

Evidence for Increased Levels of Positive and Negative Selection on the X Chromosome versus Autosomes in Humans

Krishna R. Veeramah,^{1,2} Ryan N. Gutenkunst,³ August E. Woerner,¹ Joseph C. Watkins,⁴ and Michael F. Hammer^{*1}

¹Arizona Research Laboratories Division of Biotechnology, University of Arizona

²Department of Ecology and Evolution, Stony Brook University

³Department of Molecular and Cellular Biology, University of Arizona

⁴Department of Mathematics, University of Arizona

***Corresponding author:** E-mail: mfh@email.arizona.edu.

Associate editor: Sohini Ramachandran

Abstract

Partially recessive variants under positive selection are expected to go to fixation more quickly on the X chromosome as a result of hemizygosity, an effect known as faster-X. Conversely, purifying selection is expected to reduce substitution rates more effectively on the X chromosome. Previous work in humans contrasted divergence on the autosomes and X chromosome, with results tending to support the faster-X effect. However, no study has yet incorporated both divergence and polymorphism to quantify the effects of both purifying and positive selection, which are opposing forces with respect to divergence. In this study, we develop a framework that integrates previously developed theory addressing differential rates of X and autosomal evolution with methods that jointly estimate the level of purifying and positive selection via modeling of the distribution of fitness effects (DFE). We then utilize this framework to estimate the proportion of nonsynonymous substitutions fixed by positive selection (α) using exome sequence data from a West African population. We find that varying the female to male breeding ratio (β) has minimal impact on the DFE for the X chromosome, especially when compared with the effect of varying the dominance coefficient of deleterious alleles (h). Estimates of α range from 46% to 51% and from 4% to 24% for the X chromosome and autosomes, respectively. While dependent on h , the magnitude of the difference between α values estimated for these two systems is highly statistically significant over a range of biologically realistic parameter values, suggesting faster-X has been operating in humans.

Key words: X chromosome, humans, distribution of fitness effects, autosomes, purifying selection, positive selection.

Introduction

Understanding the relative roles purifying and positive selection play in shaping the genetic diversity among species is an area of considerable interest for evolutionary and population geneticists. However, measuring the effects of natural selection on genetic variation experimentally in natural populations has been challenging, especially in vertebrates. One approach for quantifying the impact of positive selection is to compare levels of variation within and between species at nonsynonymous versus synonymous sites (Fay and Wu 2003). In addition, contrasting patterns of diversity found on the X chromosome and autosomes offers the possibility of revealing important aspects of the mechanism of natural selection. Charlesworth et al. (1987) demonstrated in a seminal paper that under certain circumstances (i.e., a dominance coefficient for beneficial alleles of < 0.5) the X chromosome is expected to acquire adaptive substitutions at an increased rate compared with the autosomes, a concept that is commonly known as the faster-X effect (Meisel and Connallon 2013). Conversely, deleterious mutations with a dominance coefficient $h < 0.5$ are expected to fix on the X chromosome at a slower rate (note: The term h used here is distinguished from the dominance coefficient for beneficial alleles, $h+$). The

underlying basis of this concept is that the X chromosome is hemizygous, and so recessive genetic variants are exposed to selection in males at a level equivalent to that of homozygous genotypes on the autosomes or on the X chromosome in females. Vicoso and Charlesworth (2009) further extended the conditions of this theory to show that when the female to male breeding ratio (β) is greater than one, as would occur in a polygynous population where males have greater variance in reproductive success than females, the faster-X effect is expected even when $h+ \geq 0.5$. It was proposed that this effect could be countered by a greater male to female mutation rate ($\nu > 1$) (Kirkpatrick and Hall 2004); however, when the rate of substitution at selected sites is normalized by the rate of substitution at neutral sites, this effect is essentially nullified (Vicoso and Charlesworth 2009).

In humans, studies of the faster-X effect have been limited to examining the ratio of the number of nonsynonymous and synonymous substitutions per site (K_a/K_s or d_N/d_S) (Lu and Wu 2005; Nielsen et al. 2005; Mank et al. 2010). Although these studies appear to show increased K_a/K_s rates on the X chromosome, the relative influence of the opposing forces of purifying and positive selection has not been explored in depth. Lu and Wu (2005) attempted to fit patterns of

synonymous divergence on the autosomes and X chromosome to a model with assumed values of $\gamma = N_e s$, h , $h +$, and β and suggested that approximately 90% of sites were under purifying selection and approximately 10% under positive selection on the X. Mank et al. (2010) were able to show general agreement between the observed d_N/d_S values for the X chromosome and those projected given a previously inferred distribution of fitness effects (DFE) for the autosomes. However, these approaches are somewhat limited in that they rely solely on divergence and ad-hoc fitting of data to model parameters.

The naïve McDonald–Kreitman (MK) test (McDonald and Kreitman 1991) utilizes both polymorphism and divergence data to estimate α , the proportion of substitutions fixed by positive selection, but is subject to certain biases. In particular, negative α estimates can be indicative of the presence of slightly deleterious mutations (Charlesworth and Eyre-Walker 2008), which contribute to polymorphism but not divergence. However, demographic events must also be accounted for if α is to be estimated reliably. For example, population growth may result in an excess of segregating slightly deleterious mutations, which can lead to an underestimation of α (Eyre-Walker 2002). As a consequence, more robust methods for quantifying levels of purifying and positive selection have been extended from the original MK test that use the entire allele frequency spectrum (AFS) and divergence to both control for population demography and jointly estimate α and the DFE (Boyko et al. 2008; Eyre-Walker and Keightley 2009; Schneider et al. 2011). In this article, we assume that the DFE is shaped only by neutral variants and those under negative selection, although variants under positive selection can also be incorporated into the DFE (Schneider et al. 2011). Assuming a gamma distribution for the shape of the DFE, estimates for the autosomes in humans demonstrate a shape value between 0.1 and 0.3 and an average $N_e s$ of 490–7,200 (Keightley and Eyre-Walker 2007; Boyko et al. 2008; Keightley and Halligan 2011), whereas α varies between 0% and 20% (Boyko et al. 2008). No similar estimates have yet been made for the X chromosome; however, a recent analysis of exome sequence data from 12 Central Chimpanzees estimated an α of 0% and 38% for the autosomes and X chromosome, respectively. The authors concluded that these results were consistent with the faster-X effect and most new beneficial mutations being at least partially recessive ($h + < 0.5$) (Hvilsom et al. 2012).

One limitation to current implementations that simultaneously infer the DFE and α is that they are based on expressions that consider the evolution of variants assuming an autosomal inheritance pattern, equal numbers of males and females, and semidominance ($h = 0.5$). Given previous work showing that the expected ratio of X chromosome and autosome diversity is sensitive to these assumptions, in this article we integrate theory on the relative evolution of the X chromosome and autosomes derived by Charlesworth and coworkers (Charlesworth et al. 1987; Vicoso and Charlesworth 2009) with an MK-based inference framework. We then apply this methodology to high-coverage whole exome sequence data from 44 Yoruba females generated as part of the 1000

Genomes Project Phase 1 release (1000 Genomes Project Consortium et al. 2012). We find that although the X chromosome appears to have evolved adaptive substitutions at a greater rate, varying the degree of recessiveness of deleterious mutations substantially alters the discrepancy between the estimated DFE and α when comparing the autosomes to the X chromosome, whereas varying β has comparatively little effect on the distribution of deleterious selection coefficients.

Results

Nucleotide Diversity and Divergence on the X and Autosomes

Table 1 lists the number of polymorphic sites (p) found in the Yoruba and the number of substitutions (d) inferred to have occurred along the human branch on the autosomes and X chromosome for individual classes of amino acid degeneracy (4-, 3-, 2-, and 0-fold degenerate sites, where we assume that no selection has occurred at 4-fold sites). At putatively neutral 4-fold sites, $\pi_{4\text{-fold}}$ for the X chromosome was approximately 69% of that of the autosomes. Although close to an expected neutral value of 75%, this value should be interpreted in evolutionary terms with caution as it is likely confounded by an increased male-to-female mutation rate, ν , which would lower the neutral X/A ratio (Miyata et al. 1987) and an increased female-to-male breeding ratio, β , which would increase the X/A ratio (Hammer et al. 2008). In addition, linkage to selected sites within or close to genes is also likely to alter X/A diversity (Hammer et al. 2010; Gottipati et al. 2011). When we attempt to control for different mutation rates by dividing by divergence from orangutan, the X/A ratio is higher than 0.75 at approximately 0.9, although increased codon-usage bias at synonymous sites on the X chromosomes (Lu and Wu 2005) may lead to an inflated estimate of the ratio of female to male N_e .

Tajima's D was negative at 4-fold sites suggesting that the population sample contained an excess of singletons, which would be consistent with a size increase in the Yoruba from a smaller ancestral population (Boyko et al. 2008; Cox et al. 2009). Tajima's D is increasingly more negative between 4- and 0-fold sites on both the autosomes and X chromosome. This pattern is consistent with increased purifying selection at nonsynonymous sites, as is the decrease in both p and d with decreasing amino acid degeneracy. However, the ratio $(p_{0,X}/p_{4,X})/(p_{0,A}/p_{4,A})$ was 0.81, consistent with a greater overall level of purifying selection, $N_e s$, acting on the X chromosome. In contrast, $(d_{0,X}/d_{4,X})/(d_{0,A}/d_{4,A})$ was 1.40. This could be a result of stronger purifying selection on the autosomes (i.e., decreasing the fraction in the denominator), greater adaptive evolution on the X chromosome (i.e., increasing the fraction in the numerator), or a combination of both. For both the autosomes and the X chromosome, the removal of CpG sites results in a substantially lower level of diversity ($\sim 50\%$) for all classes, whereas p_0/p_4 and d_0/d_4 are elevated.

Allele Frequency Spectrum

A visual inspection of the folded AFS for the autosomes and X chromosome (fig. 1a, supplementary fig. S1a, Supplementary

Table 1. Counts of Polymorphic Sites and Substitutions for the Autosomes and X-Chromosome in 44 Yoruban Females from Deep Sequencing of CCDS Genes.

Degeneracy	Total Sites	Polymorphic Sites	Fixed Sites	Poly/Tot × 1,000	Fix/Tot × 1,000	$\theta_{\pi} \times 10,000$	$\theta_W \times 10,000$	Tajima's <i>D</i>
Autosomes								
4-fold	3,191,534	15,586	17,979	4.88	5.63	9.67	7.64	−0.73
3-fold	315,364	1,133	1,134	3.59	3.60	7.12	5.42	−0.82
2-fold	3,867,890	12,717	13,344	3.29	3.45	6.51	4.91	−0.85
0-fold	13,164,744	21,573	17,701	1.64	1.34	3.25	2.08	−1.24
Autosomes excluding CpG sites								
4-fold	2,900,809	6,326	7,638	2.18	2.63	4.32	3.37	−0.76
3-fold	294,375	568	612	1.93	2.08	3.82	2.81	−0.91
2-fold	3,624,575	7,687	8,525	2.12	2.35	4.20	3.17	−0.85
0-fold	12,408,076	12,405	10,800	1.00	0.87	1.98	1.31	−1.18
X chromosome								
4-fold	121,270	421	465	3.47	3.83	6.88	5.24	−0.81
3-fold	12,771	29	38	2.27	2.98	4.50	2.44	−1.41
2-fold	151,863	310	408	2.04	2.69	4.04	2.89	−0.97
0-fold	515,533	487	663	0.94	1.29	1.87	1.09	−1.44
X-chromosome excluding CpG sites								
4-fold	112,437	188	201	1.67	1.79	3.31	2.65	−0.68
3-fold	12,140	16	26	1.32	2.14	2.61	1.56	−1.14
2-fold	144,124	203	268	1.41	1.86	2.79	2.03	−0.93
0-fold	490,649	306	443	0.62	0.90	1.24	0.76	−1.32

NOTE.—CCDS, consensus coding sequence.

Material online) demonstrates that the putatively neutral 4-fold sites contain an excess of singletons compared with a standard neutral model, consistent with the negative Tajima's *D* values in table 1. However, 0-fold sites show an even larger excess of singletons, which is likely to be a signature of slightly deleterious mutations. Both the 4-fold and 0-fold sites on the X chromosome show an excess of singletons compared with their counterparts on the autosomes. However, the smaller target region on the X chromosome results in the AFS being noticeably noisier, whereas the discrepancy between the autosomes and the X chromosome is less pronounced when removing CpG sites (fig. 1*b*). If the excess of singletons on the X chromosome is real, this may suggest that it is subject to greater net purifying selection for both synonymous and nonsynonymous sites (Lu and Wu 2005). Increased background selection due to a lower recombination rate on the X could also contribute to this pattern (see section below on forward simulations). We used our diffusion approximation expressions to explore whether demographic changes could cause this excess for neutral sites because of differences in the X chromosome and autosomal N_e (Pool and Nielsen 2007). However, we found that demographic parameters realistic for the Yobura (i.e., the N_e doubling ~150 ka) (Gravel et al. 2011) lead to the opposite trend (i.e., a greater proportion of autosomal singleton variants) (data not shown).

Extended MK-Based Analysis

We then performed an extended MK-based analysis on the autosomal data assuming a gamma distribution for the DFE with shape, *a*, and scale, *b*, parameters and a deleterious

dominance coefficient (*h*) of 0.5 (see table 2 for definitions of notation used for all parameters in our analysis). Our results largely confirm a previous analysis (Keightley and Halligan 2011) (supplementary table S1, Supplementary Material online). The best fitting demographic parameters under a two epoch model suggest an approximate doubling of the population size ($n = 2.14$) 0.37 coalescent units ago (~200 ka assuming 25 years/generation and an N_e of 10,000), whereas the estimated shape of DFE is highly leptokurtic ($a = 0.18$, 95% confidence interval [CI] = 0.17–0.19) with a *b* of approximately 4,134 (95% CI = 2,712–6,250). When the analysis is extended to the X chromosome with $\beta = 1.0$ (i.e., equal numbers of breeding females and males), we find evidence of substantially greater purifying selection compared with the autosomes, with a mean $2N_{eAS} = ab$ of 747 for the autosomes and 4,629 for the X-chromosome (table 3) (note: Scaling our estimates of selection by $2N_{eAS}$ allows direct comparison of the distribution of deleterious selection coefficients between the autosomes and X chromosome). When CpG sites are excluded, the level of purifying selection is reduced (i.e., $2N_{eAS}$ of 288 and 2,430 for the autosomes and X chromosome, respectively). However, the estimates for the X chromosome are associated with wide 95% CIs that overlap with the CIs for the autosomes (supplementary tables S2 and S3, Supplementary Material online). The wide CI values are mostly driven by large variance in estimates of the *b* parameter, whereas *a* is relatively stable. The estimated value of α is statistically significantly higher ($P < 0.01$) on the X chromosomes (48%) compared with the autosomes (10%). When we exclude CpG sites, estimated values of α remain statistically

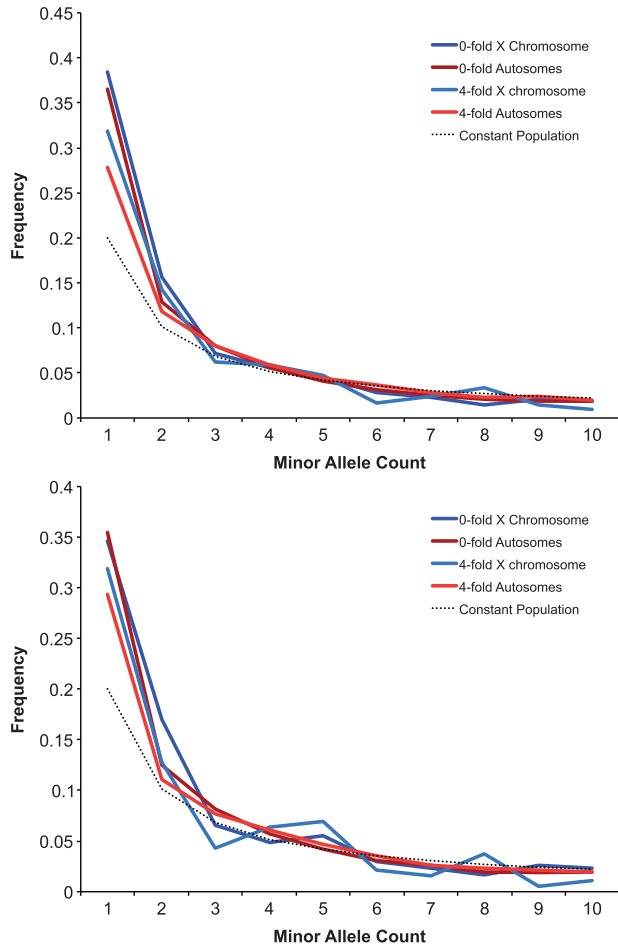


Fig. 1. Folded AFS for 44 Yoruba females for the autosomes and X chromosome at 4-fold and 0-fold sites both including (A) and excluding (B) CpG sites. The minor allele counts from 11 to 44 are not shown to aid interpretation. The full AFS are shown in [supplementary figures S1 and S2, Supplementary Material](#) online.

significantly higher on the X chromosome versus autosomes (i.e., 50% and 6%, respectively).

The genome assembly for the chimpanzee X chromosome is not as reliable as that for the autosomes, as the reference individual was a male (i.e., sequence coverage was half that of the autosomes). However, we note that our expressions involve the ratio of synonymous and nonsynonymous substitutions rather than the absolute number. Thus, if the error rate per site due to misassembly and low coverage was the same for both these classes of sites, then we might expect a priori given their vicinity to each other that our results are not highly biased (although our estimated CIs may be anticonservative).

To explore this assumption further, we repeated our analysis examining substitutions along the human lineage that have occurred since the divergence from orangutans (using the ponAbe2 reference genome). Given the reference individual was female, the orangutan should demonstrate less assembly bias for the X chromosome. We estimated an α of -3% and 34% for the autosomes and X chromosome, respectively. The negative value for the autosomes suggests that we have not fully accounted the effect of purifying selection since

Table 2. Definitions of Various Parameters Examined in This Study.

Symbol	Definition
n	Magnitude of an instantaneous change in population size some time in the past
T	The time in coalescent units ($2N_e$) of an instantaneous change in population size some time in the past
a	Shape parameter of the gamma distribution
b	Scale parameter of the gamma distribution
α	The proportion of nonsynonymous substitutions fixed by adaptive evolution
ω	Rate of adaptive substitution scaled by the rate of neutral substitution
h	The dominance coefficient for deleterious variants
$h+$	The dominance coefficient for beneficial variants
β	The female to male breeding ratio
s	The deleterious selection coefficient
γ	The net effect of selection, equal to $2N_e s = ab$
$N_{e,A}$	The effective population size for the autosomes
$N_{e,X}$	The effective population size for the X chromosome
d_0	The number of 0-fold degenerate substitutions per site
d_4	The number of 4-fold degenerate substitutions per site
p_0	The number of 0-fold degenerate polymorphisms per site
p_4	The number of 4-fold degenerate polymorphisms per site
r_A	Recombination rate per site generation

the divergence from humans using the current AFS, which is perhaps unsurprising given the much larger divergence time compared with chimpanzees. However, although both values are lower than in the analysis based on divergence with chimpanzee, the magnitude of the difference in α is almost exactly the same, confirming the suggestion that our results using chimpanzees are not highly biased.

By way of comparison, a naïve MK test using chimpanzees results in an estimated α of -40% (-33% excluding CpG sites) for the autosomes and 18% (26% excluding CpG sites) for the X chromosome, demonstrating the importance of accounting for slightly deleterious mutations and population demography. Overall, these results provide evidence for a faster-X effect in humans. However, as the relative rates of evolution for variants on the X chromosome and autosomes are sensitive to h and β , we further explore how our inference of purifying and positive selection is affected by varying these parameters.

Assessing the Impact of Varying h and β

A good fit of the AFS can be achieved when $h = 0.5$ and $\beta = 1$ ([supplementary fig. S2a–d, Supplementary Material](#) online). Therefore, we have very little power to infer h and β . Consequently, we examine the dependence of a , b , and α on the values of the deleterious dominance coefficient h for the autosomes and X chromosome and the dependence of a and b on the values of β for the X chromosome. Notably, the log likelihood for the best fit a and b parameters decreased significantly for values of h below 0.3 for the autosomes (with

Table 3. Comparison of the DFE (and related statistics) and α for the Autosomes and X Chromosome when $h = 0.5$ and $\beta = 1$.

Parameters	Autosomes	Autosomes Excluding CpG Sites	X Chromosome	X Chromosome Excluding CpG Sites
n	2.15 [2.07–2.24]	2.23 [2.12–2.39]	2.99 [2.17–4.17]	2.79 [1.52–10.00]
T	0.37 [0.30–0.46]	0.32 [0.23–0.45]	0.12 [0.05–0.17]	0.10 [0.01–0.16]
a	0.18 [0.17–0.19]	0.15 [0.13–0.17]	0.17 [0.11–0.24]	0.13 [0.08–0.23]
b	4,134 [2,712–6,250]	1,954 [905–4,964]	27,801 [3,089–999,091]	18,231 [790–999,882]
$2N_{eAS}$	747 [525–1,034]	288 [155–616]	4,629 [725–115,807]	2,431 [170–96,405]
% $2N_{eAS} < 0.1$	18.1	27.0	14.4	24.1
% $2N_{eAS} > 1.0$	72.6	61.8	78.7	67.5
% $2N_{eAS} > 10.0$	58.6	46.2	68.5	56.1
α	10% [8–14%]	6% [1–10%]	48% [39–59%]	50% [40–64%]
ω	0.024 [0.02–0.03]	0.020 [0.00–0.03]	0.16 [0.12–0.21]	0.25 [0.19–0.35]

NOTE.—95% CIs in square brackets.

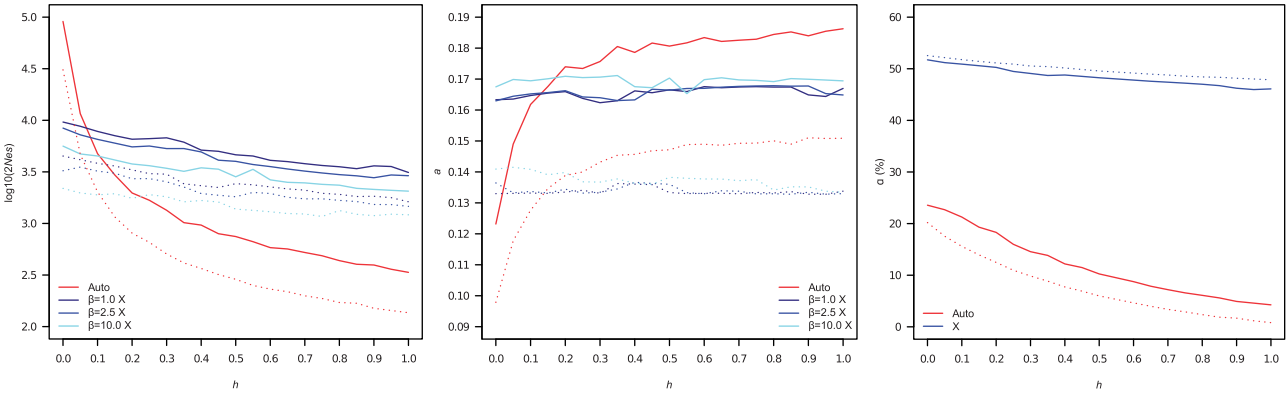


Fig. 2. Estimates of parameters of selection under various combinations of β and h for 44 Yoruba females. (A) $2N_{eAS}$, the scale parameter of the gamma distribution for the DFE, (B) a , the shape parameter of the gamma distribution for the DFE, and (C) α , the proportion of substitutions along the human lineage since divergence from a human–chimpanzee ancestor that fixed as a result of adaptive selection. Solid lines are when including CpG sites and dashed lines are when excluding CpG sites.

the log likelihood increasing by ~ 120 from $h = 0.0$ to $h = 1.0$) (see [supplementary fig. S3a](#), [Supplementary Material](#) online, for profile log likelihood curves). This appears to be a consequence of the fact that the best fit model for h values close to 0 underestimates the excess of observed singletons. The opposite trend was seen for the X chromosome with the largest log likelihood at $h = 0.0$, although the difference was not statistically significant (i.e., log likelihood ~ 0.5 higher compared with $h = 1.0$) ([supplementary fig. S3c](#), [Supplementary Material](#) online). Excluding CpG sites changed the magnitude of log likelihood values but not the trend ([supplementary fig. S3b](#) and [d](#), [Supplementary Material](#) online).

When examining the estimated parameter values of the gamma distribution for the DFE, we found that the X chromosome was relatively insensitive to changing h . For instance, the mean $2N_{eAS}$ reached its maximum value when $h = 0.0$ (9.6×10^3) and remained quite large when $h = 1.0$ (3.1×10^3), albeit with wide CIs. In comparison, the mean $2N_{eAS}$ varied substantially for the autosomes, yielding values as high as 9.1×10^4 at $h = 0.0$ and dropping down to as low as 3.4×10^2 when $h = 1.0$ ([fig. 2a](#)). On the autosomes, the shape parameter, a , was affected by h to a much lesser extent, climbing from 0.12 at $h = 0.0$ to 0.19 at $h = 1.0$, whereas for the X chromosome, a was essentially constant at 0.16–0.17

regardless of h ([fig. 2b](#)) (note: As a is relatively constant, the inferences for mean $2N_{eAS}$ and b are essentially the same, [supplementary fig. S4](#), [Supplementary Material](#) online).

Increasing β from 1 to 2.5 and 10.0 for the X chromosome decreases estimates of mean $2N_{eAS}$ by approximately 15% (min = 4%, max = 23% depending on h) and 38% (min = 26%, max = 49% depending on h), respectively ([fig. 2a](#)) (a increases but the effect is small). The decrease observed for $\beta = 10$ is slightly greater than predicted analytically ($\sim 27\%$) when projecting from fixed demographic parameters n and T and shape parameter a estimated for $\beta = 1$. This appears to reflect noise from the sparse X chromosome data that propagates through the optimization of both demographic and selection parameters, as optimizing of b while fixing the n , T , and a parameter values to those inferred for $\beta = 1$ gave very similar estimates to the analytical predictions. In general, the decrease in the level of purifying selection required to fit the data when increasing β is not as large as the effect of decreasing h for the autosomes ([fig. 2a](#)). When CpG sites were excluded from the analysis, the estimates of mean $2N_{eAS}$ were decreased by approximately 50% for both the autosomes and the X chromosome ([fig. 2a](#)).

Increasing h results in a decrease in α , with this effect again being larger for the autosomes ([fig. 2c](#)). Over the range of

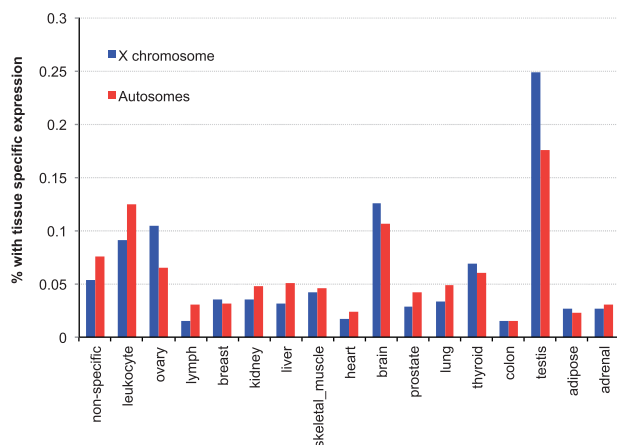


Fig. 3. Comparison of gene expression tissue specificity for X chromosome and autosomes based on RNA-Seq data from 16 human tissue generated as part of the Illumina Body Map 2.0.

dominance coefficient values, the estimated value of α decreases by 19% for the autosomes and only by 5% for the X chromosome. A similar result is obtained when removing CpG sites. As discussed in the Materials and Methods section, we note that β is not expected to affect our estimate of α . Regardless of the values of h considered, α is always statistically significantly higher ($P < 0.01$) for the X chromosome (46–52%) than for the autosomes (5–24%). Similar results were obtained when excluding CpG sites (i.e., 48–53% for the X chromosome and 1–20% for the autosomes). However, the magnitude of this difference ranges from 28% when $h = 0.0$ to 42% when $h = 1.0$. If h is allowed to differ on the autosomes and X chromosome, the difference in α can be as high as 47% when $h = 1.0$ for the autosomes and $h = 0.0$ for the X chromosome. The maximum difference when excluding CpG sites under the same parameters is 52%.

Controlling for Differing Gene Content

It has previously been shown that the human X chromosome possesses more tissue-specific expressed genes than the autosomes, as well as a bias toward male-specific genes (Lercher 2003). A recent study comparing α on the X chromosome and autosomes in house mice (Kousathanas et al. 2013) performed a detailed breakdown of gene content and demonstrated that the faster X effect was enhanced for genes with male-specific tissue expression. We are unable to perform such a nuanced analysis with our human data set because the level of diversity is substantially lower than that of house mice (i.e., N_e is ~ 50 times smaller in humans than in mice) (Phifer-Rixey et al. 2012). Therefore, we performed a more coarse examination of α where we use RNA-Seq data from 16 human tissues (the Illumina Body Map data set) to approximately match the tissue expression profile of the X chromosome to that of the autosomes by subsampling genes from the latter.

Genes called in both the 1000 Genomes high coverage exome sequencing data set and the Illumina body map (13,447 on the autosomes, 595 on the X chromosome)

were defined as either tissue specific or expressed in all tissues (fig. 3). As shown previously, the number of genes with testis-specific expression is high (Ramsköld et al. 2009; Djureinovic et al. 2014), making up approximately 17% of all genes expressed on the autosomes. This number is higher than that proposed by Djureinovic et al. (2014), who used a much more stringent criteria. Our number is more similar to the 19% estimate of Kousathanas et al. (2013), who used a similar definition as here. Interestingly, testis-specific genes appear notably enriched on the X chromosomes compared with the autosomes (25%, an increase of 8%). We resampled 10,000 genes from the autosomal set, such that the tissue-specific expression profile matched that of the X chromosome and reperformed our extended-MK analysis. The major results were largely unchanged from the original analysis, with α for the autosomes and X chromosome being 13.1% and 45.1%, respectively.

Assessing the Impact of Linkage on Inferring the DFE

Although our extended MK methodology assumes independence between sites, there is likely to be at least some linkage between sets of segregating sites that could distort the AFS through background selection and genetic draft. Using forward simulations, Messer and Petrov (2013) demonstrated that for a genomic architecture that is realistic for humans, estimates of α are relatively robust to linkage effects using the extended MK framework when neutral sites are interdigitated among selected sites (i.e., as would be the case when analyzing synonymous mutations alongside nonsynonymous mutations). Essentially, the “demographic model” estimated from the neutral sites encompasses both demographic and linkage effects and thus compensates when estimating α .

However, the caveat is that estimates of the shape and scale parameters of the DFE will be biased, with this bias potentially increasing with greater levels of adaptive evolution. As the majority of the X chromosome does not undergo recombination in males, the level of linkage may be higher than on the autosomes, and thus, the bias in estimating the DFE is potentially more severe. On the other hand, the lack of recombination in males is countered by the recombination rate in females being approximately 1.7 times that of males (Roach et al. 2010). Thus, rather than the sex-averaged recombination rate for the X chromosome (r_X), being approximately 0.67 that of the autosomes (r_A), it is likely to be closer to approximately 0.85. We conducted a series of forward simulations under a demographic scenario realistic for the Yoruba involving a doubling of the population size 150 ka (when compared with Messer and Petrov [2013] who simulated a constant sized population at equilibrium) to examine potential bias in our estimates of the DFE under different recombination rates, r .

We first examined how very large changes in r would affect our inference, exploring a range from 5×10^{-6} to 1×10^{-11} . As would be expected, decreasing the recombination rate decreases the number of segregating sites in the entire population ($N = 20,000$), but the effect is markedly stronger for synonymous sites (supplementary fig. S5a, Supplementary

Material online). Simulations with positive selection (1% of nonsynonymous sites possess a positive s of 0.001) show only slight increases and decreases in the number segregating mutations for synonymous and nonsynonymous sites, respectively, compared with a model with just purifying selection (parameterized as a gamma distribution). When analyzed using the extended MK framework, the estimated magnitude (n) and time (T) of the population size change are generally accurate with decreasing recombination rate up to 5×10^{-8} , after which they both begin to increase rapidly compared with the true value. This deviation is further increased by the inclusion of positive selection (supplementary fig. S5b and c, Supplementary Material online). Despite this, inference of the shape parameter a is quite consistent, whereas the inclusion of positive selection always leads to a slight but systematic underestimation of this parameter (supplementary fig. S5d, Supplementary Material online).

However, the effect on our inferred scale parameter b is more dramatic (supplementary fig. S5e, Supplementary Material online). Inference with recombination rates below 1×10^{-8} substantially underestimate the true value, with the effect becoming increasingly more pronounced with lower recombination rate. The inclusion of positive selection always leads to an increase in the inferred b compared with a regime with only purifying selection.

These analyses give a broad sense of how recombination may affect our inference, yet the results at $r = 1 \times 10^{-8}$ are most relevant to human populations. Therefore, we performed simulations that explored the effect of r between 1×10^{-8} and 5×10^{-9} under parameters relevant to the autosomes and X chromosome. Although there is a reduction in diversity with decreasing recombination rate for both the autosomes and X chromosome, the effect is minor, especially for nonsynonymous sites (supplementary fig. S6a, Supplementary Material online). Estimates of n and T are slightly overestimated, with the former showing a general trend of an increase with decreasing recombination rate, and positive selection increasing the bias for both. Again, estimates of a tend to be relatively accurate when there is only purifying selection and consistently underestimated when there is positive selection (supplementary fig. S6d, Supplementary Material online). When inferring b , although there is some apparent trend with decreasing recombination rate for the X chromosome and only purifying selection, noise from the estimation process and sampling tend to dominate (supplementary fig. S6e, Supplementary Material online). For both the autosomes and the X chromosome in this range of r , it appears that we underestimate the true value of b by as much as half when there is no positive selection. On the other hand, the presence of positive selection will increase our estimate, although the level of positive selection will ultimately determine whether we under- or overestimate the true value.

Discussion

Faster-X Effect in Humans

Using a quantitative framework incorporating previous theoretical work (Charlesworth et al. 1987; Vicoso and

Charlesworth 2009), we have provided evidence that the proportion of substitutions fixed due to positive directional selection (α) along the human lineage is substantially larger on the X chromosome than on the autosomes. As far as we are aware, this is the first quantitative demonstration accounting for both purifying and positive selection that the faster-X effect operates in humans. This work builds on previous divergence-only estimates that were more limited in distinguishing the relative effects of purifying and positive selection (Lu and Wu 2005; Nielsen et al. 2005; Mank et al. 2010).

Our analysis only explores the effect of the deleterious dominance coefficient, h , leaving open the question of whether the faster-X effect we observe in humans is driven in part by recessive beneficial variants ($h + < 0.5$) having a faster substitution rate on the X chromosome (Charlesworth et al. 1987). Differences in gene content between the X chromosome and autosomes could also contribute to the patterns we observe. In particular, there is evidence that the X chromosome is enriched for male-specific genes and has a narrower breadth of expression, both from our analysis of RNA-seq data and previous work (Lercher 2003), consistent with Rice's hypothesis (1984). We attempted to control for gene content by matching RNA-Seq profiles; however, this analysis ignores many other aspects of gene expression patterning. The house mouse study of Kousathanus et al. (2013) provides clear evidence that male-specific genes, and in particular those that do not undergo meiotic sex chromosome inactivation, contribute substantially to the faster X effect. Whether these findings also apply to humans is as yet unclear. Perhaps, a promising avenue in the future will be to explore individual gene content categories in other apes where gene diversity is much higher (Prado-Martinez et al. 2013) and where there should be greater power to observe such effects within these individual categories.

With this in mind, our results are consistent with a similar study of Central Chimpanzees (*Pan troglodytes troglodytes*), although they did not explicitly model the evolution of mutations on the X chromosome (Hvilsom et al. 2012). This suggests that faster-X evolution may be a common feature in the great apes. Although Hvilsom et al. (2012) suggested that their results were consistent with adaptive variants being at least partially recessive, this need not be the case if the female-to-male breeding ratio is greater than 1 (Vicoso and Charlesworth 2009). Hvilsom et al. (2012) argued that because X/A diversity at synonymous sites was substantially below the neutral expectation of 0.75, β is likely to be less than one. They reported a value of 0.4 in chimpanzees, whereas we observe a value of 0.7 in humans. However, it is well recognized that X/A diversity close to genes is likely to be skewed downward due to a higher male mutation rate (Kong et al. 2012), greater background selection on the X chromosome (Hammer et al. 2010), and greater selection at synonymous sites on the X chromosome (Lu and Wu 2005). Two independent studies have estimated β to be approximately 2.5 in humans. The first examined X/A diversity far away from genes (Hammer et al. 2008) and the second considered relative recombination rates (Lohmueller et al. 2010). In addition, X/A diversity far away from genes has been found to be more than 0.75 in seven out

of nine great apes species for which there are whole-genome sequencing data from multiple individuals (Prado-Martinez et al. 2013). Thus, the finding that β is greater than one in great apes suggests that faster-X evolution could have occurred even if some new beneficial mutations were at least partially dominant.

Our inference that α is greater on the X chromosome compared with the autosomes is robust to a range of biologically realistic values of h for deleterious mutations. Regardless of our choice of the dominance coefficient, our parameter estimates for α for the X chromosome suggest that approximately half of all substitutions occurring along the human lineage have been driven to fixation by positive selection. However, we note that the magnitude of the X-autosome difference in estimates of α , and thus the extent of the faster-X effect, depends on the choice of the h value (fig. 2c). For example, although α for the X chromosome ranges from 46% (when $h = 1$) to 52% (when $h = 0$), α ranges from 4% to 24% for the autosomes under the same h values. Therefore, the faster-X effect, as measured by the reduction in α compared with the X chromosome, ranges from 28% to 42%. Despite this variation, α is highly statistically significantly different for the X and autosomes under the entire range of parameter values.

In addition, we find larger mean $2N_{eAS}$ values for the DFE with decreasing h , with a substantially greater rate of change for the autosomes. At h close to 0, the distribution of selection coefficients estimated for the autosomes is similar to that estimated for the X chromosome. However at $h = 1$, selection coefficients are approximately 90% smaller for the autosomes compared with the X chromosome. Our intuitive explanation is that varying h has less of an effect on the X chromosome because, regardless of the level of recessiveness, hemizygosity always leads variants to be exposed to selection in males soon after their introduction. Hence, the parameter a , which is largely determined by the shape of the AFS, is essentially unaffected by changes in h for the X chromosome and b , which is largely determined by the observed reduction in nonsynonymous diversity, is only moderately affected. On the other hand, decreasing h better protects variants from the effect of purifying selection on the autosomes. Thus, to obtain the reduction of diversity seen at 0-fold versus 4-fold sites, the strength of purifying selection increases more on the autosomes to counterbalance the effects of decreasing h . This then reduces the expected number of substitutions occurring at 0-fold sites (K_a), resulting in a faster rate of increase of α on the autosomes (as α is proportional to $(d_0 - \langle K_a \rangle)^{-1}$).

Uncertainties in Estimating h

Despite its fundamental importance in both the interpretation of our results regarding relative levels of purifying and positive selection on the autosomes and X chromosome, the nature of dominance coefficients for deleterious mutations is not well understood. In humans, recessive diseases are much more prominent than dominant diseases (Wilkie 1994), though it is not clear how this informs the assignment of h values across the genome. Most hard data for estimates of h

come from experimental evidence examining the relative fitness of heterozygotes and homozygotes in *Drosophila*, *Caenorhabditis elegans*, *Arabidopsis*, and *Saccharomyces cerevisiae* (Fay and Wu 2003; Manna et al. 2011). It is difficult to draw reliable inferences from these studies, although recessiveness (i.e., $h < 0.5$) does appear to be more common. There is also substantial variation in estimates of h across variants, whereas variants with larger selective effects tend to have a lower h than variants with weaker effects (and even overdominance may be common) (Agrawal and Whitlock 2011). These studies indicate that our use of a single h value is an oversimplification. Although it would be easy to include a prior distribution for h in a Bayesian approach, currently it is difficult to envisage the shape of this distribution. Moreover, inferring the DFE from the AFS alone is unlikely to provide sufficient power to infer this distribution along with a and b (Boyko et al. 2008). We note for the autosomes that h values close to zero result in significantly lower likelihoods as they are unable to generate a similar excess of singletons as observed in our data. Presumably, this is because purely recessive variants are shielded from the effects of natural selection until they reach high population frequencies (Williamson et al. 2004). Thus, we infer that most deleterious variants segregating on the autosomes are unlikely to possess dominance coefficients close to zero.

There are currently no experimental data that shed light on the expected h value for X-linked deleterious variants in mammals (Connallon and Clark 2013), although mutation-accumulation experiments for the X chromosome in *Drosophila melanogaster* are suggestive of dominance coefficient of 0.3 (Mallet et al. 2011). Mank et al. (2010) suggested $h = 0.5$ may be a reasonable value to mimic the effects of X inactivation in therian mammals. However, even though only one of the two available alleles is expressed randomly in each cell throughout the body, this may not necessarily translate to an overall fitness effect of $h = 0.5$.

The Impact of CpG Sites

As expected, the exclusion of CpG sites results in a substantially lower level of diversity ($\sim 50\%$) for all degeneracy classes on both the autosomes and the X chromosome. This is unsurprising given the approximately 10-fold elevation in mutation rate at these sites (Hodgkinson and Eyre-Walker 2011) and the enrichment of CpG sites in mammalian coding regions (Schmidt et al. 2008). However, it is also interesting to note that the exclusion of CpG sites tends to dampen the inferred level of purifying selection we observe. The average value of $2N_{eAS}$ when excluding CpG sites is reduced by approximately 60% for autosomes and approximately 50% for the X chromosome, which is consistent with previous findings suggesting that purifying selection is greater at CpG sites (Schmidt et al. 2008). However, this exclusion only subtly changes our estimates of α , leading to a approximately 4% decrease for the autosomes and a approximately 1% increase for the X-chromosome with overlapping CIs. This suggests there is no major difference with respect to whether CpG or non-CpG sites will go to fixation due to positive selection

after controlling for their different mutation rates. Thus, our α estimates appear to be relatively robust to any differences in the underlying nucleotide context, whether between the autosomes and X chromosome, or between 4-fold and 0-fold sites.

Other Considerations

There are a number of caveats due to simplifying assumptions in our methodology that could potentially affect our inference. One immediate concern is our use of 4-fold sites as putatively neutral. There are advantages to using synonymous sites because they will be exposed to genetic drift and background selection in a similar manner as nonsynonymous sites, and hence, they can be used to control for these forces when inferring α (Messer and Petrov 2013) (see below). However, there is also evidence that purifying selection may be acting directly on 4-fold sites (i.e., via codon usage bias), and this effect may be enhanced on the X chromosome (Lu and Wu 2005). If this is the case, then our estimates of purifying selection at 0-fold sites may be an underestimate, which in turn would lead to estimates of α that are too low, as $\langle K_a \rangle$ will be too large. The fact that we obtain similar estimates of α after removing CpG sites suggests that the effect of selection at synonymous sites may not be substantial (i.e., selective constraint should be reduced at 4-fold sites when CpGs are excluded). In future studies, it would be preferable to incorporate the AFS for nongenic and intronic sites, the latter of which, being close to genes, would better capture some of the effects of genetic drift and background selection. High coverage whole-genome sequencing of many individuals from single populations is likely to make this feasible very soon, although an alternative would be to utilize methods that can accurately infer AFS from low coverage data (Nielsen et al. 2012).

One of the major assumptions of our diffusion approximation approach is that variants evolve independently at all sites. We note that linkage effects such as Hill–Robertson interference (Hill and Robertson 1966) would violate this assumption, with the effect potentially more pronounced on the X chromosome where recombination only occurs in females. We have attempted to account for linkage effects by utilizing 4-fold sites on both the autosomes and X chromosome separately, such that the fitting of a demographic model captures chromosome-specific linkage effects. Therefore, our estimates of α are likely to be relatively robust (Messer and Petrov 2013). However, our forward simulations with varying recombination rates suggest that if positive selection on the X chromosome is relatively high compared with the autosomes, then it is likely there is greater bias in our estimates of the DFE for the X chromosome, with b in particular likely to be inflated. Our simulations also suggest that the difference in recombination rate between the X chromosome and autosomes is probably not a major factor in this bias if r_X is approximately 0.85 that of r_A . We note that despite the fact that we only examined a small amount of the possible parameter space, our forward simulations required substantial computing time. To accurately infer the DFE in humans for both the autosomes and the X chromosome, future extensions

to MK-based analysis will need to find novel ways to take into account complex genomic architecture with linkage between variant sites.

Our estimates of fixation probabilities assume a constant selection coefficient. However, the selection trajectory for a given variant may change through time (e.g., selection acting on previously neutral standing variation). To take this into account would require more complex models incorporating polymorphism, divergence, dominance, and linkage (see e.g., Connallon et al. 2012). However, we note that the faster-X effect is predicted to disappear if selection is on standing variation (Orr and Betancourt 2001). Hence, a possible implication of our results is that most selection is on new mutations rather than on standing variation.

Our model also assumes the same selection coefficient at a given site for males and females. Vicoso and Charlesworth (2009) explored how antagonistic selection might affect relative substitution rates for the X chromosome and autosomes and found that advantageous male-biased variants can experience faster-X evolution if $h < 0.5$. Although we could incorporate opposing s values into our model, it is unlikely that we would be able to infer separate values of s for males and females.

Conclusion

We have integrated classical population genetics theory that contrasts autosome and X-chromosome evolution with an extended MK-based method that jointly estimates purifying and positive directional selection. This has allowed us to provide quantitative evidence that in humans the X chromosome has experienced significantly increased levels of both types of selection. However, the exact magnitude of faster-X evolution hinges on a number of mutational and evolutionary processes that require further investigation in humans and other apes, whereas additional methodological developments for inferring long-term selection in the presence of linkage, sex-specific differences, and alternative models of selection are necessary to gain a more comprehensive understanding of autosome and X-chromosome diversity.

Materials and Methods

Data and Bioinformatics Pipeline

We utilized variant call files (VCFs) assembled as part of the 1000 Genomes Phase 1 release (1000 Genomes Project Consortium et al. 2012), which contain variable sites aligned against GRCh37/hg19 across 1,092 samples from 14 populations. These data were generated using a combination of low-coverage whole-genome sequencing (2–6X) and deep (50X–100X) exome sequencing targeting approximately 15,000 consensus coding sequence genes. To minimize the effect of sequencing errors when inferring the AFS, we focused only on variant sites identified by the deep sequencing. The lower coverage data are likely to lead to highly skewed AFS, particularly in the singleton class for which there is substantial information (Han et al. 2014). Therefore, our data consisted of 163,377 targeted regions across 23,177,032 bp on the

autosomes and 7,023 targeted regions across, 1,017,890 bp on the X-chromosome.

To minimize variant calling biases due to differences in ploidy, we limited our analysis to variant sites in 44 unrelated Yoruba females. Because of the high coverage for these targeted regions, we assumed that sites not found in the VCF (but within the targeted regions) were homozygous reference (i.e., rather than not being called because of low coverage/poor sequence quality). At a mean coverage of $80 \times$ (which is the mean for the 1000 Genomes exome data examined here), the probability of observing less than 20 reads at any particular site is 2.8×10^{-16} (assuming a random distribution of reads and a Poisson distribution of coverage at any particular site), and thus, the majority of targeted sites should be well called. The choice of Yoruba as the study population was motivated by previous work showing that the AFS of African populations can be fit reasonably well with a simple two-epoch model (Boyko et al. 2008). In contrast, non-African populations may require more complicated demographic models to fit the data, whereas differential archaic introgression on the X chromosome and autosomes may also complicate our inference (Sankararaman et al. 2014). VCFtools (Danecek et al. 2011) was used to extract these data from the original VCF.

Sites were annotated as 0-fold, 2-fold, 3-fold, and 4-fold sites using SnpEff (Cingolani et al. 2012) based on Ensembl canonical transcripts (defined by SnpEff as the longest Coding DNA sequence among the protein coding transcripts in a gene). The UCSC 100-way MultiZ alignment was used to determine the chimpanzee (panTro4) allelic state. Any sites where the panTro4 nucleotide was missing or where the panTro4 or hg19 nucleotide was repeat masked was excluded from further analysis. Any sites with more than two alleles (across all 44 Yoruba and the panTro reference) or that contained an insertion or deletion were also excluded. Substitutions were determined by comparing the hg19 and panTro4 reference nucleotides, though any site fixed in the Yoruba for the nonreference allele was reverted, such that the human ancestral allele was that found in the Yoruba. Substitutions were assigned to the human or chimpanzee branch hierarchically using the nucleotide state in the phylogenetically nearest available outgroup among the Gorilla (gorGor3), Orangutan (ponAbe2), and Rhesus Macaque (rheMac3) reference genomes. If the nearest outgroup possessed a nucleotide not found in the human or chimpanzee alignment, the next nearest outgroup was used instead. If none of the three outgroups possessed a compatible nucleotide, that site was excluded from further analysis. It is possible that using a single outgroup may misassign the ancestral human–chimpanzee allele due to recurrent mutations. However, as we are only interested in the ratio of the number of substitutions between different degeneracy classes (see below), this potential source of error should have minimal impact on our final inference, assuming this particular error rate is approximately the same across degeneracy classes. We retained only substitutions inferred to have occurred along the human branch since their split from a human–chimpanzee ancestral population. To control for potential

biases caused by differing CpG content across degeneracy classes and/or chromosomes (and thus differing mutation rates), we produced a second data set that excluded all CpG sites found in any Yoruba sequence or the hg19 and panTro4 reference sequences.

Overview of Approach for Inference of Purifying and Positive Selection

To formally quantify the level of purifying and positive selection, we applied an extended MK-based method that 1) fits neutral polymorphism data in the shape of an AFS to a demographic model (2-epoch instantaneous growth), then 2) uses the estimated demographic parameters along with the AFS for putatively selected polymorphism to estimate a DFE assuming purifying selection, and finally, 3) controls for this distribution when calculating α (Eyre-Walker and Keightley 2009). We were also interested in understanding how varying h and β affected our inference (the latter only applies to the inference of negative selection coefficients with respect to the X chromosome, see below). Williamson et al. (2004) have previously shown that the shape of AFS can vary substantially with h for a single $2N_e s$ value, but there is unlikely to be sufficient information in the AFS to simultaneously infer the DFE and h (Boyko et al. 2008). In addition, β is better estimated using other methods (see Discussion). As a consequence, we considered h and β as fixed values and estimated n , T , a , b , α , ω (see table 2 for definitions) for various combinations of these two fixed parameters, following steps (1) to (3) above for each combination. We considered 21 values for h ranging from 0.0 (completely recessive) to 1.0 (completely dominant) for both the autosomes and X chromosome, as well as three separate β values for the X chromosome (1, 2.5, 10). $\beta = 2.5$ and $\beta = 10$ were chosen as they encompass the ranges found for previous estimates in modern humans (with 2.5 being closest to previous point estimates for the Yoruba) (Hammer et al. 2008; Lohmueller et al. 2010).

The steps involved in the extended MK-based framework can be performed simultaneously as in Schneider et al. (2011). Because we increased the complexity of the analysis somewhat by introducing expressions for the X chromosome, h and β , we chose to initially perform the analysis in stepwise fashion. We use a diffusion approximation implemented in the software $\partial\alpha\partial i$ and previously described by Boyko et al. (2008) rather than transition matrix methods for inference of the DFE (Eyre-Walker and Keightley 2009), but the underlying principles are similar. Below we detail the theory and implementation of our approach.

Theory

The Fitness Model

For selection coefficient s and dominance coefficient h , we have assumed the fitness model for a neutral allele A_1 and a selected allele A_2 as shown in table 4. Unlike in Charlesworth et al. (1987), we do not consider different values of s in males and females.

Table 4. The Fitness Model Assumed.

Locus	Females			Males		
Autosomal						
Genotype	A ₁ A ₁	A ₁ A ₂	A ₂ A ₂	A ₁ A ₁	A ₁ A ₂	A ₂ A ₂
Fitness	w _{11,f}	w _{12,f}	w _{22,f}	w _{11,m}	w _{12,m}	w _{22,m}
	1	1 + <i>hs</i>	1 + <i>s</i>	1	1 + <i>hs</i>	1 + <i>s</i>
X linked						
Genotype	A ₁ A ₁	A ₁ A ₂	A ₂ A ₂	A ₁		A ₂
Fitness	w _{11,f}	w _{12,f}	w _{22,f}	w _{1,m}		w _{2,m}
	1	1 + <i>hs</i>	1 + <i>s</i>	1		1 + <i>s</i>

Diffusion Approximations

For a population having variance effective population size, N_e , let A_1 and A_2 have, respectively, marginal fitness, w_1 and w_2 . Let x denote the frequency of A_2 . Then in time Δt , the allele frequency

- has a shift in mean $M(x)\Delta t = x(1-x)(w_2 - w_1)\Delta t$ due to selection, and
- has a shift in variance $V(x)\Delta t = x(1-x)\Delta t/N_e\Delta t$ due to drift.

Here and throughout, we use “drift” in the sense of a population genetic model where it represents the change in allele frequency due to random sampling.

For x , the fraction of the population having allele A_2 , the marginal fitness of A_i , $i = 1, 2$ is as follows:

$$w_i = (1-x)w_{i1} + xw_{i2}, w_{21} = w_{12}.$$

Using this formula and table 4, noting that each autosomal chromosomal segment spends 1/2 of its time in females and 1/2 of its time in males, we find for the autosomes that the difference in marginal fitness between A_2 and A_1 is

$$w_{2,A} - w_{1,A} = ((1-x)h + x(1-h))s.$$

Thus, for the diffusion approximation

$$M_{s,h,A}(x) = x(1-x)((1-x)h + x(1-h))s$$

and

$$V_{s,h,A}(x) = x(1-x)/N_{e,A},$$

where $N_{e,A}$ is the effective population size of the autosomes.

For the X chromosome, each segment spends 2/3 of its time in females and 1/3 of its time in males. The difference in marginal fitness between A_2 and A_1 is

$$w_{2,X} - w_{1,X} = 1/3(1 + 2((1-x)h + x(1-h)))s.$$

Thus, for the diffusion approximation

$$M_{s,h,X}(x) = 1/3x(1-x)(1 + 2((1-x)h + x(1-h)))s$$

and

$$V_{s,h,X}(x) = x(1-x)/N_{e,X},$$

where $N_{e,X}$ is the effective population size of the X chromosome.

Effective Population Sizes

To compare these two diffusion approximations, we must know the ratio $N_{e,X}/N_{e,A}$ of the effective population sizes. For N_f females and N_m males and for a female-to-male breeding ratio $\beta = N_f/N_m$. We assume a Poisson distribution for the number of offspring for both sexes. Our goal is to determine the effective population size for use in the drift term V , and thus, we use the variance effective population size.

$$\frac{1}{N_{e,A}} = \frac{1}{4N_m} + \frac{1}{4N_f}$$

or

$$N_{e,A} = \frac{4N_fN_m}{N_f + N_m} = \frac{4\beta}{\beta + 1}N_m = \frac{4}{\beta + 1}N_f.$$

Thus,

$$N_m = \frac{\beta + 1}{4\beta}N_{e,A} \text{ and } N_f = \frac{\beta + 1}{4}N_{e,A}.$$

On the X chromosome,

$$\frac{1}{N_{e,X}} = \frac{2}{9N_m} + \frac{4}{9N_f}$$

or

$$\begin{aligned} N_{e,X} &= \frac{9N_fN_m}{2N_f + 4N_m} = \frac{9\beta}{2\beta + 4}N_m = \frac{9\beta}{2(\beta + 2)}\frac{\beta + 1}{4\beta}N_{e,A} \\ &= \frac{9(\beta + 1)}{8(\beta + 2)}N_{e,A}. \end{aligned}$$

This gives the drift term for the evolution of A_2 on the X chromosome,

$$V_{s,h,\beta,X}(x) = x(1-x)/N_{e,X} = x(1-x)\frac{8(\beta + 2)}{9(\beta + 1)}/N_{e,A}.$$

Mutation Rates

Let μ_f and μ_m be, respectively, the mutation rates for individual males and females. Then, on the autosomes,

$$(\mu_m + \mu_f)N_f \text{ and } (\mu_m + \mu_f)N_m$$

are the rate of new mutations that appear, respectively, in breeding females and breeding males. Thus, the total rate of new mutations is

$$\begin{aligned} (\mu_m + \mu_f)(N_f + N_m) &= (\mu_m + \mu_f)(\beta + 1)N_m \\ &= (\mu_m + \mu_f)\frac{(\beta + 1)^2}{4\beta}N_{e,A}. \end{aligned}$$

For the X chromosome, mutations carried in males must have arisen in females. Here, the total rate of new mutations is

$$\mu_fN_m = \mu_f\frac{\beta + 1}{4\beta}N_{e,A}.$$

Fixation Probabilities

Let τ be the time of fixation of the allele A_2 for p_t the diffusion approximation for the proportion of A_2 at time t . Define the fixation probability

$$u(q) = P\{p_\tau = 1 \mid p_0 = q\}.$$

Then u is the solution to the differential equation

$$0 = u'(q)M(q) + \frac{1}{2}u''(q)V(q)$$

with $u(0) = 0$ and $u(1) = 1$. The solution can be written explicitly.

$$u(q) = \int_0^q G(y)dy / \int_0^1 G(y)dy \text{ where}$$

$$G(y) = \exp\left(-2 \int_0^y \frac{M(z)}{V(z)}dz\right).$$

For the autosomes,

$$\frac{M_{h,s,A}(z)}{V_{h,s,A}(z)} = N_{e,A}((1-z)h + z(1-h))s$$

$$= \gamma(h + (1-2h)z)$$

where $\gamma = N_{e,A}$. Thus (Kimura 1962),

$$G_{s,h,A}(y) = \exp(-\gamma(2hy + (1-2h)y^2))$$

For the X chromosome,

$$\frac{M_{h,s,A}(z)}{V_{h,s,A}(z)} = \frac{3(\beta+1)}{8(\beta+2)} N_{e,A}(1 + 2((1-z)h + z(1-h)))s$$

$$= \frac{3(\beta+1)}{8(\beta+2)} \gamma(1 + 2h + 2(1-2h)x).$$

Thus,

$$G_{s,h,\beta,X}(y) = \exp\left(-\frac{3(\beta+1)}{4(\beta+2)} \gamma((1+2h)y + (1-2h)y^2)\right).$$

Initial Frequencies

For the autosomes

- $q_{m,A} = \frac{1}{4N_m} = \frac{\beta}{(\beta+1)N_{e,A}}$ for a mutation in a single breeding male and
- $q_{f,A} = \frac{1}{4N_f} = \frac{1}{(\beta+1)N_{e,A}}$ for a mutation in a single breeding female.

For the X chromosome

- $q_{m,X} = \frac{1}{3N_m} = \frac{4\beta}{3(\beta+1)N_{e,A}}$ for a mutation in a single breeding male and
- $q_{m,X} = \frac{1}{3N_f} = \frac{4}{3(\beta+1)N_{e,A}}$ for a mutation in a single breeding female.

This is determined by taking the fraction of alleles carrying the mutation ($1/2N_m$, $1/2N_f$, or $1/2N_m$, $1/N_f$) and multiplying by the contribution to the population ($1/2$ or $2/3$, $1/3$), respectively, for the autosomes and the X chromosome.

Because $G(0) = 1$, we have for any of the initial frequencies q ,

$$\int_0^q G(y)dy \approx q$$

which is inversely proportional to N_e . Consequently, $u(q)$ is inversely proportional to N_e and thus the dependence of $u(q)$ on N_e appears only through the value of $\gamma = N_e s$. For example, to determine the substitution rate on the autosomes in the case $\beta = 1$ and $h = 1/2$, note that $G_{s,1/2,A}(y) = \exp(-N_e s y)$. Then, we obtain the well-known formula

$$u(q_{m,A}) + u(q_{f,A}) \approx \frac{q_{m,A} + q_{f,A}}{\int_0^1 \exp(-N_e s y) dy}$$

$$= \frac{1/2N_{e,A} + 1/2N_{e,A}}{(1 - \exp(-N_e s))/N_e s} = \frac{s}{1 - \exp(-N_e s)}.$$

Substitution Rates

Let $U_{m,A}$, $U_{f,A}$ and $U_{m,X}$, $U_{f,X}$ be the male and female fixation probabilities for a new mutation, respectively, on the autosome and the X chromosome with the initial conditions described above. Thus, these are a function of the genetic parameters s and h and, for the X chromosome, the breeding ratio β . For autosomes, $(\mu_f + \mu_m)N_f$ and $(\mu_f + \mu_m)N_m$ are the expected number of new mutations that appear, respectively, in breeding females and breeding males. Let $c_{\gamma,A} = 1/\int_0^1 G_{s,h,A}(y)dy$. The rate of substitution as a function of selection s ,

$$K_{a,A}(s) = (\mu_f + \mu_m)(N_f U_{f,A}(s) + N_m U_{m,A}(s))$$

$$= (\mu_f + \mu_m) \left(N_f \frac{1}{4N_f} c_{\gamma,A} + N_m \frac{1}{4N_m} c_{\gamma,A} \right)$$

$$= \frac{\mu_f + \mu_m}{2} c_{\gamma,A}.$$

In particular, $K_{a,A}(s)$ does not depend on β and depends on s only through the product $\gamma = N_{e,A} s$.

Because, for the X chromosome, mutations carried in males must have arisen in females, the rate of substitution as a function of selection s . As before,

$$K_{a,X}(s) \approx (\mu_f + \mu_m) \frac{1}{3} c_{\gamma,\beta,X} + \mu_f \frac{1}{3} c_{\gamma,\beta,X}$$

$$\approx \frac{2\mu_f + \mu_m}{3} c_{\gamma,\beta,X}$$

where $c_{\gamma,\beta,X} = 1/\int_0^1 G_{s,h,\beta,X}(y)dy$. Thus, $K_{a,X}(s)$ depends on s and β only through the product

$$\gamma_X = \frac{9(\beta+1)}{8(\beta+2)} N_{e,A} s = N_{e,X} s.$$

For synonymous substitutions, the substitution probability is equal to the appropriate initial probabilities. Thus, for autosomes the rate of substitution,

$$K_{s,A} = (\mu_f + \mu_m) \left(N_f \frac{1}{4N_f} + N_m \frac{1}{4N_m} \right) = \frac{1}{2} (\mu_f + \mu_m),$$

consistent with segment spending equal amounts of time in males and females.

For the X chromosome, the rate of substitution

$$K_{s,X} = (\mu_f + \mu_m)N_f \frac{1}{3N_f} + \mu_f N_m \frac{1}{3N_m} = \frac{1}{3}(2\mu_f + \mu_m)$$

consistent with segment spending 2/3 of the time in females and 1/3 in males.

Notice that for the autosomes,

$$\frac{K_{a,A}(s)}{K_{s,A}} \approx c_{\gamma,A}$$

does not depend on the mutation rates or on β . For the X chromosome, this ratio

$$\frac{K_{a,X}(s)}{K_{s,X}} \approx c_{\gamma,\beta,X}.$$

Estimating the DFE

We followed the diffusion approximation-based approach of Boyko et al. (2008), first fitting demographic parameters to an AFS for putatively neutral sites, followed by fitting the DFE to an AFS for sites putatively under selection. In this article, we consider 4-fold sites as putatively neutral and 0-fold sites as putatively under selection. Although 4-fold sites may be under low levels of selection (particularly on the X chromosome [Lu and Wu 2005]), the low coverage sequencing at nontargeted intronic, pseudogene, or intergenic regions that may be considered more neutral will likely introduce substantial distortion of the underlying AFS due to variant call errors (Nielsen et al. 2011), especially in the singleton class. In addition, the use of 4-fold sites may be advantageous in that they contain information about the effects of genetic drift and background selection, and their use has been shown to lead to relatively unbiased estimates of α when fitting to a simple 2-epoch instantaneous growth demographic model (Messer and Petrov 2013). As a consequence, we fit a 2-epoch model to 4-fold sites for the autosomes and X chromosome separately as they likely possess different linkage properties (e.g., different ratios between the physical to genetic map lengths because of the lack of recombination on the X chromosome during male meiosis). All analyses were performed on folded frequency spectra. We also examined the possibility of utilizing the unfolded frequency spectrum by applying the trinucleotide context correction for ancestral state misidentification described by Hernandez et al. (2007). We do not report these results as the inference of demographic parameters was very sensitive to the divergence value used between humans and chimpanzees, and the method may be conservative when considering positive selection.

The diffusion approximation theory described above was incorporated into $\partial\alpha\partial t$ (Gutenkunst et al. 2009). We note that $\partial\alpha\partial t$ implements the fitness model as $2s$ for homozygous individuals rather than s , thus we estimate selection as $2N_e s$ rather than $N_e s$. In addition, we report the strength of purifying selection for the autosomes and X chromosome in units of N_{eA} , so that parameter estimates in terms of s for the

autosomes and X chromosome are directly comparable. It is trivial to convert the results for the X chromosome into units of N_{eX} using the expressions described above relating β to the relative effective population sizes of N_{eX} and N_{eA} . After testing these $\partial\alpha\partial t$ implementations against forward simulations generated in simuPOP (Peng and Kimmel 2005) under a variety of h , β , and $2N_{eA}s$ values assuming autosome and X chromosome inheritance patterns, $\partial\alpha\partial t$ was used to perform inference for the autosomes and the X chromosome separately by fitting the observed AFS for 4-fold and 0-fold sites to those generated under particular demographic or demographic + selective model parameters. The 2-epoch instantaneous growth model contains two parameters, n (ratio of contemporary to ancient population size, N_a), and T (time from the present when a size change occurred in units of $2N_a$ generations). When fitting the AFS at 4-fold sites to these parameters, we performed an initial two-dimensional grid search with n varying from 0.1 to 10.0 in steps of 0.1 and T varying from 0.0 to 2.0 in steps of 0.1. These parameters were further refined using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) optimization algorithm beginning from the best grid parameter set, with parameters $\log_{10}$ scaled and bounded between 0.001 to 10.0 for n , and 0 to 10.0 for T . The likelihood of the data given the model parameters was assessed using the multinomial approach described in the $\partial\alpha\partial t$ manual where the appropriate θ_4 is scaled using $\Sigma \text{data} / \Sigma \text{model}$ when comparing data to model. For the X chromosome, we performed this analysis independently for three fixed values of β : 1, 2.5, and 10, although in theory, T should scale proportionally to N_{eX}/N_{eA} when changing β .

The observed AFS at 0-fold sites was then fit to a model assuming n and T as inferred from 4-fold sites and a DFE for new mutations under a gamma distribution with two variable parameters, the shape, a , and scale, b , and where all selection was considered deleterious (i.e., negative s). Although other distributions are possible and incorrect specification of the shape of this distribution can lead to errors in downstream inference (Kousathanas and Keightley 2013), the gamma distribution has been shown to be a relatively good fit when examining the AFS in humans (Boyko et al. 2008; Keightley and Halligan 2011). The AFS for individual $2N_{eA}s$ values were estimated and stored between 2,000 $\log_{10}$ scaled points between -300 to -1×10^{-6} , as well as a point value of 0. The AFS for a given a and b was then calculated by trapezium numerical integration. Again we performed an initial two-dimensional grid search with $\log_{10}(a)$ varying from -4.0 to 1.5 in steps of 0.25 and $\log_{10}(b)$ varying from 1.0 to 6.0 in steps of 0.25 followed by further optimization using the BFGS algorithm bounded by the same range as the grid search. Unlike for fitting the 4-fold sites, we assumed a particular $\theta_{0\text{-fold}}$ given the estimated $\theta_{4\text{-fold}}$ estimate and the relative number of available sites. Explicitly, this was calculated by multiplying $\theta_{4\text{-fold}}$ by the number of 0-fold sites/number of 4-fold sites, the latter of which is approximately 4 depending on the data set (see table 1 for exact numbers). $\theta_{0\text{-fold}}$ was thus an explicit parameter in our model, and likelihood fit was assessed using the Poisson approach described in the $\partial\alpha\partial t$ manual. We repeated this analysis independently for the three fixed β values

for the X chromosome, as well as 21 fixed h values ranging from 0.0 to 1.0 in steps of 0.05 for both the autosomes and X chromosome (i.e., a total of 63 combinations of β and h for the X chromosome). Again, our estimate of $2N_{eAS}$ should scale proportionally to N_{eX}/N_{eA} .

Estimating α and ω

We follow the approach of Eyre-Walker and Keightley (2009) in estimating α and its related measure, ω , the rate of fixation of mutations under selection given the neutral rate (Schneider et al. 2011). To begin, let $\langle K_a \rangle^-$ be the expected values over the DFE for deleterious mutations. To estimate α , we subtract the expected number of substitutions per 0-fold site from d_0 , the observed number of substitutions per 0-fold site and normalizing by d_0 . To estimate ω , we normalize by d_4 , the observed number of substitutions per 4-fold site. The expected number of substitutions per 0-fold site is calculated by multiplying the mean fixation probability of a new deleterious mutation relative to a new neutral mutation $\langle K_a \rangle^- / K_s$ by d_4 , the number of 4-fold mutations entering the population since the split from a human–chimpanzee ancestral population. Thus,

$$\alpha = \frac{d_0 - d_4 \langle K_a \rangle^- / K_s}{d_0} \text{ and } \omega = \frac{d_0 - d_4 \langle K_a \rangle^- / K_s}{d_4}.$$

$\langle K_a \rangle^-$ is estimated using the fixation probabilities described above for individual $2N_{eAS}$, h , and β values integrated over the DFE, which is parameterized by the inferred values of a and b of the gamma distribution. For each value of s , the ratio $K_a(s)/K_s$ depends on N_e only through the value of $\gamma = N_e s$. Because both α and ω depend on an expected value of the ratio $K_a(s)/K_s$, this dependence N_e also holds for both α and ω . We do not report the results of α for $\beta = 2.5$ and 10.0, as these should be the same as for $\beta = 1$. This is because the estimated DFE, as parameterized by a and b , when combined with β is essentially an estimate of the distribution of $2N_{eX}s$. This value does not change in absolute terms when β changes, it only changes in terms of $2N_{eAS}$ proportionally to the change in N_{eX}/N_{eA} ratio (in fact once b is estimated for a given β it is possible to analytically determine how b will change with increasing or decreasing β , assuming a is fixed at some value). As $2N_{eX}s$ is constant, the level of purifying selection, $\langle K_a \rangle^-$, and hence α will also be constant regardless of β .

Confidence Intervals

As in Boyko et al. (2008), 95% CIs for n , T , a , b , α , ω were generated by resampling individual AFS entries and substitutions 400 times assuming a Poisson distribution, taking observed values as the mean. When variable sites are completely unlinked this is equivalent to CIs generated by parametric bootstrap. As the number of variable sites per gene is low, this assumption should be relatively robust (perhaps less so for the X chromosome due to the decrease in the overall recombination rate) though our CIs should be considered slightly anticonservative. As this process is computationally demanding, we only evaluated CIs at $h = 0.0, 0.25, 0.5, 0.75$,

and 1.0. In addition, we did not evaluate scale values, b , greater than 10^6 for any iteration of the bootstrap procedure, and (presumably because of its low data size) it is noticeable that estimates for the X chromosome are bounded at this value and thus may actually extend above it. Software implementing all steps of the inference of these parameters and associated CIs is available from the authors (K.R.V.) upon request.

Gene Content Matching with RNA-Seq

We downloaded the RPKM (Reads per kilo base per million) data generated as part of the Illumina Body Map 2.0 (Asmann et al. 2012), where RNA-Seq was performed across 16 different tissue types (fig. 3). Gene IDs were matched between this data set and the 1000 Genome High Coverage data set. For each gene, the max RPKM value was identified across all 16 tissues. If this max value was larger than two standard deviations from the mean of the RPKM values across all other tissues, the gene was assigned as tissue specific, otherwise it was assigned as nontissue specific. We also performed the same analysis using other criteria such as three standard deviations from the mean or for the max value to be two times the median (as in Kousathanas et al. 2013), but the results were largely the same, generally changing the number of nonspecific tissue genes, with the other tissue types decreasing accordingly at the same rate relative to each other. We then identified the proportion of genes within each of the 17 classes (16 tissue-specific classes and 1 nontissue-specific class) on the X chromosome. Ten thousand genes on the autosomes were randomly chosen, such that the proportion of genes from each class was the same as the X chromosome. The extended MK analysis was then performed on this data set as described above.

Forward Simulations

Forward simulations were performed using the software SLIM (Messer 2013). As in Messer and Petrov (2013), we simulated 10 MB of chromosome consisting of 250 genes containing 8 exons 150 bp in length, 7 introns 1,500 bp in length, a 5'-UTR 550 bp in length, and 3'-UTR 250 bp in length. However, rather than sampling at different times during the simulation to obtain 100 replicates, we performed 100 independent runs for a given set of parameters and sampled at the end of the run, as we also implemented a doubling of the population size 6,000 generations (150 ky assuming 25 years/generation) before present to match a demographic model previously inferred for the Yoruba using 1000 Genomes data (Gravel et al. 2011). We only sampled variants present in exons; 25% of sites in exons were assigned as synonymous. The remaining 75% were assigned as under selection (nonsynonymous). Under a pure purifying selection model, individual sites were assigned a negative selection coefficient drawn from a gamma distribution with shape, a , equal to 0.2 and mean s equal to -0.2 with $h = 0.5$. Under a model of purifying and positive selection, 1% of sites were assigned a positive selection coefficient s of 0.001 with $h = 0.5$. Though not directly sampled, UTRs were set with the same selection regime as for the purifying selection only model. A burn-in

period of $10N$ was set prior to the size change to allow the population to reach equilibrium. When simulating based on autosomal parameters, the ancestral N was set at 10,000 before doubling to 20,000, with a mutation rate of 2.5×10^{-8} per site per generation (Nachman and Crowell 2000). When simulating based on X chromosome parameters, the ancestral N was set at 7,500 before doubling to 15,000, with a mutation rate of $2.5 \times 10^{-8} \times 0.8 = 2 \times 10^{-8}$ per site per generation to take into account male-biased mutation. The value of 0.8 was estimated using relative autosome and X chromosome divergence as described previously (Miyata et al. 1987). We recognize that this setup does not fully simulate X chromosome evolution with sex-specific inheritance patterns and hemizygosity in males, but alternative forward simulators such as SFScode (Hernandez 2008) or simuPOP (Peng and Kimmel 2005) were unable to produce long linkage blocks as implemented in Messer and Petrov (2013) in a reasonable time-frame, whereas analysis using genes as single linkage blocks (and with all genes being independent) as performed previously (Boyko et al. 2008) are unlikely to accurately capture the true extent of linkage effects. For the coarse analysis, recombination rates were varied from 5×10^{-6} to 1×10^{-11} in steps of $0.5 \log_{10}$ units. For the finer scale analysis, recombination rates were varied from 1×10^{-9} to 5×10^{-9} in steps of 1×10^{-9} . Recombination rates were set as constant across the length of a chromosomal segment. For a particular parameter set (autosome or X chromosome, purifying selection or purifying selection + positive selection and a certain recombination rate), the 100 iterations were combined and an AFS created for the full diploid population, before being resampled down to 100 chromosomes (i.e., 50 diploid individuals) using the $\partial\alpha\partial i$ projection function. These AFS were then run through out extended MK pipeline as described above for 1000 Genomes data set.

Supplementary Material

Supplementary figures S1–S6 and tables S1–S3 are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

Acknowledgments

This work was supported by the National Institutes of Health to M.F.H. (R01_HG005226) and by the National Science Foundation to R.N.G. (DEB-1146074). The authors thank Michael Nachman for comments on the manuscript. They also thank the three anonymous reviewers whose comments and suggestions substantially improved both the robustness and clarity of the manuscript.

References

- 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56–65.
- Agrawal AF, Whitlock MC. 2011. Inferences about the distribution of dominance drawn from yeast gene knockout data. *Genetics* 187: 553–566.
- Asmann YW, Necela BM, Kalari KR, Hossain A, Baker TR, Carr JM, Davis C, Getz JE, Hostetter G, Li X, et al. 2012. Detection of redundant fusion transcripts as biomarkers or disease-specific therapeutic targets in breast cancer. *Cancer Res.* 72:1921–1928.
- Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR, et al. 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.* 4:e1000083.
- Charlesworth B, Coyne JA, Barton NH. 1987. The relative rates of evolution of sex chromosomes and autosomes. *Am Nat.* 130:113–146.
- Charlesworth J, Eyre-Walker A. 2008. The McDonald-Kreitman test and slightly deleterious mutations. *Mol Biol Evol.* 25:1007–1015.
- Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 6:80–92.
- Connallon T, Clark AG. 2013. Sex-differential selection and the evolution of X inactivation strategies. *PLoS Genet.* 9:e1003440.
- Connallon T, Singh ND, Clark AG. 2012. Impact of genetic architecture on the relative rates of X versus autosomal adaptive substitution. *Mol Biol Evol.* 29:1933–1942.
- Cox MP, Morales DA, Woerner AE, Sozanski J, Wall JD, Hammer MF. 2009. Autosomal resequence data reveal Late Stone Age signals of population expansion in sub-Saharan African foraging and farming populations. *PLoS One* 4:e6366.
- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27:2156–2158.
- Djureinovic D, Fagerberg L, Hallstrom B, Danielsson A, Lindskog C, Uhlen M, Ponten F. 2014. The human testis specific proteome defined by transcriptomics and antibody-based profiling. *Mol Hum Reprod.* 20: 476–488.
- Eyre-Walker A. 2002. Changing effective population size and the McDonald-Kreitman test. *Genetics* 162:2017–2024.
- Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol Biol Evol.* 26:2097–2108.
- Fay JC, Wu C-I. 2003. Sequence divergence, functional constraint, and selection in protein evolution. *Annu Rev Genomics Hum Genet.* 4: 213–235.
- Gottipati S, Arbiza L, Siepel A, Clark AG, Keinan A. 2011. Analyses of X-linked and autosomal genetic variation in population-scale whole genome sequencing. *Nat Genet.* 43:741–743.
- Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, 1000 Genomes Project, Bustamante CD. 2011. Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci U S A.* 108:11983–11988.
- Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5:e1000695.
- 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.
- 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.
- Han E, Sinsheimer JS, Novembre J. 2014. Characterizing bias in population genetic inferences from low-coverage sequencing data. *Mol Biol Evol.* 31:723–735.
- Hernandez RD. 2008. A flexible forward simulator for populations subject to selection and demography. *Bioinformatics* 24: 2786–2787.
- Hernandez RD, Williamson SH, Bustamante CD. 2007. Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Mol Biol Evol.* 24:1792–1800.
- Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genet Res.* 8:269–294.
- Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet.* 12:756–766.

- Hvilsom C, Qian Y, Bataillon T, Mailund T, Sallé B, Carlsen F, Li R, Zheng H, Jiang T, Jiang H, et al. 2012. Extensive X-linked adaptive evolution in central chimpanzees. *Proc Natl Acad Sci U S A*. 109: 2054–2059.
- Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* 177: 2251–2261.
- Keightley PD, Halligan DL. 2011. Inference of site frequency spectra from high-throughput sequence data: quantification of selection on non-synonymous and synonymous sites in humans. *Genetics* 188: 931–940.
- Kimura M. 1962. On the probability of fixation of mutant genes in a population. *Genetics* 47:713–719.
- Kirkpatrick M, Hall DW. 2004. Male-biased mutation, sex linkage, and the rate of adaptive evolution. *Evolution* 58:437–440.
- 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.
- Kousathanas A, Halligan DL, Keightley PD. 2013. Faster-X adaptive protein evolution in house mice. *Genetics* 196:1131–1143.
- Kousathanas A, Keightley PD. 2013. A comparison of models to infer the distribution of fitness effects of new mutations. *Genetics* 193: 1197–1208.
- Lercher MJ. 2003. Evidence that the human X chromosome is enriched for male-specific but not female-specific genes. *Mol Biol Evol*. 20: 1113–1116.
- Lohmueller KE, Degenhardt JD, Keinan A. 2010. Sex-averaged recombination and mutation rates on the X chromosome: a comment on Labuda et al. *Am J Hum Genet*. 86:978–980; author reply980–981.
- Lu J, Wu C-I. 2005. Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. *Proc Natl Acad Sci U S A*. 102:4063–4067.
- Mallet MA, Bouchard JM, Kimber CM, Chippindale AK. 2011. Experimental mutation-accumulation on the X chromosome of *Drosophila melanogaster* reveals stronger selection on males than females. *BMC Evol Biol*. 11:156.
- Mank JE, Vicoso B, Berlin S, Charlesworth B. 2010. Effective population size and the faster-X effect: empirical results and their interpretation. *Evolution* 64:663–674.
- Manna F, Martin G, Lenormand T. 2011. Fitness landscapes: an alternative theory for the dominance of mutation. *Genetics* 189: 923–937.
- McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351:652–654.
- Meisel RP, Connallon T. 2013. The faster-X effect: integrating theory and data. *Trends Genet*. 29:537–544.
- Messer PW. 2013. SLiM: simulating evolution with selection and linkage. *Genetics* 194:1037–1039.
- Messer PW, Petrov DA. 2013. Frequent adaptation and the McDonald–Kreitman test. *Proc Natl Acad Sci U S A*. 110:8615–8620.
- Miyata T, Hayashida H, Kuma K, Mitsuyasu K, Yasunaga T. 1987. Male-driven molecular evolution: a model and nucleotide sequence analysis. *Cold Spring Harb Symp Quant Biol*. 52:863–867.
- Nachman MW, Crowell SL. 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics* 156:297–304.
- Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol*. 3:e170.
- Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J. 2012. SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS One* 7:e37558.
- Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet*. 12: 443–451.
- Orr HA, Betancourt AJ. 2001. Haldane's sieve and adaptation from the standing genetic variation. *Genetics* 157:875–884.
- Peng B, Kimmel M. 2005. simuPOP: a forward-time population genetics simulation environment. *Bioinformatics* 21:3686–3687.
- Phifer-Rixey M, Bonhomme F, Boursot P, Churchill GA, Pialek J, Tucker PK, Nachman MW. 2012. Adaptive evolution and effective population size in wild house mice. *Mol Biol Evol*. 29:2949–2955.
- Pool JE, Nielsen R. 2007. Population size changes reshape genomic patterns of diversity. *Evolution* 61:3001–3006.
- Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, et al. 2013. Great ape genetic diversity and population history. *Nature* 499: 471–475.
- Ramsköld D, Wang ET, Burge CB, Sandberg R. 2009. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. Jensen LJ, editor. *PLoS Comput Biol*. 5:e1000598.
- Rice WR. 1984. Sex chromosomes and the evolution of sexual dimorphism. *Evolution* 38:735–742.
- Roach JC, Glusman G, Smit AFA, Huff CD, Hubley R, Shannon PT, 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.
- Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, Patterson N, Reich D. 2014. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* 507:354–357.
- Schmidt S, Gerasimova A, Kondrashov FA, Adzhubei IA, Kondrashov AS, Sunyaev S. 2008. Hypermutable non-synonymous sites are under stronger negative selection. Schierup MH, editor. *PLoS Genet*. 4:e1000281.
- Schneider A, Charlesworth B, Eyre-Walker A, Keightley PD. 2011. A method for inferring the rate of occurrence and fitness effects of advantageous mutations. *Genetics* 189:1427–1437.
- Vicoso B, Charlesworth B. 2009. Effective population size and the faster-X effect: an extended model. *Evolution* 63:2413–2426.
- Wilkie AO. 1994. The molecular basis of genetic dominance. *J Med Genet*. 31:89–98.
- Williamson SH, Fledel-Alon A, Bustamante CD. 2004. Population genetics of polymorphism and divergence for diploid selection models with arbitrary dominance. *Genetics* 168:463–475.