The Role of Phylogenetically Conserved Elements in Shaping Patterns of Human Genomic Diversity

August E. Woerner,[1,2] Krishna R. Veeramah,[3] Joseph C. Watkins,[4] and Michael F. Hammer*,[1]

[1]ARL Division of Biotechnology, University of Arizona, Tucson, AZ

[2]Center for Human Identification, University of North Texas Health Science Center, Fort Worth, TX

[3]Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY

[4]Department of Mathematics, University of Arizona, Tucson, AZ

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

Associate editor: John Novembre

Abstract

Evolutionary genetic studies have shown a positive correlation between levels of nucleotide diversity and either rates of recombination or genetic distance to genes. Both positive-directional and purifying selection have been offered as the source of these correlations via genetic hitchhiking and background selection, respectively. Phylogenetically conserved elements (CEs) are short (~100 bp), widely distributed (comprising ~5% of genome), sequences that are often found far from genes. While the function of many CEs is unknown, CEs also are associated with reduced diversity at linked sites. Using high coverage (>80×) whole genome data from two human populations, the Yoruba and the CEU, we perform fine scale evaluations of diversity, rates of recombination, and linkage to genes. We find that the local rate of recombination has a stronger effect on levels of diversity than linkage to genes, and that these effects of recombination persist even in regions far from genes. Our whole genome modeling demonstrates that, rather than recombination or GC-biased gene conversion, selection on sites within or linked to CEs better explains the observed genomic diversity patterns. A major implication is that very few sites in the human genome are predicted to be free of the effects of selection. These sites, which we refer to as the human "neutralome," comprise only 1.2% of the autosomes and 5.1% of the X chromosome. Demographic analysis of the neutralome reveals larger population sizes and lower rates of growth for ancestral human populations than inferred by previous analyses.

Key words: **phylogenetic conserved elements, recombination, background selection, genetic hitchhiking, diversity, null model.**

Introduction

Recombination is a fundamental force of molecular evolution. Nearby sites on the chromosome are likely coinherited, resulting in correlated evolutionary trajectories. This linkage disequilibrium (LD) between sites reduces the efficiency of natural selection, decreasing fixation rates for positively selected alleles while increasing rates for negatively selected alleles (Hill and Robertson 1966). Neutral alleles in LD with selected alleles can hitchhike to fixation under a selective sweep (Maynard Smith and Haigh 1974) or to extinction under background selection (Charlesworth et al. 1993). Both background selection and hitchhiking (herein termed linked selection) distort the underlying ancestral recombination graph (ARG), increasing terminal branch lengths, altering allele frequency spectra (AFS) and generally reducing genetic diversity (Braverman et al. 1995; Tachida 2000; Williamson and Orive 2002; O'Fallon et al. 2010; Nicolaisen and Desai 2012, 2013).

Recombination mitigates the effects of linked selection by decoupling selected alleles from the genomic background. In their landmark study, Begun and Aquadro (1992) found that the rate of recombination (*R*) is significantly correlated with nucleotide diversity (π) in *Drosophila*, a finding that has been replicated and expanded upon in a variety of species (Nachman 2001; Spencer et al. 2006; Cai et al. 2009; Lohmueller et al. 2011; McGaugh et al. 2012). Similarly, human diversity is also positively correlated with the minimum genetic distance to genes (*G*) (Hammer et al. 2010; Gottipati et al. 2011; Prado-Martinez et al. 2013; Arbiza et al. 2014). The implication of these bodies of work is that linked selection plays a major role in constraining diversity across the genome. Cellular processes such as GC-biased gene conversion (Marais 2003) or the mutagenicity of recombination (Pratto et al. 2014; Arbeithuber et al. 2015; Francioli et al. 2015) may also contribute to these correlations (but see McGaugh et al. 2012).

An important question that has not been carefully addressed in humans is how much of the effects of linked selection is driven by evolutionarily constrained regions within and outside of genes. Phylogenetically conserved elements are loci that have far fewer substitutions than would otherwise be expected in a neutral region over a given phylogeny. Classes of phylogenetically conserved elements include ultraconserved elements, which within humans are defined as having 100% identical orthologs between humans, mouse, and rat (Bejerano et al. 2004). More relaxed definitions of conserved elements, such as *phastCons* elements (Siepel et al. 2005), permit low levels of nucleotide substitution. Conserved elements are not simply mutational coldspots, but instead appear selective (Bejerano et al. 2004; Cooper et al. 2005; Siepel et al. 2005; Chen et al. 2007; Katzman et al. 2007; McVicker et al. 2009; Halligan et al. 2011). In addition, diversity is reduced within and adjacent to both coding and noncoding *phastCons* elements (Hernandez et al. 2011; Halligan et al. 2013). Unlike genes, conserved elements are generally short (~100 bp) and widely distributed sequences (comprising ~5% of the genome), suggesting that linked selection to phylogenetically conserved elements may also play role in constraining diversity across the genome.

We investigate the relationships between π, *R*, and *G* at genic and nongenic sites across the genome, making use of high coverage whole genome sequence data from two human populations: the Yoruba from Ibadan, Nigeria, and Northern Europeans from Utah. We test the hypothesis that recombination is mutagenic within the nonrepetitive portions of the genome that we consider in our estimates of diversity. We also assess the extent to which GC-biased gene conversion contributes to the relationships among diversity, *R* and *G*. By modeling linkage to phylogenetically conserved sequences, we assess their role in shaping genomic patterns of diversity, and then validate this model by testing for linked selection in regions predicted to be unlinked to *phastCons* elements. This leads to a test of "neutrality" for genomic regions.

Results

Diversity and Linkage on the X Chromosome and the Autosomes

To prepare a data set to estimate local genetic diversity, we partitioned the human genome into nonoverlapping 10-kb loci (see Supplementary Material for justification of locus size). After masking out several classes of repeats, we computed nucleotide diversity (π) and divergence (*D*) to a human-orang ancestor sequence within each locus for both the Yoruba (YRI, $n = 9$) and Northern Europeans (CEU, $n = 9$) based on high coverage (~80×) whole genomes. Next, we assessed the effects of linkage on diversity at each locus as a function of *G*, the minimum genetic distance to genes in centimorgans (cM), and *R*, the length of the locus in cM (supplementary fig. S1, Supplementary Material online).

To visualize how *G* and *R* jointly affect π/D (π/D is an estimator of effective population size, with the division by *D* serving to control for variation in the local mutation rate; see Charlesworth 2009 for review), we bin loci using the marginal distributions of both *G* and *R*, by deciles in the autosomes and quartiles on the X chromosome. For each bin, we display the median value of π/D (fig. 1 and supplementary fig. S2, Supplementary Material online). As expected, we find that diversity increases with the genetic distance to genes (left to right columns), consistent with previous work (Hammer et al. 2010; Gottipati et al. 2011; Arbiza et al. 2014). Surprisingly, π/D also increases with *R* (bottom to top rows) for both types of chromosomes regardless of the value of *G*. This gradient is even evident in the last column of figure 1, which is composed of loci ≥0.46 cM from genes on the autosomes and ≥0.13 cM from genes on the X chromosome. The expected disequilibrium coefficient r^2 is quite small between these loci and the nearest gene. Taking the expected value of *r* to be $N_e\mu$ (assuming conservative values; $N_e = 10{,}000$ and an inheritance scalar of 3 on the X chromosome), these expected values are <0.005 on the autosomes and < 0.025 on the X chromosome. Thus, the last column of figure 1*a* shows the case of loci essentially unlinked from genes. Both π (supplementary fig. S3, Supplementary Material online) and *D* (supplementary fig. S4, Supplementary Material online) examined separately show the same general trends, though with a much more pronounced effect for π.

To assess statistical significance of the observed trends, we used iterated reweighted least squares (IRLS) regression on the model $\pi/D \sim G + R$ for the autosomes (*A*) and the X chromosome (*X*) separately, with estimated standardized slope coefficients β providing a measure of effect size (see Supplementary Material). Diversity on the X chromosome was scaled both by 0.75 (i.e., assuming a 1:1 sex ratio) and by the more conservative value of 0.95 (approximating the upper bound of Arbiza et al. 2012). All β are significantly >0 ($P < 0.001$ across all comparison). Further they satisfy $\beta_{XG} > \beta_{XR} > \beta_{AR} > \beta_{AG}$ ($P < 0.001$ across all comparisons in the YRI, $P < 0.05$ for the CEU) (Materials and Methods, table 1). The slope on the autosomes for *R* (β_{AR}) is significantly larger than that for *G* (β_{AG}). The reverse inequality holds on the X chromosome, with both coefficients being larger on the X chromosome than on the autosomes. We note that $\beta_{XG} > \beta_{AG}$ is the primary finding of Hammer et al. (2010), where the interpretation is that linked selection on genes reduces diversity more rapidly on the X chromosome than on the autosomes.

The Correlation between *R* and π/D

There are three possible explanations for the observed positive correlation between *R* and π/D: 1) recombination is itself mutagenic (Pratto et al. 2014; Arbeithuber et al. 2015; Francioli et al. 2015), 2) GC-biased gene conversion (gBGC) may influence allele frequencies (Marais 2003), or 3) selection at noncoding (e.g., regulatory) sites are having a significant impact on local genetic diversity (recall that we only consider the distance from genes in the analysis above). Below, we examine the evidence for each.

The Local Recombination Rate of De Novo Mutations

A direct measure of local mutation rate can be obtained by examining the distribution of de novo mutations identified via whole genome sequencing of families. We obtained two large genomic data sets consisting of 4,917 (Kong et al. 2012) and 11,010 (Francioli et al. 2015) such mutations, which we refer to as the Decode and GoNL data sets, respectively. To mirror our estimates of diversity, we masked repeats (see Supplementary Material), leaving 1,467 Decode and 2,166 GoNL mutations in the nonrepetitive portions of the genome. We then tested if the recombination rate at sites

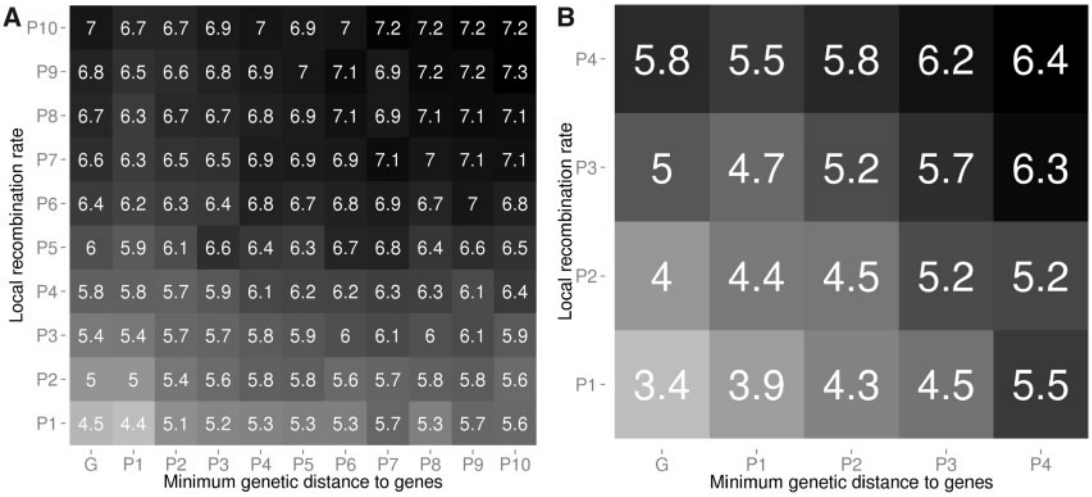

Fig. 1. Heatmap of the median π/D ($\times 100$) in YRI. The *x*-axis shows the minimum genetic distance to genes and the *y*-axis the local recombination rate. Darker cells correspond to higher π/D. Each cell corresponds to a pair of percentiles (P) in the distance to genes and the local recombination rate, with column G corresponding to loci in genes, and the remaining columns being outside of genes. (*a*) Autosomes. (*b*) X chromosome.

Table 1. Beta Coefficients and Confidence Intervals Inferred from Modeling $\pi/D \sim R + G$ on the X Chromosome (X) and the Autosomes (A).

NeX/NeA	Population	Coefficient	β coefficient from IRLS Regression	Lower CI (0.025)	Upper CI (0.975)	*P* value (between rows)
0.75	YRI	X_G	0.32	0.28	0.36	<0.001
0.75	YRI	X_R	0.15	0.12	0.19	<0.001
0.75	YRI	A_R	0.07	0.07	0.08	<0.001
0.75	YRI	A_G	0.04	0.03	0.04	
0.75	CEU	X_G	0.20	0.16	0.24	<0.001
0.75	CEU	X_R	0.10	0.07	0.13	0.002
0.75	CEU	A_R	0.05	0.05	0.06	<0.001
0.75	CEU	A_G	0.03	0.03	0.03	
0.95	YRI	X_G	0.26	0.23	0.29	<0.001
0.95	YRI	X_R	0.12	0.10	0.15	<0.001
0.95	YRI	A_R	0.07	0.07	0.08	<0.001
0.95	YRI	A_G	0.04	0.04	0.04	
0.95	CEU	X_G	0.16	0.13	0.19	0.001
0.95	CEU	X_R	0.08	0.05	0.10	0.028
0.95	CEU	A_R	0.05	0.05	0.06	<0.001
0.95	CEU	A_G	0.03	0.03	0.04	

Note.—Within a population the beta coefficients are ordered, and the *P* value is given comparing a given coefficient to the coefficient in the row that is given in the table.

with de novo mutations exceeded that of the genomic average.

To control for the effect of sequence context on mutation rate (e.g., enrichment of mutation rates at CpG sites, which in turn are generally enriched in recombination hotspots; Nachman 2001) we generated a null distribution by extracting genome-wide sites that matched the trinucleotide distribution of the de novo mutations, yielding 67,394,212 and 60,392,816 sites in our Decode data set and GoNL data set, respectively.

For nonrepetitive regions the Decode de novo data set had a mean local recombination rate of 1.48 cM/Mb (SD: 4.90), while its matching null set had a mean of 1.47 cM/Mb (SD: 5.06). The GoNL de novo data set had a mean local recombination rate of 1.57 cM/Mb (SD: 5.53), while its matching null set had a mean of 1.47 cM/Mb (SD: 5.06). Pooling the data sets yielded a mean de novo local recombination rate of 1.53 cM/Mb (SD: 5.06), which was not significantly different than the pooled null sets ($P = 0.48$, 2-tailed Welch *t*-test). Repeating this procedure on the whole genome, including repetitive sequences but limited to the range of the HapMap genetic map, yielded a mean local recombination rate of de novo mutations of 1.54 cM/Mb (SD: 5.56) (1.51 cM/Mb SD: 5.38 for Decode, 1.56 cM/Mb SD: 5.63 for GoNL). The pooled trinucleotide-matching null set for Decode ($n = 227{,}640{,}296$) and GoNL ($n = 206{,}352{,}696$) had a mean local recombination rate of 1.37 cM/Mb ($SD = 4.72$), which is significantly different from the combined set of 15,927 (4,917 Decode and 11,010 GoNL) de novo mutations ($P = 0.00013$, 2-tailed Welch *t*-test). Thus, the expected local recombination rate of de novo mutations is significantly greater than that of the genomic average suggesting that recombination is associated with mutagenesis, consistent with previous works (Pratto et al. 2014; Arbeithuber et al. 2015;

Francioli et al. 2015). However, when constrained to the nonrepetitive portions of the genome, this effect no longer has significant statistical support.

GC-Biased Gene Conversion

Another cellular process that may induce a positive correlation between diversity and R (fig. 2) is GC-biased gene conversion (gBGC). With gBGC, heterozygous genotypes near recombination-initiated double-strand breaks are preferentially converted from weakly bonded (W: A, T) to strongly bonded (S: G, C) states (Strathern et al. 1995; Marais 2003). This GC-biased DNA repair is posited to create a fixation bias whose effects have been shown in the AFS for the populations considered here (Katzman et al. 2011). As gBGC induces fixation biases for S→W and W→S mutations, we performed our heatmap and regression analyses limited to S→S and W→W transversions, which are gBGC-independent. Visualizations of the joint effects of R and G on π/D under this condition largely recapitulate our original findings (fig. 3 and supplementary fig. S5, Supplementary Material online), though the exclusion of so many variable sites increases the variance in our estimate of π/D. As before, IRLS regression analysis allows us to infer that all slope coefficients were significantly positive in both populations on both the X chromosome and the autosomes ($P < 0.001$), and that $\beta_{XR} > \beta_{AR} > \beta_{AG}$ in the YRI ($P < 0.001$ and in the CEU $P < 0.02$, supplementary table S1, Supplementary Material online). Curiously, in both the YRI and the CEU, we could not conclude that $\beta_{XG} > \beta_{XR}$. This may be due to reduced power as a result of considering ~1/6 the number of variable sites. We also note that our consideration of only S→S and W→W transversions serves as an additional control on the mutagenic effects of recombination, which is believed to be restricted to transitions (Arbeithuber et al. 2015). Overall, we conclude that neither gBGC nor the mutagenicity of recombination are driving the trends seen in figure 1.

The Linkage to *phastCons* Elements

If selection is driving the trend between π/D and R, it follows that this trend must largely stem from selection at nongenic sites. To evaluate this hypothesis, we examined 725,430 phylogenetically conserved *phastCons* elements inferred over a primate phylogeny downloaded from the UCSC genome database. Primate *phastCons* elements are short (Mean: 145 bp, SD: 141 bp) genomic regions that have a marked lack of substitutions over the given phylogeny. Even though only ~27% of these elements overlap coding regions, they have long-term selective effects. We first computed the (minimum) genetic distance of each of our 10-kb loci to conserved exonic (*phastCons*) elements (CEEs) and to conserved nonexonic (*phastCons*) elements (CNEs) in the YRI. To evaluate the effects of linkage to these elements, we removed loci that overlapped either CNEs or CEEs, leaving 58,314 and 5,365 loci on the autosomes and X chromosome, respectively. As *phastCons* elements are so numerous, nearly all of the remaining loci were extremely close to at least one conserved element (see McVicker et al. 2009), with loci on the autosomes and the X chromosome having a mean minimum distance of 0.0067 cM (SD: 0.0261) and 0.0093 cM (SD: 0.0264), respectively. We then used IRLS regression to model the relative effects of linkage using the minimum genetic distances to CEE elements (D_{CEE}) and to CNE elements (D_{CNE}). We modeled $\pi/D \sim D_{CEE} + D_{CNE}$ separately for the X chromosome and the autosomes. Consistent with previous works (Hernandez et al. 2011; Halligan et al. 2013), π/D is positively correlated with both D_{CEE} and D_{CNE} ($P < 0.001$, nonparametric bootstrap). Furthermore, our slope coefficient estimates are large for both the autosomes ($\beta_{DCNE} = 0.057$, 95CI: 0.048–0.066 and $\beta_{DCEE} = 0.057$, 95CI: 0.052–0.065) and the X chromosome ($\beta_{DCNE} = 0.128$, 95CI: 0.116–0.154 and $\beta_{DCEE} = 0.175$, 95CI: 0.119–0.200). As a reference, when constrained to just nongenic loci that also do not overlap *phastCons* elements, modeling $\pi/D \sim G + R$ yields estimated slope coefficients of $\beta_G = 0.041$, $\beta_R = 0.069$ and $\beta_G = 0.192$, $\beta_R = 0.079$ for the autosomes and X chromosome, respectively. Thus, the effect sizes (slope coefficients) for linkage to *phastCons* elements are approximately the same size as that of either linkage to genes or for the local rate of recombination (also see table 1), perhaps suggesting a comparable level of importance between linkage to *phastCons* elements and the trends apparent in figure 1. The coefficients β_{DCNE} and β_{DCEE} were not significantly different from each other within the X chromosome ($P = 0.37$, nonparametric bootstrap) or within the autosomes ($P = 0.81$, nonparametric bootstrap), suggesting that linkage to the nearest CNE and CEE have approximately equally impacts on diversity, though CNEs are roughly three times more common than CEEs.

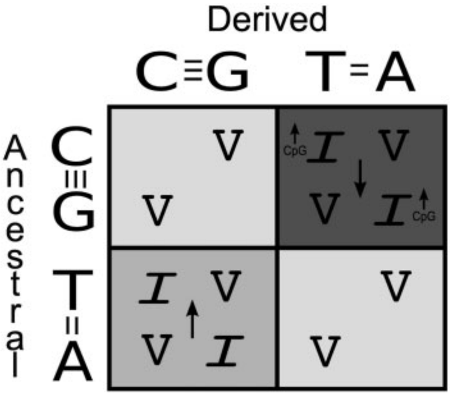

FIG. 2. The role of base composition on different mutation types. GC-biased gene conversion (gBGC) increases the fixation probability of strongly bonded base pairs (S, G or C) over weakly bonded base pairs (W, T or A). Thus, gBGC can increase diversity (upward arrow, bottom left) or decrease diversity (downward arrow, top right) for transitions (I) and transversions (V) alike. Similarly, recombination-induced mutagenesis is almost entirely exclusive to CpG mutations, which are S→W transitions (Arbeithuber et al. 2015). To avoid both phenomena, we repeated our experiments using the beige class of S→S and W→W transversions (beige).

Modeling Linkage to Conserved Elements

We hypothesize that conserved elements are the primary drivers of the trends seen in figure 1. One way to assess this hypothesis is to identify regions of the genome that are

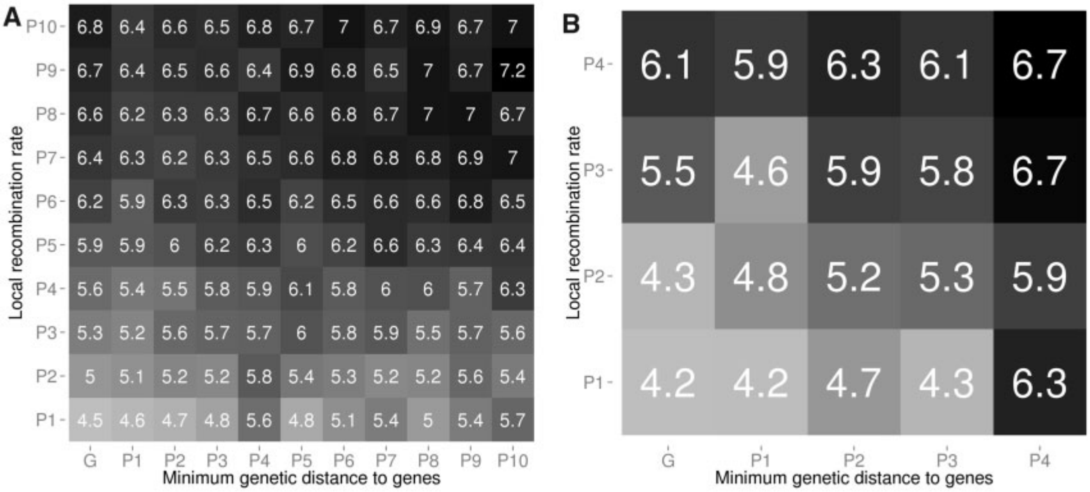

Fig. 3. Heatmap of the median π/D ($\times 100$) in YRI considering only W→W and S→S transversions. The *x*-axis shows the minimum genetic distance to genes and the *y*-axis the local recombination rate. Darker cells corresponding to higher π/D. Each cell corresponds to a pair of percentiles (P) in the distance to genes and the local recombination rate, with column G corresponding to loci in genes, and the remaining columns being outside of genes. (*a*) Autosomes. (*b*) X chromosome.

completely unlinked to *phastCons* elements and assess if the trends in figure 1 are still apparent. This solution, however, is lacking as most genomic bases reside in or near a conserved element (McVicker et al. 2009), which further suggests that thresholding on a minimum genetic distance to *phastCons* elements is unlikely to yield large enough distances and/or sufficient numbers of loci to adequately explore this hypothesis. A more feasible approach is to model the effects of linked selection on diversity across every nucleotide in the human genome; thus not only would minimum genetic distances be considered but the cumulative effects of all putatively selective sites. Loci predicted to be (nearly) unlinked from *phastCons* elements could then be tested for, say, correlations between diversity and recombination as a means to assess our original hypothesis.

We used the approach of Halligan et al. (2013) (their Model C) to model the effects of linkage to CEE and CNE elements throughout the human genome. Unlike other genomic models of linked selection (McVicker et al. 2009; Elyashiv et al. 2016), which are directly derived from theory on background selection and genetic hitchhiking (Hudson and Kaplan 1995; Nordborg et al. 1996; Barton 1998; Gillepsie 2000), Model C is a phenomenological model that describes the diversity trough apparent around putative sites of selection, making it perhaps more robust to the violation of assumptions on the modes, strengths, and types of selection than other more heavily parameterized approaches (Elyashiv et al. 2016). Further, Model C is especially appropriate for the inference of neutral sequence as its predictions need only be accurate for describing a lack of linked selection (i.e., when there is little predicted linked selection).

In Model C diversity at linked sites decays exponentially around conserved elements, and diversity at a given site is taken as a product over all linked selective sites (Materials and Methods). Model C predicts diversity at every locus considering all *phastCons* elements within 1 cM of that locus, and we normalize this prediction into ν (Materials and Methods). In other words, the value of ν is a measure of the local reduction in π/D explained by linkage to nearby *phastCons* elements, with lower values indicating lower (expected) diversity.

The mean value of ν on the X chromosome is 0.85 (SD: 0.13), which is significantly smaller than the mean value on the autosomes (0.89, SD: 0.12, $P < 2.2\text{e-}16$, Welch two-tailed *t*-test). To evaluate the overall distribution of ν, the empirical cumulative distribution function of ν was computed on both the autosomes and the X chromosome (fig. 4). Consistent with the estimates of the mean ν, the X chromosome appears to generally have a lower ν (e.g., ~25% of X chromosome loci have a $\nu < 0.75$, while this is true of only ~11% of autosomal loci), suggesting more selective constraint on the X chromosome. This is consistent with previous findings (McVicker et al. 2009; Hammer et al. 2010; Gottipati et al. 2011; Veeramah et al. 2014).

Determining Sites Unaffected by Selection

We developed an approach to determine a cutoff for the minimum value for ν that removes the effects seen in figure 1 under the simplifying assumption that linked selection acts solely to reduce diversity. Specifically, at this cutoff, we lose all statistically significant positive correlations between any pair of choices among π/D, and R, G, and ν. This can be framed as an approach to the problem of finding so-called "neutral" sequence, that is, those loci where the distribution of allele frequencies are dictated solely by the rate of mutation and genetic drift and not by any direct or indirect (i.e., linked) effects of selection. We term the total set of all neutral loci in a genome the neutralome. We note that a lack of correlation between R, G, (and even ν), and π/D is a necessary condition of a sequence to be neutral.

Using statistical tests of both linear and of monotonic relationships, we find that thresholding on $\nu \geq 99\%$ on the X chromosome and $\nu \geq 99.9\%$ on the autosomes is sufficient to remove any significant correlations between

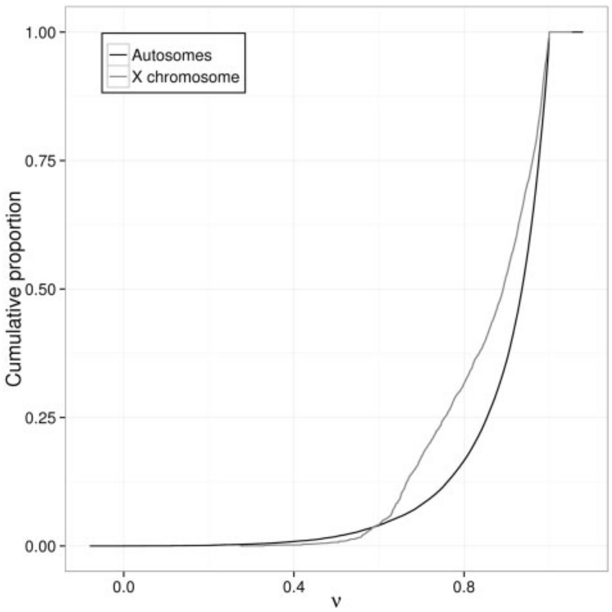

Fig. 4. The empirical cumulative distribution function (ECDF) of ν. ν is a prediction of nucleotide diversity relative to a neutral rate of 1.0. The predictions are off of nonlinear least squares fitting of observed diversity levels to linkage to *phastCons* elements from Model C of Halligan et al. (2013). In general, the X chromosome (black) shows lower ν (consistent with more constraint) than the autosomes (gray).

diversity, *R*, *G*, and ν (Materials and Methods). These results persist both when the data from the autosomes and the X chromosome are evaluated separately, and when they are pooled. Pooling increases our power to detect correlations while mitigating concerns about the smaller sample size of the X chromosome. Visual inspection of relationship between π/D and *R* further supports our assertion that the neutralome removes not just the significance, but also the trends seen in figure 1 (fig. 5 and supplementary figs. S6–S8, Supplementary Material online). The neutralome's size, by genomic standards, is quite small, with only 3,606 and 787 10-kb loci appearing neutral on the autosomes and X chromosome, respectively.

The AFS and Estimates of Θ in Different Genomic Regions

To evaluate the properties of our neutralome, we compared it against regions predicted to be neutral by Neutral Region Explorer (NRE) (Arbiza et al. 2012). NRE finds neutral regions parametrically, allowing the user to select specific criteria (e.g., thresholds on the minimum genetic distance to genes, local recombination rate, and background selection coefficients; McVicker et al. 2009) that may yield regions unaffected by selection. We downloaded regions identified as neutral by NRE using all parameters set to their defaults (background coefficient >0.95, > 0.4 cM from genes, >0.9 cM/Mb). We considered all 10-kb loci from the YRI that overlapped the genomic coordinates identified as neutral by NRE yielding a candidate set of 6,140 NRE loci (276 on the X chromosome). In total 659 loci (154 on the X chromosome) were both identified by NRE and are in the neutralome. As the size of the NRE set was larger than the neutralome, and perhaps may be less "neutral" as a result, a second NRE set (NRE, bg97) was constructed by further constraining the background selection coefficient of McVicker et al. (2009) to a minimum of 0.97. This more restricted set is composed of 4,212 loci (164 on the X chromosome), which shares 533 loci with the neutralome (92 on the X chromosome).

For all sets of loci, we computed two estimates of the population mutation rate, θ_W and π, dividing each by divergence to correct for mutation rate heterogeneity, as well as Tajima's *D* (1989), Fu and Li's *D* (1993), and Thomson's estimator of the time to the most recent common ancestor (TMRCA) (Hudson 2007). The means of each of these parameters were compared using 2-tailed Welch *t*-tests.

The neutralome has significantly higher mean autosomal π/D (0.083) than loci identified by NRE using the default parameters (0.076, $P < 1\text{e-}18$) or with the more conservative NRE, bg97 set (0.077, $P < 1\text{e-}12$), though no significant differences were found on the X chromosome (fig. 6 and supplementary table S3, Supplementary Material online). The lack of significance on the X chromosome may again in part reflect a lack of power due to a limited sample size. The neutralome also has significantly higher autosomal Tajima's *D* ($P < 1\text{e-}4$) and Fu and Li's *D* ($P < 1\text{e-}10$) (fig. 7) compared with the NRE set, indicating an excess of rare variants in the NRE loci compared with the neutralome. When constrained to the NRE, bg97 set Tajima's *D* becomes marginally significantly different ($P = 0.054$), while Fu and Li's *D* remains significantly different ($P < 1\text{e-}10$). Lastly, we see a striking reduction in TMRCA for both sets of NRE loci ($P < 1\text{e-}20$ for NRE, $P < 1\text{e-}13$ NRE, bg97). This reduction may stem from positive correlations between TMRCA and θ.

To assess if the neutralome impacts demographic inference, we fit the AFS to a 2-epoch instantaneous growth model using ∂a∂I (Gutenkunst et al. 2009) on the autosomes in the YRI. We estimated the ancestral population size (Na), the growth rate (*n*), and the time of the start of growth (*T*) assuming a mutation rate of 2.5×10^{-8} (Nachman and Crowell 2000). These parameters were estimated for both NRE sets, the neutralome, and for a reference, with 4-fold degenerate sites. Further, several filtering and scaling adjustments were made in this process. Models were fit considering all segregating sites, and again using sites inconsistent with CpG mutations (table 2). To ensure that these inferences were not driven solely by differences in θ, the number of callable bases for the NRE set and the NRE, bg97 set were scaled by the ratio of mean π/D compared with that of the neutralome (9.23% for NRE, 7.7% for NRE, bg97, Theta scaling, Materials and Methods, table 2).

When considering all sites, Na was significantly higher in the neutralome than in regions found by NRE and in 4-fold degenerate sites (table 2, bootstrapped $P < 0.05$ across all comparisons). Excluding CpGs (removing ∼1/3 of all sites) removed all significant effects with respect to the NRE loci, though Na was marginally significantly different between both NRE sets and the neutralome. Consistent with our analysis of Fu and Li's *D*, *n* in the neutralome was significantly less than the NRE loci (marginally so for the NRE, bg97 loci) across

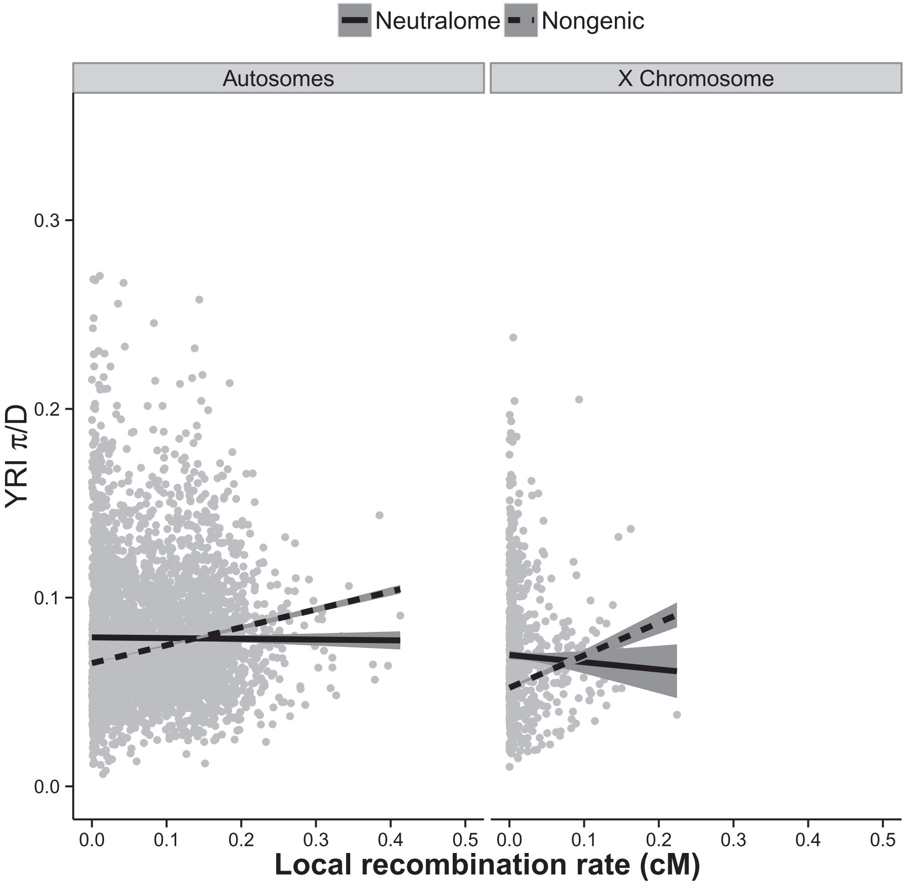

Fig. 5. Scatterplot of YRI π/D versus the local recombination rate in the neutralome. The gray dots represent loci in the neutralome in the autosomes (left) and X chromosome (right). Each line represent the IRLS regression line with 95% CI fit to the neutralome (solid) and the nongenic portion of the genome (dashed).

the same comparisons (table 2), consistent with an excess of rare variants in the NRE loci and in 4-fold degenerate sites. The time of growth (T) varied across our choices of "neutral" samples, though only rarely were the differences to the NRE loci significant. Attempting to correct for differences in the underlying θ removed all significant effects save for n in the original NRE data set (though Na in the NRE data set obtained marginal significance at $P = 0.055$). As the inference on n is independent of the number of (scaled) bases (and thus is invariant to the correction), this perhaps points to differences in the population mutation rate driving many of the differences in the estimated parameters.

Discussion

The Pervasive Effects of Linked Selection

Using high coverage whole genomes from two human populations, we show that even if only ~5–15% of sites in the human genome are directly targeted by selection (Chinwalla et al. 2002; Cooper et al. 2005; Siepel et al. 2005; Meader et al. 2010; Ponting and Hardison 2011; Rands et al. 2014; Schrider and Kern 2017), the indirect action of selection at linked sites causes genetic diversity in up to 99% of the genome to be demonstrably reduced from neutral expectations. On genomic scales, the effects of linked selection up to this point have solely been framed either with respect to how linked a particular locus is to the nearest genic source of selection (Hammer et al. 2010; Gottipati et al. 2011; Arbiza et al. 2014), or with correlations between diversity, gene-density and rates of recombination (Cai et al. 2009; Lohmueller et al. 2011). We examined the dependence of π/D against two genomic measures sensitive to selection at linked sites, the minimum genetic distance to genes and the local rate of recombination, and show that genetic diversity is influenced by both measures at fine scales throughout the genome (fig. 1). While both R and G are positively correlated with diversity, the effect of R is far greater than that of G on the autosomes (table 1). This effect is neither driven by GC-biased gene conversion nor the mutagenicity of recombination (fig. 3). Instead we find that these patterns are primarily caused by linked selection at nongenic sites, an argument that is bolstered when we consider the sizable effect of R on diversity in loci that are generally unlinked from all genic sites of selection (the last column of fig. 1*a*).

Our hypothesis of the effect of linked selection at both genic and nongenic loci is consistent with other aspects of our findings. First, linked selection may have sizable impacts on diversity, while it has at most a modest effect on patterns of divergence (Birky and Walsh 1988; but see McVicker et al. 2009). Consistent with this both π and D are correlated with R and G, with the effect much stronger for π than D (supplementary figs. S3 and S4, Supplementary Material online).

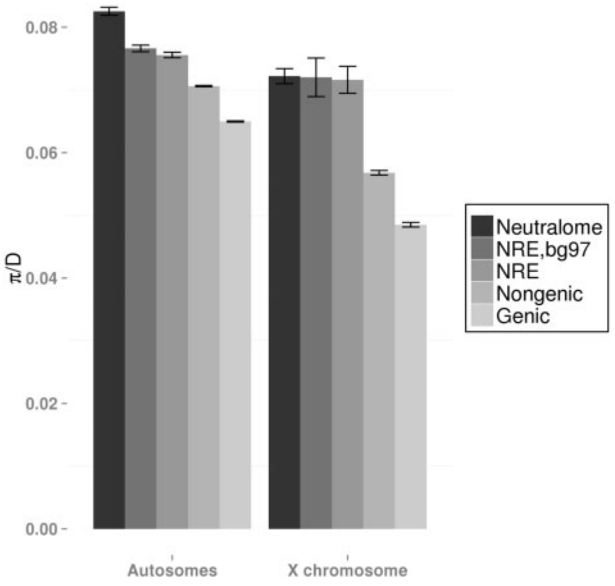

Fig. 6. Mean π/D (± SEM) in several regions of the autosomes (left) and the X chromosome (right). NRE loci are regions predicted to be neutral by Neutral Region Explorer, while the neutralome refers to loci identified in this study. The NRE loci considered were obtained either using the default settings (NRE, with a minimum background selection coefficient of 0.95), or with a minimum background selection coefficient of 0.97 (NRE, bg97). Genic and nongenic loci are also included for comparison purposes. Lower π/D is consistent with higher levels of background selection and/or genetic hitchhiking.

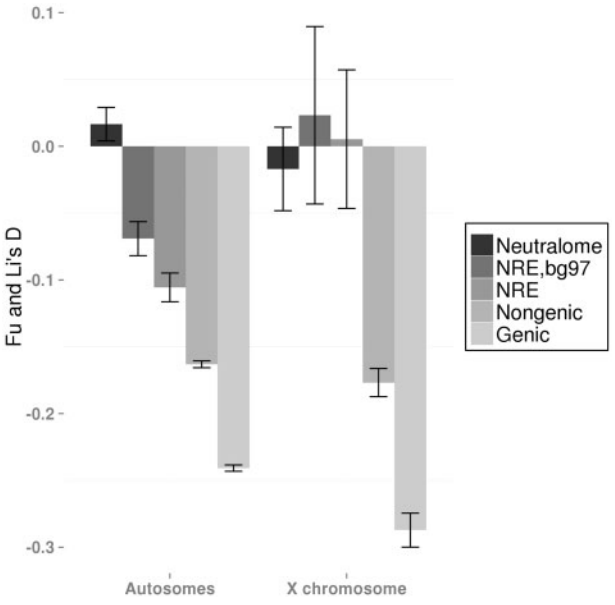

Fig. 7. Mean Fu and Li's D (± SEM) in several regions of the autosomes (left) and the X chromosome (right). NRE loci are regions predicted to be neutral by Neutral Region Explorer, while the neutralome refers to loci identified in this study. The NRE loci considered were obtained either using the default settings (NRE, with a minimum background selection coefficient of 0.95), or with a minimum background selection coefficient of 0.97 (NRE, bg97). Relative to a neutral model with no growth, negative Fu and Li's D signifies an excess of singletons, while positive values indicate a reduction in singletons. Weak background selection and genetic hitchhiking may both lead to negative Fu and Li's D, as will demographic effects such as population growth.

Examination of these patterns on the X chromosome versus the autosomes also supports our linked-selection hypothesis. As the X chromosome is exposed to selection in males, beneficial recessive alleles are much more likely to be driven to fixation, while alleles with equivalent selection coefficients are more likely be lost to drift on the autosomes (Maynard Smith and Haigh 1974; Charlesworth 1996). Veeramah et al. (2014) estimated that 46–51% of nonsynonymous mutations are driven to fixation by positive selection on the X chromosome, compared with 4–24% on the autosomes. Given this vast disparity in the rates of genic adaptive substitution between the X chromosome and the autosomes in humans, and the hitchhiking that results from this, it is perhaps unsurprising that the effect size for G on the X chromosome (β_{XG}) is greater than that of the autosomes (β_{AG}) (table 1), consistent with previous works (Hammer et al. 2010; Gottipati et al. 2011; Arbiza et al. 2014). In addition, for linked selection that includes nongenic sites, the effect for R on the X chromosome (β_{XR}) is larger than that of the autosomes (β_{AR}) (table 1), suggesting that this faster X effect may be acting on regulatory elements in addition to genic elements, though perhaps to a lesser degree as $\beta_{XG} > \beta_{XR}$. Attributing $\beta_{XR} > \beta_{AR}$ to other nonselective causes, however, is more difficult, and perhaps requires a mechanistic sex-biased explanation consistent with the X chromosome spending 2/3 of its time in females (but see Goldberg and Rosenberg 2015).

In considering the role that nongenic selective sites may have on population genetic diversity, we examined the effects of being linked to phylogenetically conserved *phastCons* elements inferred in primates. Using the same metrics as with G, we show that the effect size for being linked to the nearest conserved nonexonic elements (CNE) is as great as it is for conserved exonic elements (CEE) on both the X chromosome and the autosomes. This not only supports the hypothesis that linked selection to conserved (putatively regulatory) sites is as important as exonic sites but as CNEs outnumber CEEs nearly 3:1, it may suggest that the cumulative effects of being linked to regulatory sequence may be greater than the cumulative effects for being linked to protein-coding sequence.

The Human Neutralome

While our findings are consistent with the action of selection at linked sites, the extent that the linked selection to *phastCons* elements drives the patterns seen in figure 1 remains to be seen. Prior works have identified loci that are (effectively) unlinked from genes simply by setting a threshold on the genetic distance between a given locus and its closest genic nucleotide (Hammer et al. 2010; Gottipati et al. 2011; Arbiza et al. 2014). This thresholding strategy is unlikely to hold for *phastCons* elements, especially when tests of statistical significance are employed. As *phastCons* elements are widely distributed, any given locus is likely quite close (in this study, mean minimum genetic distance is 0.0067 cM) to a single such element. Thus, their small size and distributed nature necessitates a more comprehensive measure on the

Table 2. Demographic Parameters Estimated Considering Four Definitions of Neutral Sites.

Sample	Sample	Parameter	Point Estimate	0.025 CI	0.975 CI	*P* Value (vs. Neutralome)
All Sites	Neutralome	Na	10,777	9,410	11,366	
All Sites	NRE, bg97	Na	9,346	8,324	9,867	0.032
All Sites	NRE	Na	9,097	8,682	9,387	0.018
All Sites	4-Fold	Na	6,344	5,620	6,846	0.006
All Sites	Neutralome	n	1.67	1.60	1.88	
All Sites	NRE, bg97	n	1.81	1.74	2.00	0.062
All Sites	NRE	n	1.84	1.80	1.91	0.035
All Sites	4-Fold	n	2.58	2.41	2.82	0.006
All Sites	Neutralome	T (gen)	15,691	12,092	22,607	
All Sites	NRE, bg97	T (gen)	14,657	11,947	19,314	0.386
All Sites	NRE	T (gen)	12,312	10,806	14,354	0.078
All Sites	4-Fold	T (gen)	8,467	6,650	11,279	0.001
Excluding CpGs	Neutralome	Na	7,196	4,203	7,694	
Excluding CpGs	NRE, bg97	Na	5,800	1,854	6,710	0.075
Excluding CpGs	NRE	Na	6,331	5,826	6,632	0.096
Excluding CpGs	4-Fold	Na	2,343	959	2,745	0.009
Excluding CpGs	Neutralome	n	1.65	1.57	2.77	
Excluding CpGs	NRE, bg97	n	2.01	1.76	6.22	0.075
Excluding CpGs	NRE	n	1.86	1.80	1.99	0.097
Excluding CpGs	4-Fold	n	2.75	2.43	6.17	0.027
Excluding CpGs	Neutralome	T (gen)	11,103	8,031	23,729	
Excluding CpGs	NRE, bg97	T (gen)	14,112	9,816	26,402	0.766
Excluding CpGs	NRE	T (gen)	9,483	7,926	11,828	0.248
Excluding CpGs	4-Fold	T (gen)	4,137	2,683	8,341	0.009
Theta scaling	Neutralome	Na	10,777	9,410	11,366	
Theta scaling	NRE, bg97	Na	10,068	8,967	10,629	0.111
Theta scaling	NRE	Na	9,937	9,483	10,253	0.055
Theta scaling	Neutralome	n	1.67	1.60	1.88	
Theta scaling	NRE, bg97	n	1.81	1.74	2.00	0.062
Theta scaling	NRE	n	1.84	1.80	1.91	0.035
Theta scaling	Neutralome	T (gen)	15,691	12,092	22,607	
Theta scaling	NRE, bg97	T (gen)	15,790	12,870	20,806	0.545
Theta scaling	NRE	T (gen)	13,447	11,802	15,678	0.198

NOTE.—Sites from the neutralome, 4-fold degenerate sites, those identified by NRE, and those with a more stringent background selection coefficient (bg97) were assessed with ∂a∂I in the YRI. A three-parameter model instantaneous growth model was fit, inferring three parameters: An ancestral population size (Na), a proportional size change (*n*), and the time (*T*, in generations) of that change. ∂a∂I was assessed using all sites, all sites save CpG mutations, and with all sites and changing the number of sampled bases to reflect differences in π/D in the NRE data sets (Theta scaling). *P* values and confidence intervals were assessed by bootstrap, with the former relative to the neutralome.

amounts of linked selection that takes into account not just the closest nucleotides in the genetic map, but the cumulative effects of all linked nucleotides. To this end, we applied the approach of Halligan et al. (2013) to infer ν for each of our 10-kb loci (see Supplementary Material and supplementary figs. S9–S13, Supplementary Material online), with reductions in ν stemming solely from linkage to the local distribution of *phastCons* elements. The genomic distribution of ν shows that while ~70% genomic bases are weakly linked to sites of selection ($\nu = 80$–99%), only ~2% are apparently decoupled from sites of selection ($\nu > 99.9\%$ on the autosomes) (fig. 4).

In this study, we sought to identify loci that are generally unlinked to *phastCons* elements (i.e., with ν close to 1), to assess the hypothesis that these elements are instrumental in driving the trends seen in figure 1. Determining whether loci with high ν exhibit these same relationships is delicate. Notably, *phastCons* elements are unlikely to be the sole targets of selection in the genome, thus some loci with high ν may nevertheless be linked to a target of selection. For example, genic sites that do not intersect *phastCons* elements are not included in our estimate of ν. A simpler, and testable, approach is to assume that selective sites occur in a background that contains some appreciable density of *phastCons* elements. With this assumption, it follows that if ν is sufficiently large—that is, we are generally far from the linked effects of *phastCons* elements—then perhaps the trends in figure 1 will no longer be apparent. Thresholding solely on the value of ν is sufficient to remove both gradients apparent in figure 1, as well as any significant correlations with ν itself (see Supplementary Material, fig. 5, and supplementary figs. S6–S8 and tables S2 and S4–S6, Supplementary Material online). The high thresholds on ν lead to a diminutive human neutralome and demonstrate that linked selection, or perhaps some other (perhaps mechanistic) force that is tightly linked to *phastCons* elements, is the primary contributor to the correlation between π/D and *R* (and *G*).

The Neutralome versus Other Definitions of Neutrality

Demographic inferences are based on the assumption that the sequences considered are "neutral." We show that regions identified as neutral by NRE—regions that consider both linkage to genes and background selection in their definitions—lead to statistically significantly different values for the summary statistics than that of the neutralome. Some of these effects are less apparent when one considers sites with less predicted background selection (NRE, bg97), though reductions in power from a limited sample size likely also contribute to this finding. On the autosomes, the neutralome has higher effective population size than the loci identified by NRE (fig. 6), further, it has fewer low-frequency polymorphisms, as measured by Tajima's D (supplementary table S3, Supplementary Material online) and Fu and Li's D (fig. 7). This suggests that linked selection is not only reducing θ in the NRE loci but it is altering the distribution of the AFS (Tachida 2000; Williamson and Orive 2002; Nicolaisen and Desai 2012, 2013). While incorrect estimates of θ may only bias parameter estimates by a constant factor, changes to the AFS may fundamentally alter the demographic inference. To assess this, we fit a simple 2-epoch instantaneous growth model using ∂a∂I (Gutenkunst et al. 2009) considering sites in the neutralome, and two nested sets of loci identified NRE, as well as in 4-fold degenerate sites. All three demographic parameters varied considerably across these definitions of "neutral" sites, with the use of regions with greater selective impacts (4-fold degenerate sites and loci found by NRE) inferring a smaller ancestral effective population size, as well as larger population growth. Taken together this suggests the linked selection biases demographic inference in regions either classically considered to be neutral (4-fold sites), as well as sites that have been carefully chosen to minimize the effects of linked selection (NRE). While skews in the allele frequency spectrum, and the demographic biases that result, may seem to have a limited scope, demographic models serve as powerful null models for inferences on selective processes (Keightley and Eyre-Walker 2012; Singh et al. 2013; Veeramah et al. 2014; Hsieh et al. 2016; Uricchio et al. 2016). Thus, it follows that biased null demographic models may induce bias in the testing of alternative models of selection, though it should be noted that a "correct" null demographic model is not always necessary to make inference on selection (Messer and Petrov 2013; Tataru et al. 2017). However, the correct null model will always depend on the specifics of the hypotheses tested, and in some cases, say, distinguishing positive from negative selection, a null model of background selection may instead be preferred (Comeron 2017).

Genomic patterns of diversity are shaped by a combination of forces, including those that act within the cell (e.g., mutation and recombination), as well as those that act at the individual and population levels (e.g., demographic processes and natural selection). One of the key findings of this study is that due to the combined effects of selection and recombination, very few sites in the genome are unaffected by natural selection. This highlights the importance of considering only regions that are unlinked to *phastCons* elements when making inferences on the demographic history of human populations.

Materials and Methods

Samples and Locus Preparation

We used a total of 18 high coverage whole genomes, 9 West African samples (YRI), and 9 European samples (CEU), made publicly available by Complete Genomics (Drmanac et al. 2010). We computed nucleotide diversity (π) across the genome in nonoverlapping 10-kb windows of the hg19 genome. Divergence was computed from the 46-way Multiz vertebrate alignments made available from the UCSC genome browser website (Kent et al. 2002). Using the rhesus macaque to polarize sites, we computed the average divergence between each population and the ancestor of humans and orangutans. Both π and D were computed using a mixture of male and female samples. To perform such calling on the X chromosome, we utilized the ploidy information imbedded in the .tsv files for all samples. Similar to our analysis of de novo mutations (below), sites that fell in microsatellites, simple repeats, repetitive elements, segmental duplications, self-chain regions, as well as regions identified as copy number, structural variants or as uncallable by Complete Genomics, were removed from our analysis. After masking, 10-kb windows with ≤ 1 kb of callable sequence in either population or that were outside of our genetic map were also removed, giving a total sample size of 244,661 autosomal loci and 12,491 X chromosome loci.

We used the population averaged (YRI+CEU) genetic map from HapMap (Frazer et al. 2007), and scaled the X chromosome by ⅔ to compute all genetic distances. The UCSC known genes track (downloaded on 12/04/2015) (Hsu et al. 2006), which includes both protein coding and RNA genes, was analyzed using a custom perl script. For each locus, we computed in centimorgans (cM), the genetic length R and the minimum distance to genes G.

phastCons Element Analysis

phastCons elements inferred in primates were downloaded from the UCSC genome database (Kent et al. 2002), and their positions were encoded relative to the hg19 genome. The primates used in this inference are: chimpanzees (panTro2) gorilla (gorGor1) rhesus (rheMac2), baboon (papHam1), marmoset (calJac1), tarsier (tarSyr1), mouse lemur (micMur1), and bushbaby (otoGar1). We applied Ensembl's (version 72) perl API (Yates et al. 2016) to find exons from canonical protein-coding genes and divided *phastCons* conserved elements into two categories; 196,044 elements that intersected protein-coding exons (conserved exonic elements, or CEEs), and 529,386 elements that did not (conserved nonexonic elements, or CNEs). Using our 10-kb loci, including those that overlap genes, we computed the minimum genetic distance to CEE and CNE elements using the HapMap genetic map (Frazer et al. 2007).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Data Access

The code used to generate our estimates of ν across the genome is available at https://github.com/Ahhgust/Neutralome. A bed file of the human neutralome is available at: http://hammerlab.biosci.arizona.edu/Neutralome/neutralome.html

Acknowledgments

This work was supported by the US National Institutes of Health to M.F.H. (R01_HG005226) and the National Science Foundation Graduate Research grant (DGE-1143953) to A.E.W.

References

Arbeithuber B, Betancourt AJ, Ebner T, Tiemann-Boege I. 2015. Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proc Natl Acad Sci U S A.* 112(7):2109–2114.

Arbiza L, Gottipati S, Siepel A, Keinan A. 2014. Contrasting X-linked and autosomal diversity across 14 human populations. *Am J Hum Genet.* 94(6):827–844.

Arbiza L, Zhong E, Keinan A. 2012. NRE: a tool for exploring neutral loci in the human genome. *BMC Bioinformatics* 14:1.

Barton NH. 1998. The effect of hitch-hiking on neutral genealogies. *Genet Res.* 72(2):123–133.

Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* 356(6369):519–520.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. *Science* 304(5675):1321–1325.

Birky CW, Walsh JB. 1988. Effects of linkage on rates of molecular evolution. *Proc Natl Acad Sci U S A.* 85(17):6414–6418.

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* 140(2):783–796.

Cai JJ, Macpherson JM, Sella G, Petrov DA. 2009. Pervasive hitchhiking at coding and regulatory sites in humans. *PLoS Genet.* 5(1):e1000336.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* 134(4):1289–1303.

Charlesworth B. 1996. Background selection and patterns of genetic diversity in *Drosophila melanogaster*. *Genet Res.* 68(2):131–149.

Charlesworth B. 2009. Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet.* 10(3):195–205.

Chen CT, Wang JC, Cohen BA. 2007. The strength of selection on ultraconserved elements in the human genome. *Am J Hum Genet.* 80(4):692–704.

Chinwalla AT, Cook LL, Delehaunty KD, Fewell GA, Fulton LA, Fulton RS, Graves TA, Hillier LW, Mardis ER, McPherson JD. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420(6915):520–562.

Comeron JM. 2017. Background selection as null hypothesis in population genomics: insights and challenges from Drosophila studies. *Philos Trans R Soc B* 372(1736):20160471.

Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* 15(7):901–913.

Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, et al. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 327(5961):78–81.

Elyashiv E, Sattath S, Hu TT, Strutsovsky A, McVicker G, Andolfatto P, Coop G, Sella G. 2016. A genomic map of the effects of linked selection in Drosophila. *PLoS Genet.* 12(8):e1006130.

Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I, van Duijn CM, Swertz M, Wijmenga C, van Ommen G, et al. 2015. Genome-wide patterns and properties of de novo mutations in humans. *Nat Genet.* 47(7):822–826.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449(7164):851–861.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. *Genetics* 133(3):693–709.

Gillepsie JH. 2000. Genetic drift in an infinite population. The pseudohitchhiking model. *Genetics* 155:909–919.

Goldberg A, Rosenberg NA. 2015. Beyond 2/3 and 1/3: the complex signatures of sex-biased admixture on the X chromosome. *Genetics* 201(1):263–279.

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(8):741–743.

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(10):e1000695.

Halligan DL, Kousathanas A, Ness RW, Harr B, Eöry L, Keane TM, Adams DJ, Keightley PD. 2013. Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents. *PLoS Genet*. 9(12):e1003995.

Halligan DL, Oliver F, Guthrie J, Stemshorn KC, Harr B, Keightley PD. 2011. Positive and negative selection in murine ultraconserved noncoding elements. *Mol Biol Evol.* 28(9):2651–2660.

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(10):830–831.

Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, Sella G, Przeworski M. 2011. Classic selective sweeps were rare in recent human evolution. *Science* 331(6019):920–924.

Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genet Res.* 8(3):269–294.

Hsieh P, Veeramah KR, Lachance J, Tishkoff SA, Wall JD, Hammer MF, Gutenkunst RN. 2016. Whole-genome sequence analyses of Western Central African Pygmy hunter-gatherers reveal a complex demographic history and identify candidate genes under positive natural selection. *Genome Res.* 26(3): 279–290.

Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. 2006. The UCSC known genes. *Bioinformatics* 22(9):1036–1046.

Hudson RR, Kaplan NL. 1995. Deleterious background selection with recombination. *Genetics* 141:1605–1607.

Hudson RR. 2007. The variance of coalescent time estimates from DNA sequences. *J Mol Evol.* 64(6):702.

Katzman S, Capra JA, Haussler D, Pollard KS. 2011. Ongoing GC-biased evolution is widespread in the human genome and enriched near recombination hot spots. *Genome Biol Evol.* 3:614–626.

Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, Wilson RK, Salama SR, Haussler D. 2007. Human genome ultraconserved elements are ultraselected. *Science* 317(5840):915.

Keightley PD, Eyre-Walker A. 2012. Estimating the rate of adaptive molecular evolution when the evolutionary divergence between species is small. *J Mol Evol.* 74(1–2):61–68.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res.* 12(6):996–1006.

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, igurdsson 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(7412):471–475.

Lohmueller KE, Albrechtsen A, Li Y, Kim SY, Korneliussen T, Vinckenbosch N, Tian G, Huerta-Sanchez E, Feder AF, Grarup N, et al. 2011. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. *PLoS Genet.* 7(10):e1002326.

Marais G. 2003. Biased gene conversion: implications for genome and sex evolution. *Trends Genet.* 19(6):330–338.

Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genet Res.* 23(01):23–35.

McGaugh SE, Heil CS, Manzano-Winkler B, Loewe L, Goldstein S, Himmel TL, Noor MA. 2012. Recombination modulates how selection affects linked sites in Drosophila. *PLoS Biol*. 10(11):e1001422.

McVicker G, Gordon D, Davis C, Green P. 2009. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet.* 5(5):e1000471.

Meader S, Ponting CP, Lunter G. 2010. Massive turnover of functional sequence in human and other mammalian genomes. *Genome Res.* 20(10):1335–1343.

Messer PW, Petrov DA. 2013. Frequent adaptation and the McDonald–Kreitman test. *Proc Natl Acad Sci U S A.* 110(21):8615–8620.

Nachman MW, Crowell SL. 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics* 156(1):297–304.

Nachman MW. 2001. Single nucleotide polymorphisms and recombination rate in humans. *Trends Genet*. 17(9):481–485.

Nicolaisen LE, Desai MM. 2012. Distortions in genealogies due to purifying selection. *Mol Biol Evol.* 29(11):3589–3600.

Nicolaisen LE, Desai MM. 2013. Distortions in genealogies due to purifying selection and recombination. *Genetics* 195(1):221–230.

Nordborg M, Charlesworth B, Charlesworth D. 1996. The effect of recombination on background selection. *Genet Res*. 67(2):159–174.

O'Fallon BD, Seger J, Adler FR. 2010. A continuous-state coalescent and the impact of weak selection on the structure of gene genealogies. *Mol Biol Evol.* 27(5):1162–1172.

Ponting CP, Hardison RC. 2011. What fraction of the human genome is functional? *Genome Res.* 21(11):1769–1776.

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(7459):471–475.

Pratto F, Brick K, Khil P, Smagulova F, Petukhova GV, Camerini-Otero RD. 2014. Recombination initiation maps of individual human genomes. *Science* 346(6211):1256442.

Rands CM, Meader S, Ponting CP, Lunter G. 2014. 8.2% of the human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. *PLoS Genet.* 10(7):e1004525.

Schrider DR, Kern AD. 2017. Soft sweeps are the dominant mode of adaptation in the human genome. *Mol Biol Evol.* 34(8):1863–1877.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15(8):1034–1050.

Singh ND, Jensen JD, Clark AG, Aquadro CF. 2013. Inferences of demography and selection in an African population of *Drosophila melanogaster*. *Genetics* 193(1):215–228.

Spencer CC, Deloukas P, Hunt S, Mullikin J, Myers S, Silverman B, Donnelly P, Bentley D, McVean G. 2006. The influence of recombination on human genetic diversity. *PLoS Genet*. 2(9):e148.

Strathern JN, Shafer BK, McGill CB. 1995. DNA synthesis errors associated with double-strand-break repair. *Genetics* 140(3):965–972.

Tachida H. 2000. DNA evolution under weak selection. *Gene* 261(1):3–9.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3):585–595.

Tataru P, Mollion M, Glémin S, Bataillon T. 2017. Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. *Genetics* 207(3):1103–1119.

Uricchio LH, Zaitlen NA, Ye CJ, Witte JS, Hernandez RD. 2016. Selection and explosive growth alter genetic architecture and hamper the detection of causal rare variants. *Genome Res.* 26(7):863–811.

Veeramah KR, Gutenkunst RN, Woerner AE, Watkins JC, Hammer MF. 2014. Evidence for increased levels of positive and negative selection on the X chromosome versus autosomes in humans. *Mol Biol Evol.* 31(9):2267–2282.

Williamson S, Orive ME. 2002. The genealogy of a sequence subject to purifying selection at multiple sites. *Mol Biol Evol.* 19(8):1376–1384.

Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, et al. 2016. Ensembl 2016. *Nucleic Acids Res.* 44(D1):D710–D716.