7	6.68	0.95	9.66	5.46	12.4	11.13
teal spacer	0	3	2.84	0	4.89	2.21	5.71	4.76
violet arms	0	13	11.54	1.14	20.08	10.63	28.57	26.44
violet spacer	0	7	6.84	0	10.29	6.22	14.72	13.52
yellow arms	2	5	4.87	2.65	5.02	4.9	5.08	3.34
yellow spacer	0	2	1.83	2.3	2.09	1.95	1.95	1.04
yellow spacer	0	3	2.69	2.72	3.1	2.72	3.15	2.1

*Average Median Value

+We include an additional purple amplicon mapped in the panTro4 assembly

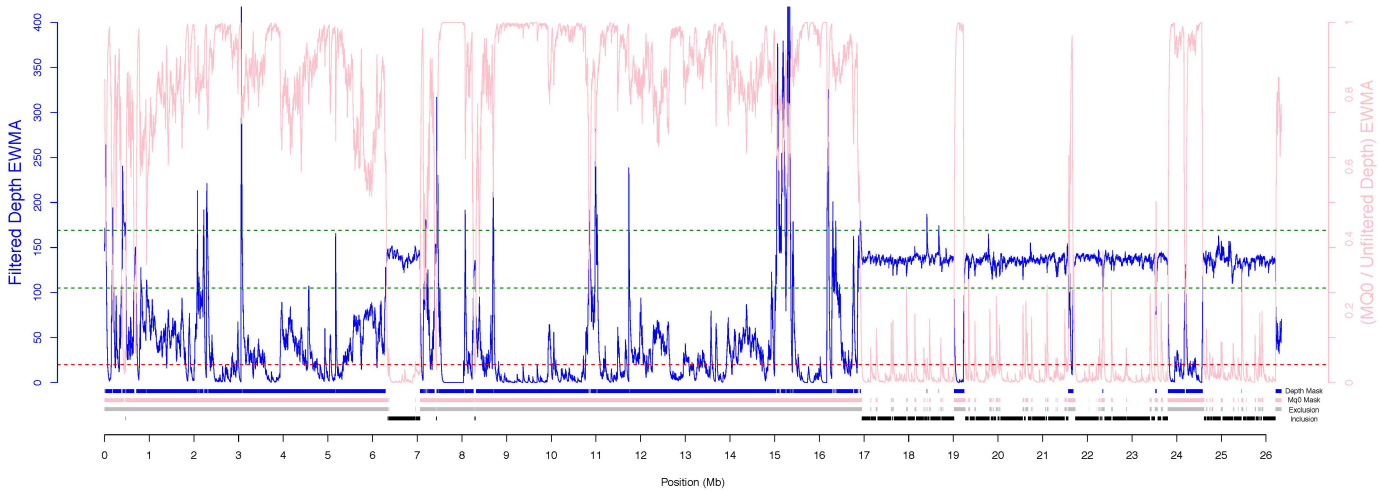
B. Ampliconic genes

	Hughes et al. 2010 (Human)	**Hughes et al. 2010 (Chimpanzee)**	**Clint**	**Human**	**Eastern (Average)**	**Western (Average)**	**Nigerian-Cameroon (Average)**	**Bonobo (Average)**
BPY2	3	2	1.81	3.61	1.1	1.86	2.07	1.79
CDY	4	5	3.98	4	5.86	4.51	5.41	3.59
DAZ	4	4	4.37	0.19	1.76	4.11	3.85	2.26
RBMY	6	6	10.91	10.32	14.34	9.16	24.12	32.28
TSPY	35	6	5.46	38.85	16.6	10.27	26.35	32.1
VCY	2	2	0.77	0	0.44	1.01	1.41	0.05

UCSC Genome Browser Gene Identifier: *CDY* (NM_001145039), *DAZ* (ENSPTRG00000022523), *BPY2* (NM_004678.2), *RMBY (*NM_001006121.2), *TSPY* (NM_001077697.2), *VCY* (NM_004679.2)

Figure 1

A.


Filtered Depth EWMA
(MQ0 / Unfiltered Depth) EWMA
Depth Mask
Mq0 Mask
Exclusion
Inclusion
Position (Mb)

B.

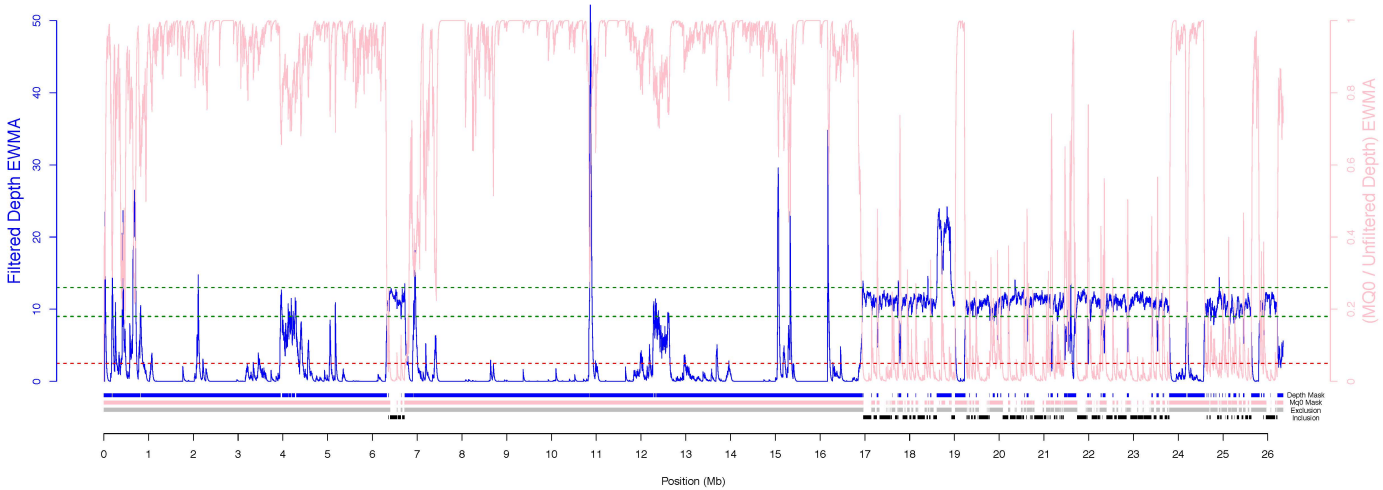
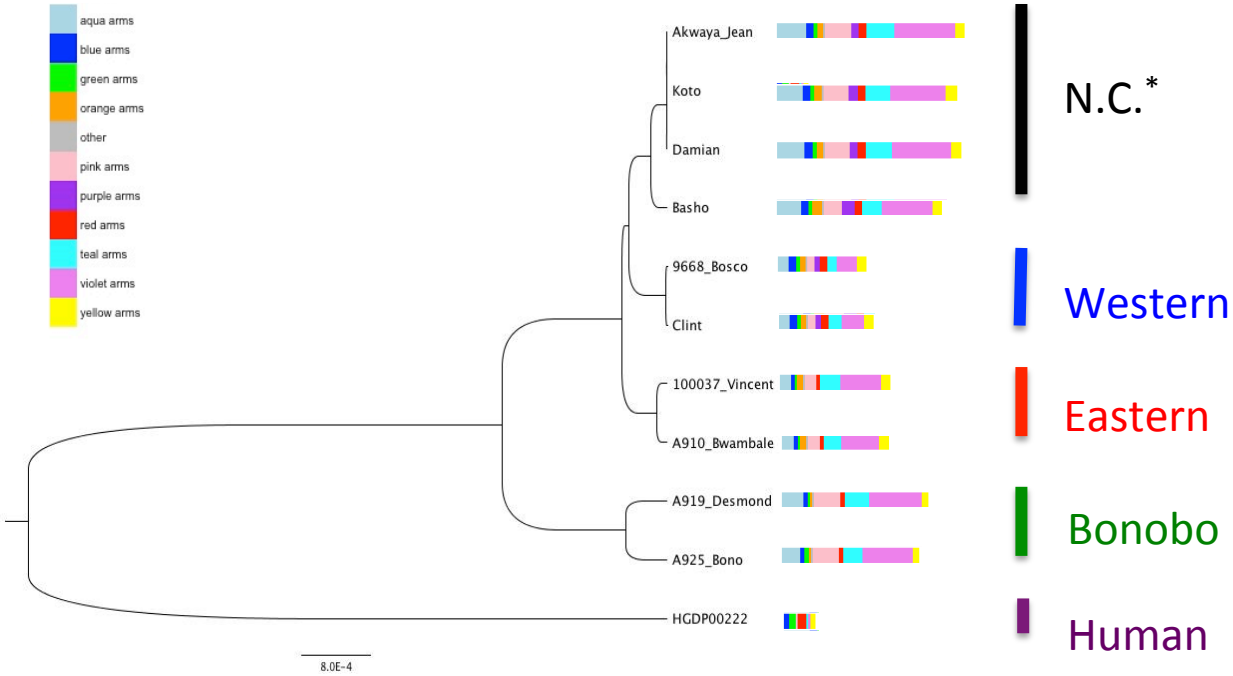

Filtered Depth EWMA
(MQ0 / Unfiltered Depth) EWMA
Depth Mask
Mq0 Mask
Exclusion
Inclusion
Position (Mb)

Figure 2



*N.C. Nigerian-Cameroon

Figure 3

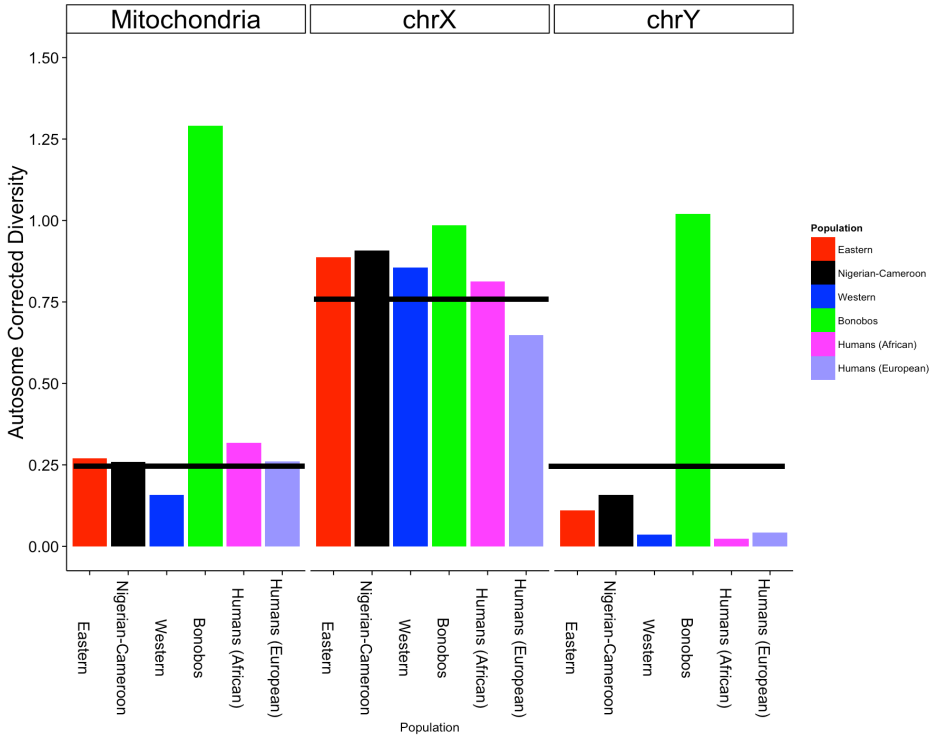
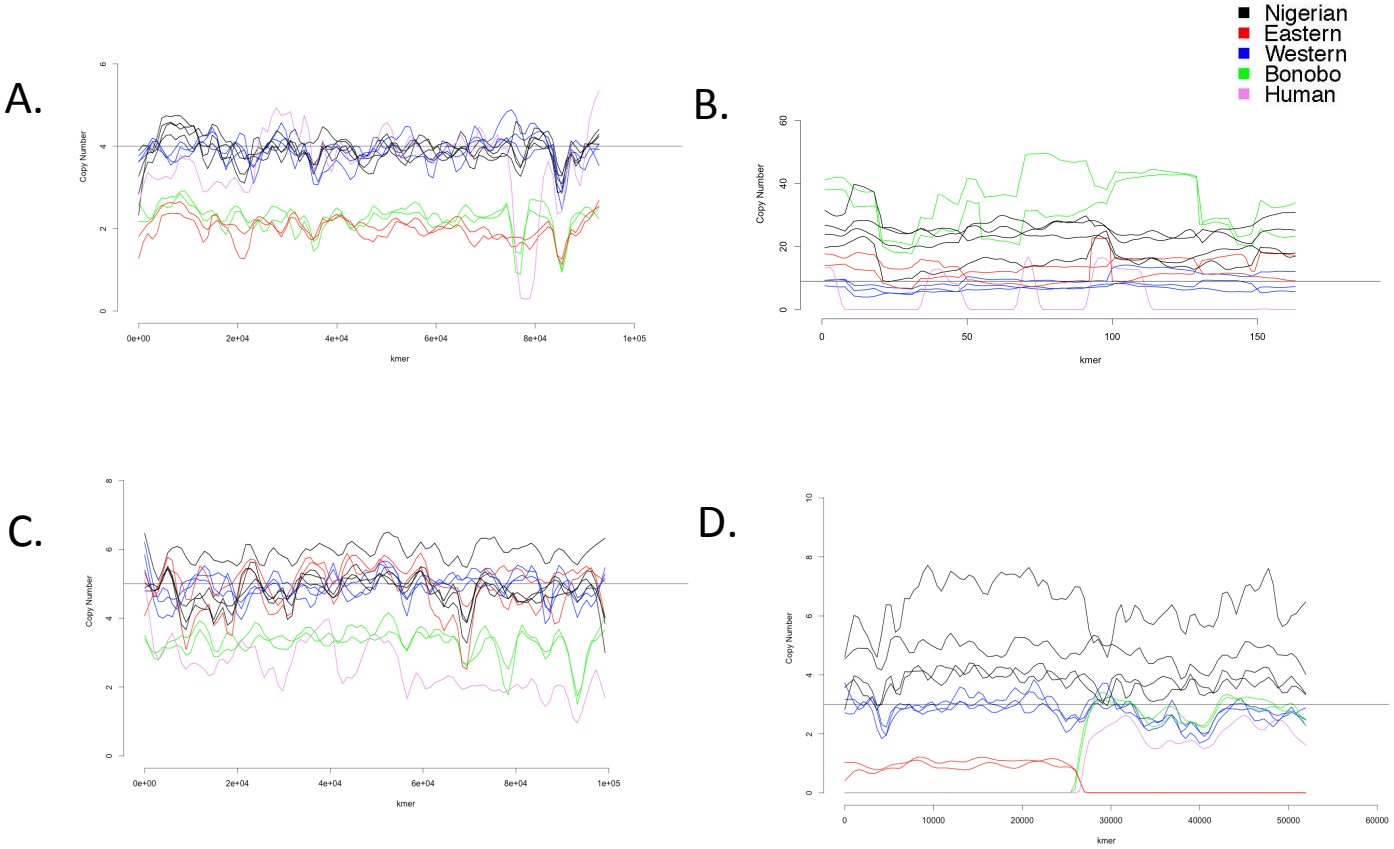

Mitochondria
chrX
chrY
Autosome Corrected Diversity
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Eastern
Nigerian-Cameroon
Western
Bonobos
Humans (African)
Humans (European)
Population
Population
Eastern
Nigerian-Cameroon
Western
Bonobos
Humans (African)
Humans (European)

Figure 4


A.
B.
C.
D.
Nigerian
Eastern
Western
Bonobo
Human
Copy Number
kmer