Impact of mutation rate and selection at linked sites on fine-scale DNA variation across the homininae genome

David Castellano[1,3,4]
Adam Eyre-Walker[2]
Kasper Munch[1]

1. Bioinformatics Research Centre
Aarhus University
C.F. Møllers Allé 8 DK-8000 Aarhus C
Denmark

2. School of Life Sciences
University of Sussex
Brighton BN1 9QG
United Kingdom

3. Current affiliation:
Centre for Genomic Regulation (CRG),
The Barcelona Institute of Science and Technology,
Dr. Aiguader 88,
Barcelona, 08003, Spain

4. Corresponding author: E-mail: david.castellano@crg.eu

Abstract

DNA diversity varies across the genome of many species. Variation in diversity across a genome might arise for one of three reasons; regional variation in the mutation rate, selection and biased gene conversion. We show that both non-coding and non-synonymous diversity are correlated to a measure of the mutation rate, the recombination rate and the density of conserved sequences in 50KB windows across the genomes of humans and non-human homininae. We show these patterns persist even when we restrict our analysis to GC-conservative mutations, demonstrating that the patterns are not driven by biased gene conversion. The positive correlation between diversity and our measure of the mutation rate seems to be largely a direct consequence of regions with higher mutation rates having more diversity. However, the positive correlation with recombination rate and the negative

correlation with the density of conserved sequences suggests that selection at linked sites affect levels of diversity. This is supported by the observation that the ratio of the number of non-synonymous to non-coding polymorphisms is negatively correlated to a measure of the effective population size across the genome. Furthermore, we find evidence that these genomic variables are better predictors of non-coding diversity in large homininae populations than in small populations, after accounting for statistical power. This is consistent with genetic drift decreasing the impact of selection at linked sites in small populations. In conclusion, our comparative analyses describe for the first time how recombination rate, gene density, mutation rate and genetic drift interact to produce the patterns of DNA diversity that we observe along and between homininae genomes.

Introduction

The level of genetic variation is known to vary across the genome of many species and this depends on genomic characteristics such as recombination, gene density and mutation rate. This was first demonstrated by Begun and Aquadro (Begun & Aquadro 1992) who showed that putatively neutral genetic diversity was correlated to the rate of recombination across the genome of *Drosophila melanogaster*. This has subsequently been observed in species as diverse as humans and tomatoes (reviewed by Cutter & Payseur [2013]).

Variation in diversity across the genome of a given species might arise for one of three reasons; variation in the mutation rate, selection, and biased gene conversion. The mutation rate can affect the level of diversity both directly and indirectly. Directly, the level of genetic diversity is expected to depend upon the rate of mutational input; the higher the mutation rate, the more diversity there is expected to be. It can also have an indirect effect by increasing the frequency of selection at linked sites, which is described below. Natural selection can also affect the level of genetic diversity both directly and indirectly. Direct selection tends to either decrease or increase diversity at the sites at which it is acting, depending on whether the selection is either negative or positive, particularly if there is balancing selection. However, in general, selection tends to act indirectly reducing diversity at linked sites through the processes of genetic hitch-hiking (HH) (Smith & Haigh 1974) and background selection (BGS) (Charlesworth et al. 1993). Genetic hitch-hiking also has the effect of moving a locus away from a state of quasi-equilibrium; after a selective sweep, deleterious genetic variation approaches its equilibrium value more rapidly than neutral variation, leading to a disproportionate amount of the diversity being deleterious (Gordo & Dionisio 2005; Pennings et al. 2014; Brandvain & Wright 2016; Do et al. 2015; Castellano et al. 2018). Finally, GC biased gene conversion (gBGC) has the potential to affect diversity and the efficiency of

selection. In humans and all organisms that have been studied, biased gene conversion appears to be GC-biased. This is counter to the pattern of mutation, which is AT-biased. In such a system in which gBGC and mutation are in opposite direction, an increase in gBGC (from no gBGC) is expected to slightly increase the levels of diversity before reducing them, when gBGC becomes strong (McVean & Charlesworth 1999). gBGC can actually hamper natural selection when the G or C alleles are less favorable for the fitness than the A or T alleles (Glémin 2010; Capra et al. 2013; Necşulea et al. 2011; Lachance & Tishkoff 2014), particularly at recombination hotspots (Glémin et al. 2015).

If selection at linked sites is pervasive and acts genome-wide, this should be visible as correlations between DNA diversity and factors affecting the intensity of selection at linked sites, such as recombination rate, gene density and mutation rate. In this work we return to the question of whether selection at linked sites has an effect on levels of DNA sequence diversity and the efficiency of purifying selection along the autosomes of humans and our closest living relatives, the homininae subfamily: humans, bonobos, chimpanzees and gorillas. Since these species diverged very recently but differ substantially in the effective population size (N_e) (Prado-Martinez et al. 2013) we can also investigate how very similar genomes react to regional variation in mutation, selection and recombination when exposed to different amounts of genetic drift. The difference in the N_e between the largest and smallest population in our dataset, gorillas and bonobos, is more than two-fold (Prado-Martinez et al. 2013). This allows us to test the population genetic prediction that natural selection is less efficient in small populations than large ones (Ohta 1992). Thus, we predict that recombination rate, gene density and mutation rate should be better predictors of the level of neutral diversity in large than in small populations because selection at linked sites will contribute more to the patterns of diversity in larger populations (Corbett-Detig et al. 2015).

In other words, in very small populations most of the regional variation in diversity will come from variation in mutation rate because selection (direct and indirect) will be very inefficient, while in large populations both selection at linked sites and mutation rate variation will contribute to the patterns of neutral variation.

Furthermore, the role that mutation, selection and gBGC play in determining the levels of genetic diversity and the efficiency of natural selection across the non-human homininae genomes remains unresolved. In humans, it was observed many years ago that levels of diversity at putatively neutral sites are correlated to the rate of recombination (Lercher & Hurst 2002; Hellmann et al. 2005). Since the rate of substitution is also correlated to the rate of recombination it seemed likely that at least part of the correlation between diversity and the rate of recombination was due to a mutagenic effect of recombination. There is now good evidence that recombination is mutagenic in humans (Pratto et al. 2014; Francioli et al. 2015; Arbeithuber et al. 2015; Halldorsson et al. 2019) and recent analyses of the correlation between diversity and the rate of mutation, as inferred from rates of *de novo* mutations in human trios, suggests that much, but not all, of the variation in diversity across the human genome can be explained by variation in the rate of mutation at the 100KB and 1MB scale (Smith et al. 2018). However, several lines of evidence suggest that selection at linked sites may also affect neutral and selected diversity across the human genome. First, it has been observed that levels of diversity are negatively correlated to gene density (Payseur & Nachman 2002). Second, levels of non-coding diversity are lower near functional DNA elements in humans and non-human primates (McVicker et al. 2009; Enard et al. 2014; Nam et al. 2017). Third, the rate of non-synonymous to synonymous substitution is positively correlated to gene (and exon) density (Bullaughey et al. 2008). Fourth, Hussin et al. (2015) showed that exons in regions of low recombination are significantly enriched for deleterious

and disease-associated variants consistent with variation in the intensity of selection at linked sites generating variation in the efficiency of purifying selection along the genome.

Briefly, we find that our three genomic variables: recombination rate, mutation rate and the density of conserved sites are correlated to each other. We show that the levels of both putatively neutral non-coding and putatively selected non-synonymous variation are correlated to those genomic variables in most homininae, but that the relative importance of each genomic variable is different for non-coding and non-synonymous polymorphisms. Interestingly, we confirm that linked selection is probably contributing more to the patterns of non-coding diversity in large homininae populations than in small ones. We also find evidence that indicates variation in the efficiency of negative selection likely generated by interference among deleterious mutations in the genome of all the homininae. Finally, we find little impact of gBGC in our analyses and conclusions when interrogating only GC-conservative mutations.

Materials and Methods

Population genomic data

SNP calls from the autosomes are retrieved from Prado-Martinez et al. (2013) for five great ape populations: *Homo sapiens, Pan paniscus, Pan troglodytes ellioti, Pan troglodytes verus* and *Gorilla gorilla gorilla.* Hereafter we refer to these species as humans, bonobos, Nigeria-Cameroon chimpanzees, western chimpanzees and gorillas, respectively. In Prado-Martinez et al. (2013) all reads are mapped to the human reference genome (hg18), we lift over the original VCF to hg19/GRCh37.75 coordinates to take advantage of more recent functional annotations (see below). To avoid errors introduced by miss-mapping due to paralogous variants and repetitive sequences, we also restrict all analyses to a set of sites with a unique mapping to the human genome as in Cagan et al. (2016). Additionally, we also require positions to have at least 5-fold coverage in all individuals per species. Only the resulting set of sites are used in further analyses (Supplementary Table 1). The final list of analyzed positions is available upon request.

Genome annotation and identification of putatively neutral non-coding sites

Genomes are annotated using the SnpEff and SnpSift software (Cingolani et al. 2012) (version 4.3m, last accessed June 2017) and the human database GRCh37.75. We extract 0-fold degenerate sites from the codon information provided by SnpEff (4-fold, 2-fold and 3-fold degenerate sites are discarded). We assume that the degeneracy and gene annotations are identical across species. In order to obtain putatively neutral non-coding sites, we applied stringent filters: (1) We keep sites annotated only as intronic or intergenic (splicing sites, UTRs, coding or transcribed non-coding genes are discarded). (2) We remove GERP elements (Davydov et al. 2010) and positions with a PhastCons score > 5% in primates and/or in 100 vertebrate species (Siepel et al. 2005). In this way, we remove conserved sites at different

phylogenetic depths. Note that GERP elements were calculated in a multiple species alignment of the human genome to 33 other mammalian species (the most distant mammalian species is Platypus), while the PhastCons scores used here are based on a multiple species alignment of the human genome to 9 other primates (the most distant is bushbaby) and a multiple species alignment of the human genome to 99 other vertebrate species (the most distant is zebrafish). (3) Some sites might have become functional very recently. Thus, we remove *DNase* I hypersensitivity sites across multiple human tissues (Song et al. 2011) (downloaded from: http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/openchrom/jan2011/combined_peaks/). (4) Predicted transcription factor binding sites detected in humans are also excluded (Cingolani et al. 2012). (5) Hypermutable *CpG* sites in humans and the rest of species are excluded to remove variation in mutation rate due to variation in GC-content. This last filter is also applied to coding sites and mutation rate estimates (see below).

Genomic windows and statistics

We split the autosomes in non-overlapping windows of 50KB, and for each window we estimate: (1) diversity at putatively neutral non-coding (NC) sites, (2) 0-fold degenerate (N) sites, (3) GC-conservative substitution rate at NC sites (our main proxy of the mutation rate), (4) the density of conserved sites (DCS), (5) the rate of recombination (RR, crossing overs) and (6) the rate of *de novo* mutations (DNMs) (an alternative proxy of the mutation rate).

Recombination maps and the density of conserved sites (DCS)

Population recombination rate estimates for non-human great apes are retrieved from (Stevison et al. 2016; Auton et al. 2012) and human population recombination rates from

HapMap (Myers et al. 2005) and deCODE (Kong et al. 2002). Blocks for each non-human genome that are syntenic with human are identified as in Stevison et al. (2016). To estimate the density of conserved sites we use as before GERP elements (Davydov et al. 2010) and PhastCons scores (Siepel et al. 2005), but this time we label and count all GERP elements and/or positions with a PhastCons score > 50% in primates and/or positions with a PhastCons score > 50% in 100 vertebrate species in a given 50KB window. We discard the number of unsequenced nucleotides (N) in the human reference genome to estimate the DCS.

Polymorphism and mutation rate estimates

For a fair comparison between species, we downsample our population genomic data to 8 haploid chromosomes per position. Note that there are only 5 individuals of western chimpanzees. Positions called in less than 8 chromosomes are excluded. For each window we count the total number of analyzable polymorphic sites ($L_{P,N}$ and $L_{P,NC}$) and the number of segregating sites (S_N and S_{NC}) to get the Watterson estimator (θ, Watterson [1975]) for NC and N sites, respectively. We do this for all point mutations and for GC-conservative point mutations alone.

For divergence estimates, we count the total number of analyzable divergent non-coding sites ($L_{D,NC}$) from a multiple species alignment between one randomly sampled Nigeria-Cameroon chimpanzee, western chimpanzee, bonobo, gorilla and human chromosome. This multiple species alignment is generated from Prado-Martinez et al. (2013) original VCF file. Then, to estimate our proxy of the mutation rate in each window (d_{NC}), and given that there are few GC-conservative substitutions per window, we sum all GC-conservative substitutions (D_{NC}) occurring in the homininae tree and divide it by $L_{D,NC}$. Thus, our proxy of the mutation rate is the same for all species and it is unaffected by gBGC. For

some validation analyses we also use the rate of DNMs per window in humans (Jónsson et al. 2017; Wong et al. 2016; Francioli et al. 2015) as an alternative proxy of the regional mutation rate. We only consider non-CpG DNMs but this time we consider both GC-conservative and non-GC-conservative DNMs.

Hypergeometric sampling and grouping

We are interested in the effect of the mutation rate (d_{NC}), the rate of recombination (RR) and the density of conserved sites (DCS) on the efficiency of negative selection across the homininae genomes. We consider first the log (θ_N/θ_{NC}) and its relationship to the mutation rate, the rate of recombination and the DCS, and then its relationship to a measure of the local effective population size, $\log(\theta_{NC}/d_{NC})$. $\log(\theta_N/\theta_{NC})$ and $\log(\theta_{NC}/d_{NC})$ are undefined if θ_N, θ_{NC} or d_{NC} are zero, we, therefore, combined data across windows in the following manner. We split the number of non-coding polymorphisms (S_{NC}) in a window into three parts using a hypergeometric distribution. We used $\theta_{NC,1}$ to rank and bin windows into 50 groups, $\theta_{NC,2}$ is our unbiased measure of the non-coding diversity while $\theta_{NC,3}$ is used to estimate the ratio θ_N/θ_{NC}. We used two methods to combine data across windows. In both methods we split S_{NC} into three statistically independent estimates using a hypergeometric distribution. In the first method, we include all windows that have non-coding sites, irrespective of whether they have coding sites. In the second method, we only include windows with coding sites. The second method yields about 43% of the data-points of the first method due to the requirement that windows have both coding and non-coding sites. The rationale for using two methods is that for the non-coding analyses conserved non-coding sites might be a source of selection at linked sites. We present results from method 1 for non-coding results and from method 2 for non-synonymous results. There are some regions with very high θ_{NC} and d_{NC} values. These have a disproportionate influence over the stats. We exclude the top 1% diversity and

divergence regions which turn out to overlap with well-known regions under strong balancing selection, like the MHC locus, and structurally complex regions of the genome which are probably alignment and read mapping errors (data not shown).

Expected relationship between θ_N/θ_{NC} and θ_{NC} under free recombination and limited recombination

θ_N and θ_{NC} are expected to be correlated through variation in the mutation rate and/or the N_e. If we assume that the distribution of fitness effects (DFE) of new deleterious mutations follows a gamma distribution, then under free recombination the slope (b) of the relationship between θ_N/θ_{NC} and θ_{NC} in a log-log scale informs us about the source of this variation (Welch et al. 2008). If there is no variation in the N_e and all variation in θ_N and θ_{NC} is due to variation in mutation rate, then we expect b = -1. In contrast, if all the variation in θ_N and θ_{NC} comes from variation in the N_e, then we expect b = -β (Welch et al. 2008), where β is the shape parameter of the distribution. Finally, b = 0 if θ_N and θ_{NC} are independent.

Forward simulations with background selection (that is limited recombination plus deleterious and neutral mutations) and variation in the N_e among loci have shown that the shape of the deleterious DFE can be successfully estimated with the slope between $\log(\theta_N/\theta_{NC})$ and $\log(\theta_{NC})$ after correcting for variation in the mutation rate (James et al. 2017). Castellano et al. (2018) showed that the slope is overestimated when HHs are incorporated due to the faster recovery of the levels of deleterious variation compared to the levels of neutral variation.

Statistical analyses

All statistical analyses are performed within the R framework (version 3.4.4). Here we want to explain how our dependent variable, the number of SNPs in a given window, which is discrete and over-dispersed data (Supplementary Figure 1), is related to our three genomic variables. To do that we implement a negative binomial regression by means of the R function glm.nb(). We model the log of the expected number of non-coding or non-synonymous SNPs as a function of the predictor variables; recombination rate, d_{NC}, DCS and the number of non-coding or 0-fold degenerate sites in a given 50KB window, respectively. We can interpret the negative binomial regression coefficient as follows: for a one unit change in the predictor variable, the difference in the logs of expected counts of the response variable is expected to change by the respective regression coefficient, given the other predictor variables in the model are held constant. To assess the relative importance of each genomic variable, we also report standardized regression coefficients to make variances of dependent and independent variables 1. These standardized coefficients refer to how many standard deviations the log of the expected number of SNPs will change, per standard deviation increase in the predictor variable. To explain the variation in our statistic of the efficiency of negative selection, $\log(\theta_N/\theta_{NC})$, we use a standard multiple linear regression using the R function lm(). As before the standardized regression coefficients are used to assess the relative importance of each genomic variable. Finally, to estimate the variance inflation factor we use the R function vif() and to perform the bivariate correlations we use the R function cor.test() and the non-parametric Spearman method.

Results

Genomic determinants of non-coding and non-synonymous diversity

We are interested in how genetic variation is distributed along the homininae genome and in particular, the role that selection at linked sites and mutation rate variation might play in this distribution. To see whether these patterns are consistent across all homininae we used the data of Prado-Martinez et al. (2013) from Nigeria-Cameroon chimpanzees, western chimpanzees, bonobos and gorillas. To allow an unbiased comparison with the other homininae we use the cosmopolitan human sample from the Great Apes Project. We used the same DCS and mutation rate for all species, while the rate of recombination estimates are population-specific and come from publicly available recombination rate maps (Stevison et al. 2016; Auton et al. 2012). We estimate the level of diversity at non-coding sites and 0-fold degenerate sites in 50KB windows across the homininae autosomes. We exclude non-coding sites that are inferred to be subject to natural selection based on the conservation of sites across species and other potentially functional annotations such coding and non-coding genes, UTRs, *DNase I* hypersensitivity sites and transcription factor binding sites. Supplementary Table 1 shows the summary statistics of our analyzed dataset.

We expect the level of genetic diversity at both selected and neutral sites to depend on the mutation rate (d_{NC}), the rate of recombination (RR) and the density of conserved sites (DCS), because each of these factors are expected to affect the diversity either directly, or indirectly. To estimate the mutation rate there are two options: using *de novo* mutations (DNMs) that have been discovered by the sequencing of trios or using the divergence between species. Neither of these methods is perfect. We currently have too few DNMs to estimate the mutation rate reliably at the 50KB scale and attempts to predict the mutation rate of DNMs based on genomic features have so far proved to be unreliable (Smith et al. 2018). The

divergence between species is also not a completely satisfactory measure of the mutation rate either, for several reasons. We have used GC-conservative substitutions (i.e. A<>T and G<>C), since these are not affected by gBGC, a process known to affect substitution rates (Duret & Arndt 2008; Smith et al. 2018), and the rate of different types of mutation appear to be strongly correlated at the 100KB and 1MB scales, suggesting that GC-conservative mutations should therefore be a reasonable measure of the overall mutation rate (Smith et al. 2018). However, the mutation rate appears to evolve at large scales (Terekhanova et al. 2017; Smith et al. 2018), and some of the variation in the substitution rate is due to variation in the depth of the genealogy in the ancestors of the homininae (Phung et al. 2016).

We find each genomic variable is individually correlated to non-coding diversity if we run a bivariate analysis (Supplementary Figure 2-6) and in a negative binomial regression all three genomic variables come out as significant (Table 2A). Although the three genomic features are correlated to each other (Table 1), their variance inflation factors are small (VIF~1) suggesting that multicollinearity is negligible.

	d_{NC}	DCS
RR	0.07 ***	-0.04 ***
d_{NC}		-0.24 ***

Table 1. Spearman correlations between genomic variables in humans.

We confirm the positive correlation between the mutation rate and the rate of recombination in humans but also, for the first time, in the other non-human homininae (Table 1; Supplementary Table 2). We also find a negative correlation between the DCS and the rate of recombination in humans (as in McVean et al. [2004]; Kong et al. [2010]) and the rest of homininae (Table 1; Supplementary Table 2). We find this negative correlation is driven by

conserved coding sites but not by conserved non-coding sites in agreement with the lower rate of recombination seen in exons and nearby non-coding regions (Supplementary Analyses). Interestingly, there is a strong negative correlation between the DCS and our measure of the mutation rate. This finding is further investigated in humans using three large publicly available *de novo* mutation (DNM) datasets from trios (Francioli et al. 2015; Wong et al. 2016; Jónsson et al. 2017) (Supplementary Analyses). We find that the density of putatively neutral DNMs is either significantly positively or uncorrelated to the DCS, depending on which dataset of DNMs is considered. Hence, we do not know whether the negative correlation between d_{NC} and DCS is a consequence of co-variation in the mutation rate and the DCS, or that d_{NC} is reduced in regions with high DCS because, despite our stringent filtering, we might not be masking all sites under purifying selection.

Standardised regression coefficients suggest that the rate of recombination is slightly more important than mutation rate and the DCS in determining levels of putatively neutral diversity in humans and non-human homininae (Table 2A), with the exception of a weak correlation between non-coding diversity and the rate of recombination in western chimpanzees. Despite the limited number of non-synonymous SNPs per window (Supplementary Figure 1) the results at 0-fold degenerate sites are remarkably stable across our set of analyzed species. Interestingly, the DCS is the strongest correlate as judged by standardised regression coefficients (Table 2B). The DCS has more than twice the impact on non-synonymous diversity as the rate of recombination and our measure of the mutation rate has slightly more than half of the impact of RR (note a lack of correlation with d_{NC} in bonobos for non-synonymous diversity). In other words, the density of conserved elements dominates non-synonymous diversity, followed by the rate of recombination and mutation rate, with the exception again of western chimpanzees where the rate of recombination is the weakest predictor variable.

A)

Species	RR	d_{NC}	DCS	R^2	$R^{2,L-}$	$R^{2,L+}$
Human	0.0030 ***	0.0023 ***	-0.0026 ***	67%	12%	63%
NC chimpanzee	0.0032 ***	0.0026 ***	-0.0019 ***	71%	14%	67%
W chimpanzee	0.0008 ***	0.0037 ***	-0.0019 ***	54%	7%	52%
Bonobo	0.0044 ***	0.0025 ***	-0.0019 ***	68%	10%	65%
Gorilla	0.0024 ***	0.0016 ***	-0.0017 ***	74%	14%	70%

B)

Species	RR	d_{NC}	DCS	R^2	$R^{2,L-}$	$R^{2,L+}$
Human	0.0988 ***	0.0561 ***	-0.2817 ***	27%	0.64%	24%
NC chimpanzee	0.1047 ***	0.0605 ***	-0.2460 ***	27%	0.62%	23%
W chimpanzee	0.0748 **	0.1029 ***	-0.3728 ***	23%	0.40%	21%
Bonobo	0.1128 ***	0.0085	-0.3703 ***	24%	0.47%	22%
Gorilla	0.1034 ***	0.0660 ***	-0.3714 ***	25%	0.16%	22%

Table 2. The standardised regression coefficients from the full model regressing the number of SNPs against the recombination rate (RR), d_{NC} and DCS along with the R^2 for the full model, and a R^2 for the model without L and with only L, where L is the number of analyzed sites for A) non-coding SNPs and B) non-synonymous SNPs versus recombination rate, d_{NC} and the DCS. * *P*-value<0.05, ** *P*-value<0.01, *** *P*-value<0.001

Both non-synonymous and non-coding diversity increases significantly with our measure of the recombination rate and decreases with the density of conserved elements. The level of both types of diversity also increases with mutation rate, but mutation rate is a stronger determinant of non-coding diversity than non-synonymous diversity.

Effective population size

The species N_e can influence the strength of the relationship between recombination rate, mutation rate and the density of conserved sites and neutral diversity. For example, we might expect that the higher the species N_e, the greater the contribution of selection at linked sites to the patterns of diversity (Corbett-Detig et al. 2015). This is because the N_e determines both the rate of new (effectively) selected mutations that enter the population each generation and the population recombination rate. The interaction between these two variables and the distribution of fitness effects (DFE) will determine the strength of selection at linked sites and the levels of neutral diversity at fine-scale. Assuming that the DFE is similar between these species and constant along the genome, then we should expect a greater impact of selection at linked sites in large than small populations. In other words, our genomic variables should be better predictors of the level of putatively neutral non-coding diversity in gorillas (large N_e) than bonobos (low N_e) because genetic drift will be stronger in the latter. Figure 1 shows that this is the case after accounting for statistical power (by downsampling species diversity to that found in the less diverse species, bonobos). Despite the limited number of data points, there is a marginally significant positive correlation between the amount of explained variance in non-coding diversity and our proxy of the species N_e (Spearman coefficient r = 0.87, *P*-value = 0.053). The values of the correlation leaving one population out goes from r = 0.95, *P*-value = 0.051 (when bonobos or western chimpanzees are excluded) to r = 0.74, *P*-value = 0.26 (when humans are excluded) showing that the correlation is not driven by an outlier.

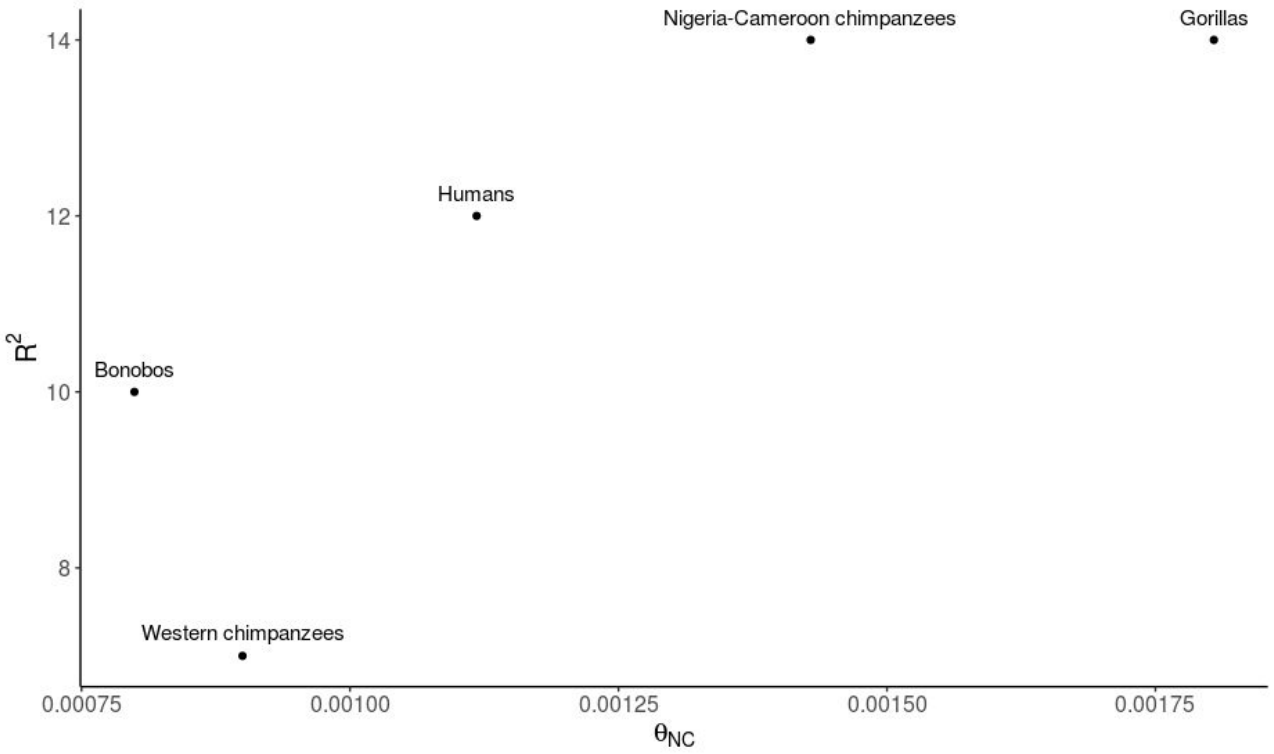

Figure 1. Relationship between non-coding diversity (our proxy of the N_e) and the amount of variance in non-coding diversity explained by our three genomic variables after down-sampling non-coding SNPs to yield the same average number of SNPs per 50KB as we find in bonobos (the species with the lowest genetic diversity).

Biased gene conversion

Furthermore, the influence of the rate of recombination on diversity might be mediated by gBGC. Although biased gene conversion is not expected to affect diversity greatly, we repeated our analyses restricting the analysis to GC-conservative SNPs. As expected our results are largely unaffected by this restriction; non-coding diversity remains correlated to all factors (Table 3A). For non-synonymous diversity we again find qualitatively similar results to those using all mutations, although the correlation between non-synonymous diversity and our estimate of the mutation is generally non-significant (Table 3B).

A)

Species	RR	d_{NC}	DCS	R^2	$R^{2,L-}$	$R^{2,L+}$
Human	0.0119 ***	0.0227 ***	-0.0104 ***	53%	11%	49%
NC chimpanzee	0.0142 ***	0.0212 ***	-0.0076 ***	60%	14%	55%
W chimpanzee	0.0014	0.0301 ***	-0.0073 ***	41%	7%	39%
Bonobo	0.0199 ***	0.0204 ***	-0.0083 ***	48%	8%	46%
Gorilla	0.0120 ***	0.0152 ***	-0.0089 ***	65%	14%	61%

B)

Species	RR	d_{NC}	DCS	R^2	$R^{2,L-}$	$R^{2,L+}$
Human	0.2374 *	0.1669	-1.0218 ***	16%	0.41%	14%
NC chimpanzee	0.3491 **	0.2527 *	-1.0033 ***	18%	0.42%	16%
W chimpanzee	0.2040	0.3047	-1.4525 ***	14%	0.26%	13%
Bonobo	0.3335 *	0.1897	-1.3009 ***	14%	0.24%	13%
Gorilla	0.2977 *	0.1518	-1.6346 ***	14%	0.07%	12%

Table 3. The standardised regression coefficients from the full model regressing the number of SNPs against the recombination rate (RR), d_{NC} and DCS along with the R^2 for the full model, and a R^2 for the model without L and with only L, where L is the number of analyzed sites for A) non-coding SNPs and B) non-synonymous GC-conservative SNPs versus recombination rate, d_{NC} and the DCS.

To investigate in more detail whether gBGC is having an effect we considered the correlation between diversity and the recombination rate, but down-sampling non-GC-conservative mutations to yield the same average number of mutations per 50KB as we find in the GC-conservative dataset. GC-conservative mutations represent 15-18% of all SNPs in the homininae subfamily (Supplementary Table 1). Although the effect is very small (Figure 2), the correlation between θ_{NC} and the rate of recombination is significantly greater for non-GC-conservative mutations than GC-conservative mutations in all homininae (as judged by

a sign test, *P*-value < 0.01). This suggests that gBGC is having a significant but very small effect on the correlation between recombination rate and non-coding diversity. The strength of the correlation between θ_N and the rate of recombination is again similar between mutation types and the difference is only significant in humans (*P*-value = 0.043).

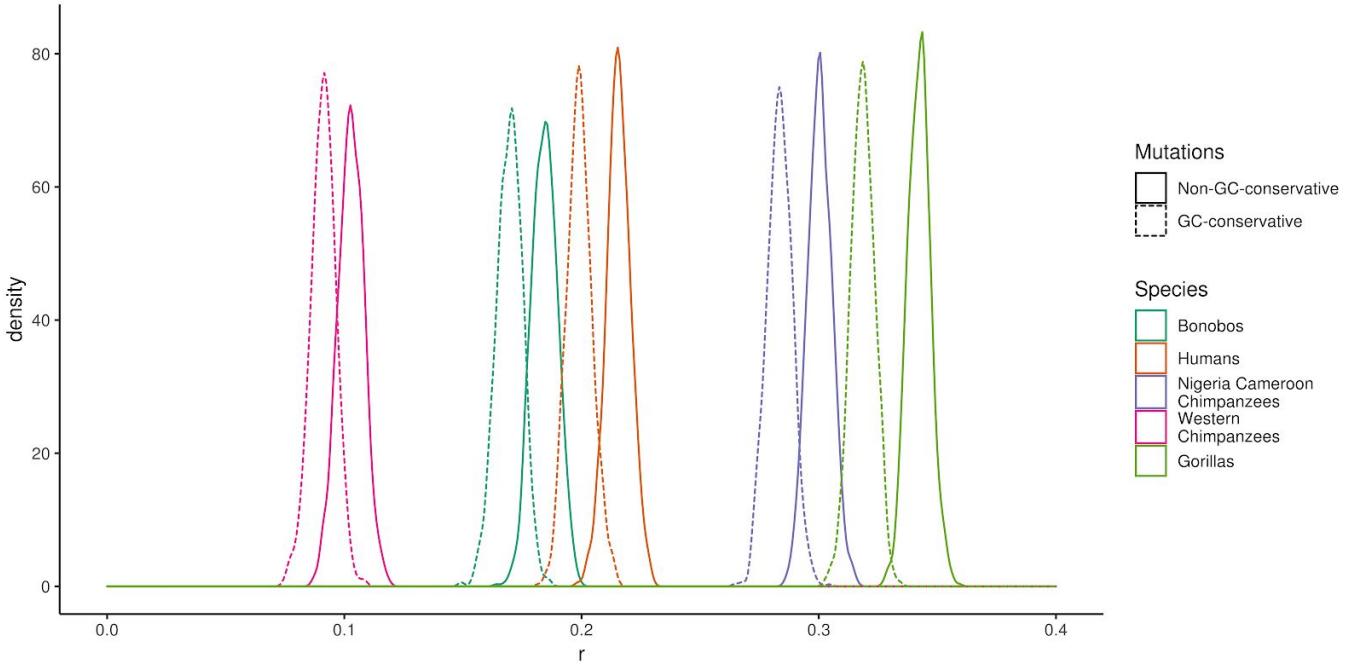

Figure 2. Relative effect of gBGC on the relationship between recombination rate and non-coding diversity. Distribution of the Spearman rank correlation coefficients (r) across 1000 bootstrap replicates for non-GC-conservative mutations (downsampled to match GC-conservative diversity) and GC-conservative mutations.

Efficiency of negative selection

Finally, we sought to investigate whether there is an influence of selection at linked sites on the efficiency of negative selection along the genomes by considering whether θ_N/θ_{NC} is correlated to the recombination rate, d_{NC} and the DCS. If there is variation in the effects of selection at linked sites across the homininae genome then we would expect θ_N/θ_{NC} to be negatively correlated to recombination rate, and positively correlated to the mutation rate and the DCS. Because we have very few coding sites in each window, and hence very few non-synonymous SNPs, we grouped windows together into 50 groups each containing on average ~1800 non-coding SNPs, and ~25 non-synonymous SNPs (in humans). We grouped windows by splitting the non-coding SNPs in each window into three independent estimates of the non-coding diversity using a hypergeometric sampling; we used one estimate to rank and group windows, the other as our unbiased estimate of the diversity in a group of windows and the third to estimate the ratio θ_N/θ_{NC}. We find that θ_N/θ_{NC} is generally negatively correlated to the rate of recombination and d_{NC}, and positively correlated to the DCS, when the correlations are performed individually (Table 4). The unexpected negative correlation with d_{NC} might be due to the positive correlation between the rate of recombination and d_{NC} (Table 1; Supplementary Table 2). In fact, none of these factors remain significant when we perform a multiple regression (Supplementary Table 3). This might be because there are strong correlations between recombination rate, d_{NC} and DCS when we group windows introducing multicollinearity – the VIF > 4.5 for each variable.

A)

Species	RR	d_{NC}	DCS
Human	-0.16 (-0.32, 0.02)	-0.19 (-0.35, 0.00)	0.15 (-0.02, 0.32)
NC chimpanzee	-0.24 (-0.40, -0.08)	-0.22 (-0.39, -0.05)	0.20 (0.02, 0.37)
W chimpanzee	-0.31 (-0.49, -0.13)	-0.27 (-0.44, -0.09)	0.20 (0.03, 0.38)
Bonobo	-0.18 (-0.36, -0.01)	-0.18 (-0.35, -0.01)	0.01 (-0.18, 0.20)
Gorilla	-0.14 (-0.32, 0.02)	-0.10 (-0.29, 0.07)	0.16 (-0.01, 0.33)

B)

Species	RR	d_{NC}	DCS
Human	-0.26 (-0.44, -0.07)	-0.21 (-0.39, -0.03)	0.12 (-0.09, 0.34)
NC chimpanzee	-0.13 (-0.32, 0.06)	-0.14 (-0.34, 0.06)	0.07 (-0.13, 0.28)
W chimpanzee	0.01 (-0.20, 0.22)	-0.10 (-0.29, 0.11)	0.01 (-0.20, 0.21)
Bonobo	-0.21 (-0.42, 0.02)	-0.19 (-0.39, 0.01)	0.12 (-0.10, 0.34)
Gorilla	-0.03 (-0.23, 0.15)	-0.03 (-0.23, 0.14)	-0.08 (-0.27, 0.14)

Table 4. The Spearman rank correlation coefficients of θ_N/θ_{NC} versus RR, d_{NC} and the DCS A) for all SNPs and B) for GC-conservative SNPs. Given in parentheses are the 95% confidence intervals by bootstrapping.

Recombination rate, d_{NC} and the DCS are likely to be crude predictors of the effects of selection at linked sites compared to the realized level of neutral diversity in a given window, after accounting for mutation rate variation. We, therefore, investigated whether θ_N/θ_{NC} is correlated to a measure of the effects of linked selection, the effective population size (N_e) of a window, estimated by dividing the non-coding diversity by our estimate of the mutation rate: i.e. θ_{NC}/d_{NC}. We find that θ_N/θ_{NC} is significantly negatively correlated to our measure of the local N_e in all homininae species (Figure 3). The slope of this relationship in a log-log scale is expected to equal to the negative value of the shape parameter of the DFE, if the DFE is

gamma distributed (Welch et al. 2008). The slope of this relationship in humans is -0.2 which is consistent with estimates of the DFE made from the site frequency spectrum (Boyko et al. 2008; Eyre-Walker & Keightley 2009; Kim et al. 2017). The other non-human homininae show equivalent slopes (Table 5) suggesting that, in these species, the DFE for new deleterious mutations is quite stable. As expected, when we consider GC-conservative mutations the slope estimates become noisier than when all mutations are used. For the two chimpanzee subspecies and gorillas the slope is not significantly different from 0, and for humans and bonobos the confidence intervals of the slope goes below -1.

Species	**All mutations**	**GC-conservative mutations**
Human	-0.20 (-0.31, -0.08)	-0.59 (-1.09, -0.13)
NC chimpanzee	-0.20 (-0.33, -0.07)	-0.17 (-0.60, 0.22)
W chimpanzee	-0.21 (-0.31, -0.11)	-0.04 (-0.43, 0.35)
Bonobo	-0.31 (-0.46, -0.15)	-0.71 (-1.20, -0.18)
Gorilla	-0.14 (-0.24, -0.03)	-0.05 (-0.38, 0.29)

Table 5. Mean slope (b) of $\log(\theta_N/\theta_{NC}) \sim \log(\theta_{NC}/d_{NC})$ linear regression for all SNPs and GC-conservative SNPs. Given in parentheses are the 95% confidence intervals by bootstrapping. Results grouping 50KB windows into 50 bins by non-coding diversity.

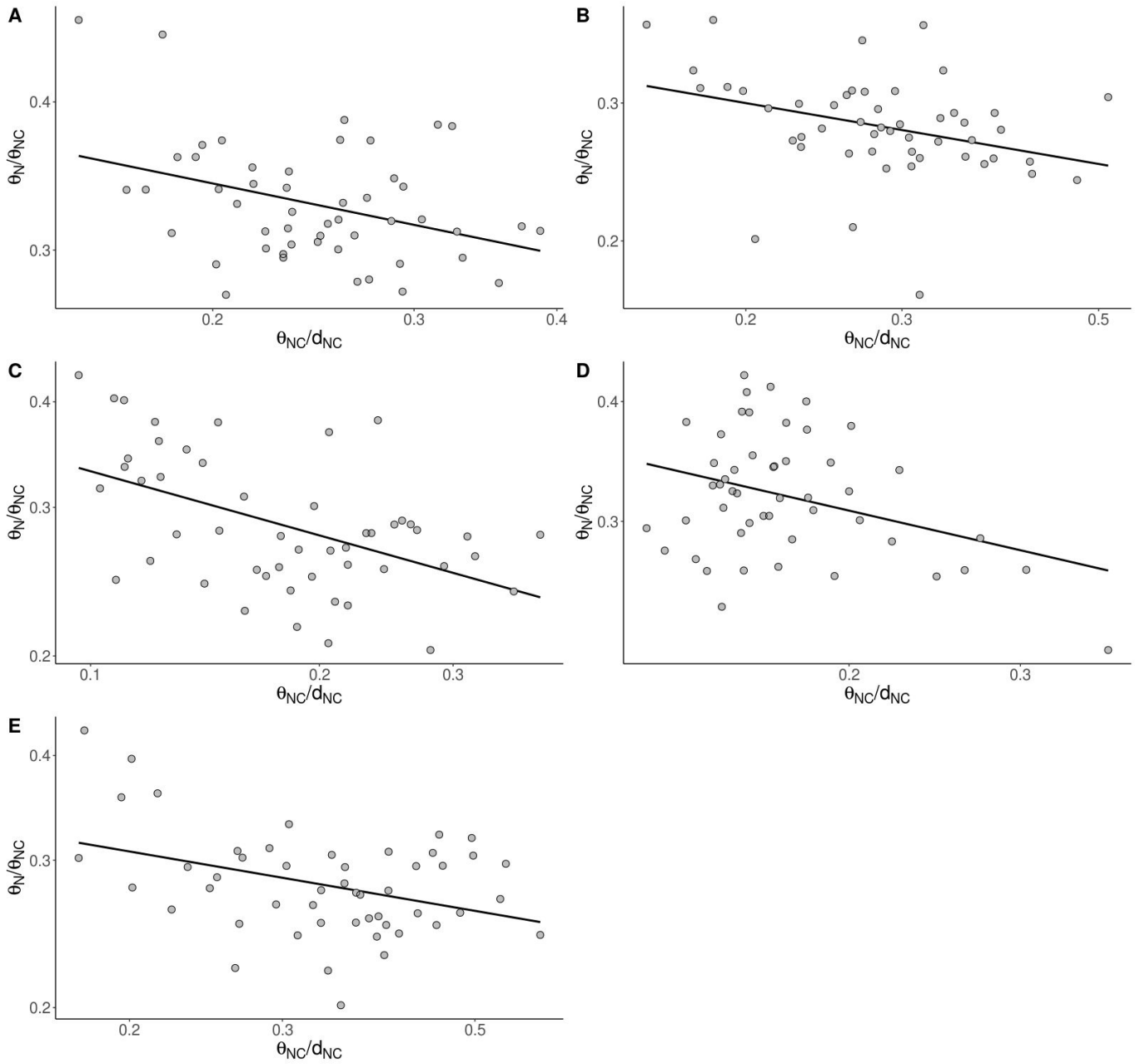

Figure 3. Relationship between θ_N/θ_{NC} and θ_{NC}/d_{NC} in a log-log scale for all mutations in A) humans (R^2 = 13%, *P*-value < 0.01), B) Nigeria-Cameroon chimpanzees (R^2 = 12%, *P*-value < 0.05), C) western chimpanzees (R^2 = 25%, *P*-value < 0.001), D) bonobos (R^2 = 12%, *P*-value < 0.05) and E) gorillas (R^2 = 14%, *P*-value < 0.01). Results grouping 50KB windows into 50 bins by non-coding diversity.

This relationship can be studied across species as in (Chen et al. 2017). Figure 4 shows the slope for all mutations and GC-conservative mutations. We do not correct for mutation rate variation because we assume that the genome-wide mutation rate is the same between species (although see Besenbacher et al. [2018]).

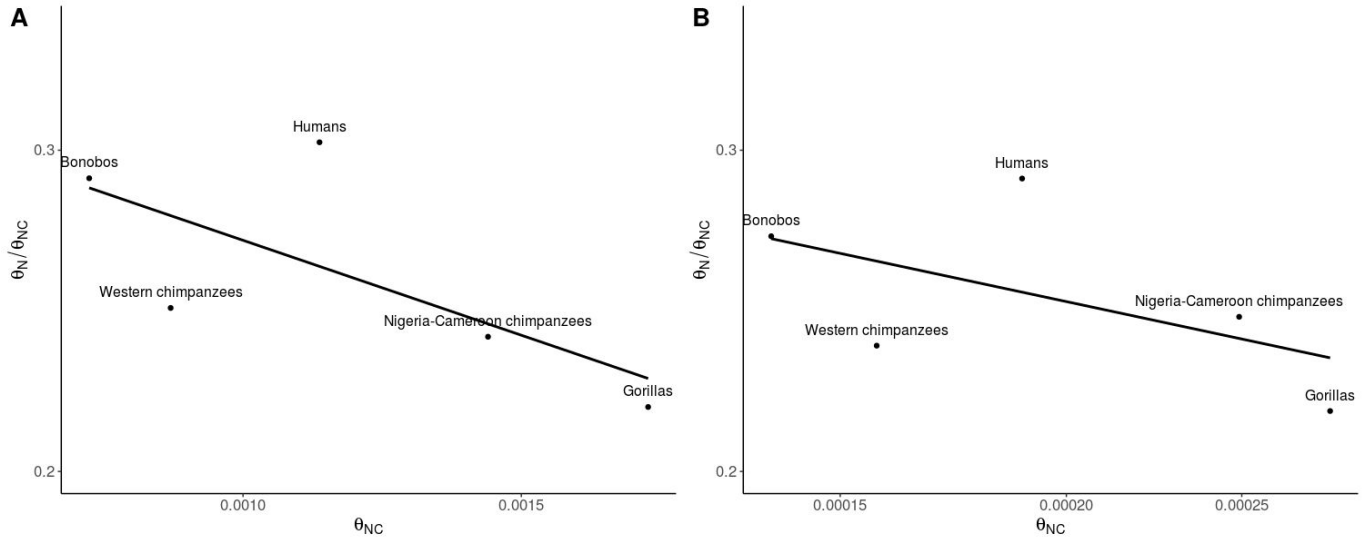

Figure 4. Relationship between θ_N/θ_{NC} and θ_{NC} in a log-log scale for A) all mutations (b =-0.29, R^2 = 50%, *P*-value = 0.18) and B) GC-conservative mutations (b = -0.21 , R^2 = 30%, *P*-value = 0.33).

Probably due to the few data points the slope between $\log(\theta_N/\theta_{NC})$ and $\log(\theta_{NC})$ across species is non-significant but it is close to the slope observed within a genome suggesting again a very leptokurtic DFE of new deleterious mutations. However, it is worth to note that along the genome the only possible source of variation in the efficiency of negative selection is selection at linked sites, not drift. Confusingly, selection at linked sites, particularly background selection, and genetic drift can be both modeled by a scalar reduction in the N_e (but see Messer & Petrov [2013] for the non-scalar reduction of the N_e in the presence of HH). In contrast, the source of the variation in the efficiency of negative selection between homininae species is probably driven mainly by genetic drift. Of note, when the N_e is very large, then an extreme version of selection at linked sites, genetic draft (Gillespie 2000) can

reduce the differences in the efficiency of natural selection and diversity between populations, but genetic draft is not a likely scenario in the homininae subfamily.

Discussion

We have investigated what genomic variables affect the level of genetic diversity across the homininae genome. We find highly significant effects of recombination rate, mutation rate and the density of conserved elements on levels of putatively neutral genetic diversity in most species, even if we restrict the analysis to GC-conservative polymorphisms, and hence remove the effects of gBGC. The positive correlation between diversity and our measure of the mutation rate is not surprising, given that we expect regions of the genome with high rates of mutation to have high levels of diversity, and previous analyses have suggested that much of the variation in diversity at the 100KB level can be explained in terms of variation in the mutation rate (Smith et al. 2018). Non-coding diversity is also positively correlated to the rate of recombination and negatively correlated to the DCS. These three genomic variables are stronger predictors of the levels of non-coding diversity in large N_e homininae, like gorillas than in the low N_e ones, such as bonobos. Hence, we find evidence of a greater impact of selection at linked sites in large homininae populations due to a decrease in the amount of genetic drift confirming previous reports across a broader panel of species (Corbett-Detig et al. 2015).

The level of non-synonymous diversity for both all mutations and GC-conservative mutations is strongly negatively correlated to the density of conserved sites and positively correlated to the recombination rate and our measure of the mutation rate in most homininae. Mutation rate is the weakest correlate, this might reflect a lack of power since we have few

non-synonymous SNPs in each of our windows or genuine differences in the coding and non-coding mutation rate. There is an alternative explanation. There might be an indirect attenuating effect when increasing mutation rate – the higher the rate of mutation the more selected mutations are produced and the greater the effect of selection at linked sites. It is unknown whether this side-effect of mutation will more strongly affect neutral or selected variants. From a perspective of interference selection, we wonder then why the density of conserved sites is a stronger correlate of non-synonymous diversity than the rate of recombination. We hypothesize that the observed fast evolution of fine-scale recombination rate patterns (Serre et al. 2005; Duret & Arndt 2008; Laayouni et al. 2011; Stevison et al. 2016) can explain this weaker correlation because the density of conserved sites is expected to evolve much slowly. This negative correlation with the density of conserved sites can also be explained through co-variation in the mutation rate and the density of conserved sites along the genome. To explore this mutationist hypothesis, we compiled *de novo* mutation (DNM) data from three studies that have discovered large numbers of DNMs in the sequencing of trios (Supplementary Analyses). We then investigated the relationship between putatively selected DNMs and the density of conserved sites. We find a significantly negative correlation between putatively selected DNMs and the density of conserved sites in one DNM dataset (Jónsson et al. 2017), a significantly positive correlation in another dataset (Wong et al. 2016), and no correlation in (Francioli et al. 2015) dataset. Such contradictory patterns have been observed before for other genomic variables and likely arise as a consequence of biases in the ascertainment of DNMs (Smith et al. 2018). Thus, it is not possible to conclude that the lower mutation rate in functionally rich regions is driving our strong negative correlation between θ_N and the density of conserved sites. The only result which we found consistent across DNMs datasets is that genome-wide, the ratio between putatively selected and putatively neutral DNM rates is ~⅔ (Supplementary Analyses). This result might suggest extreme negative

selection upon new mutations, biases in the ascertainment of DNMs shared across studies (Smith et al. 2018) and/or substantial differences in mutation rates between our set of putatively neutral and putatively selected sites. There is a third, non-mutually exclusive interpretation to explain the importance of the density of conserved sites on the levels of non-synonymous diversity. That is co-variation of the DFE of new amino acid mutations and gene density along the genome. In other words, if genes in highly gene-dense regions are more constrained (their DFE has a lower proportion of new nearly neutral and slightly deleterious mutations and a greater proportion of new strongly deleterious mutations) than "isolated" genes, then this will explain the negative correlation between the gene density and θ_N without invoking selection at linked sites or a lower mutation rate in functionally rich regions. The degree to which gene density (plus the associated regulatory sequences) and the DFE co-vary is therefore an interesting question for further investigation. However, the significantly negative slope between $\log(\theta_N/\theta_{NC}) \sim \log(\theta_{NC}/d_{NC})$ (see below) makes us think that selection at linked sites is indeed reducing the efficiency of purifying selection across the homininae genome.

It should be appreciated that the correlations between recombination rate, the density of conserved sites and genetic diversity may be an artifact of deficiencies in multiple regression. Recombination is known to be mutagenic in humans (Lercher & Hurst 2002; Hellmann et al. 2005; Pratto et al. 2014; Francioli et al. 2015; Arbeithuber et al. 2015; Halldorsson et al. 2019) and since d_{NC} is an imperfect measure of the mutation rate, this may allow the rate of recombination to remain significant in a multiple regression. Similarly, there is a negative correlation between the density of conserved sites and our measure of the mutation rate, d_{NC}. This could be due to regions of the genome with a high density of conserved sites having lower mutation rates, or due to the fact that we have not successfully masked all sites subject

to selection (i.e. regions of the genome with high density of conserved sites might have a higher density of other sites subject to recent selection, and these sites will decrease d_{NC}). To explore this possibility we have also investigated the relationship between the density of putatively neutral DNMs in humans and the density of conserved elements. Again, we find that the density of DNMs at putatively neutral sites is either significantly positively or uncorrelated to the density of conserved sites, depending on which dataset of DNMs is considered (Supplementary Analyses). As a consequence, we do not know whether the negative correlation between d_{NC} and density of conserved sites is a consequence of variation in the mutation rate with the density of conserved sites, or that d_{NC} is reduced in regions with a high density of conserved sites because there are sites subject to selection that we have not masked.

However, we find additional evidence that selection at linked sites affects the efficiency of purifying selection. First, we find that our measure of the efficiency of purifying selection, θ_N/θ_{NC}, is negatively correlated to our estimate of the rate of recombination and positively correlated to the density of conserved sites in a bivariate analysis. Unfortunately, when we group data we find that our genomic variables are strongly correlated to each other and so it is not possible to disentangle which variable is actually correlated to θ_N/θ_{NC} – i.e. there is a problem of multicollinearity in our multiple regression. Second, we find that θ_N/θ_{NC} is negatively correlated to a measure of the effective population size of a window, θ_{NC}/d_{NC}. We find that the slope of the relationship in a log-log scale is consistent with the estimated shape parameter of the DFE in humans (assuming the DFE is gamma distributed). This is in contrast to what has been observed in *Drosophila melanogaster*, in which the slope of the relationship between $\log(\theta_N/\theta_S)$ versus $\log(\theta_S)$ is significantly steeper than expected given an estimate of the DFE estimated from the site frequency spectrum; a similar pattern is apparent between

species (Chen et al. 2017; James et al. 2017). Castellano et al. (2018) considered a number of explanations for this and concluded that it was most likely due to genetic hitch-hiking; they showed by simulation that hitch-hiking increases the slope of the relationship, because deleterious genetic variation recovers more rapidly after a hitch-hiking event than neutral genetic variation (Gordo & Dionisio 2005; Pennings et al. 2014; Brandvain & Wright 2016; Do et al. 2015). This is consistent with the high rates of adaptive evolution observed in Drosophila (Smith & Eyre-Walker 2002; Andolfatto 2005; Eyre-Walker & Keightley 2009; Enard et al. 2014). Rates of adaptive evolution seem to be lower in humans than in Drosophila (Gossmann et al. 2012; Galtier 2016) which is therefore consistent with the fact that the slope is similar to the shape parameter of the gamma distribution. However, note that Zhen et al. (2018) have reported higher rates of adaptive evolution in humans than in Drosophila when accounting for ancestral population size. It is likely that most of the variation in θ_N/θ_{NC} along these genomes reflects variation in the efficiency of selection caused by selection at linked sites, mainly driven by weakly deleterious mutations. We have also shown some evidence for the variation in the efficiency of negative selection between species consistent with the Nearly Neutral Theory (Ohta 1992). Thus, our results show that genetic drift and selection at linked sites are two non-mutually exclusive sources of stochasticity acting at different scales. Importantly, the relative contribution of these two stochastic evolutionary forces to the patterns of DNA diversity within and between species is eventually determined by the effective population size.

Finally, we would like to highlight that interference among slightly deleterious mutations can potentially lead to indirect selection on recombination modifiers (Otto & Lenormand 2002; Coop & Przeworski 2007). This, therefore, might explain the rapid evolution of the fine-scale recombination rate patterns that have been observed in humans and non-human homininae

(Stevison et al. 2016). Although this strong divergence primarily results from the rapid evolution of the hotspot-defining PRDM9 protein, it is possible that other local recombination modifiers contribute as well. Note that the recombinogenic PRDM9 protein appears to be responsible for around 40% of the recombination hotspots in humans and mice (Myers et al. 2008; Webb et al. 2008; Baudat et al. 2010). The mechanism(s) responsible for the rest of the hotspots remain to be identified. Thus, we think indirect selection on recombination modifiers might be one important mechanism to explain the rapid turnover of a substantial amount of recombination hotspots. A prediction of this model is that the turnover of recombination hotspots should be faster in regions of the genome with a greater number of potentially selected mutations, particularly in regions under a high rate of strongly beneficial and many slightly deleterious mutations (Hartfield et al. 2010).

Conclusions

We show that non-coding and non-synonymous diversity are positively correlated to both mutation rate and recombination rate, while the density of conserved sites is associated with low levels of genetic diversity in all homininae. This result is not affected by gBGC. The positive correlation with the rate of recombination and the negative correlation with the density of conserved sites is consistent with variation in the intensity of selection at linked sites along the genome. While the positive correlation with the mutation rate just indicates that the higher the number of new mutations per generation the higher the level of genetic diversity. We find a negative correlation between the ratio of the number of non-synonymous to non-coding polymorphisms and a measure of the effective population size across the homininae genome which suggests pervasive interference selection, mainly, among weakly deleterious variants. However, there is evidence that genetic drift is decreasing the impact of selection at linked sites in small homininae populations compared to larger ones. Finally, an interesting implication of our study is that indirect selection on recombination modifiers might help to explain the rapid turnover of a substantial fraction of recombination hotspots in humans and their closest living relatives.

Acknowledgments

We thank D. Weghorn and three anonymous reviewers for comments. This work was supported by the Danish Council For Independent Research (grant number 4181-00358).

Bibliography

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in Drosophila. Nature. 437:1149–1152. doi: 10.1038/nature04107.

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. 112:2109–2114. doi: 10.1073/pnas.1416622112.

Auton A et al. 2012. A fine-scale chimpanzee genetic map from population sequencing. Science. 336:193–198. doi: 10.1126/science.1216872.

Baudat F et al. 2010. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science. 327:836–840. doi: 10.1126/science.1183439.

Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature. 356:519–520. doi: 10.1038/356519a0.

Besenbacher S, Hvilsom C, Marques-Bonet T, Mailund T, Schierup MH. 2018. Direct estimation of mutations in great apes reveals significant recent human slowdown in the yearly mutation rate. doi: 10.1101/287821.

Boyko AR et al. 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. 4:e1000083. doi: 10.1371/journal.pgen.1000083.

Brandvain Y, Wright SI. 2016. The Limits of Natural Selection in a Nonequilibrium World. Trends Genet. 32. doi: 10.1016/j.tig.2016.01.004.

Bullaughey K, Przeworski M, Coop G. 2008. No effect of recombination on the efficacy of natural selection in primates. Genome Res. 18:544–554. doi: 10.1101/gr.071548.107.

Cagan A et al. 2016. Natural Selection in the Great Apes. Mol. Biol. Evol. 33:3268–3283. doi: 10.1093/molbev/msw215.

Capra JA, Hubisz MJ, Kostka D, Pollard KS, Siepel A. 2013. A model-based analysis of GC-biased gene conversion in the human and chimpanzee genomes. PLoS Genet. 9:e1003684. doi: 10.1371/journal.pgen.1003684.

Castellano D, James J, Eyre-Walker A. 2018. Nearly Neutral Evolution across the Drosophila melanogaster Genome. Mol. Biol. Evol. doi: 10.1093/molbev/msy164.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics. 134:1289–1303. https://www.ncbi.nlm.nih.gov/pubmed/8375663.

Chen J, Glémin S, Lascoux M. 2017. Genetic Diversity and the Efficacy of Purifying Selection across Plant and Animal Species. Mol. Biol. Evol. 34:1417–1428. doi: 10.1093/molbev/msx088.

Cingolani P et al. 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 . 6:80–92. doi: 10.4161/fly.19695.

Coop G, Przeworski M. 2007. An evolutionary view of human recombination. Nat. Rev. Genet.

8:23–34. doi: 10.1038/nrg1947.

Corbett-Detig RB, Hartl DL, Sackton TB. 2015. Natural Selection Constrains Neutral Diversity across A Wide Range of Species Barton, NH, editor. PLoS Biol. 13:e1002112. doi: 10.1371/journal.pbio.1002112.

Cutter AD, Payseur BA. 2013. Genomic signatures of selection at linked sites: unifying the disparity among species. Nat. Rev. Genet. 14:262–274. doi: 10.1038/nrg3425.

Davydov EV et al. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput. Biol. 6:e1001025. doi: 10.1371/journal.pcbi.1001025.

Do R et al. 2015. No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. Nat. Genet. 47. doi: 10.1038/ng.3186.

Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet. 4:e1000071. doi: 10.1371/journal.pgen.1000071.

Enard D, Messer PW, Petrov DA. 2014. Genome-wide signals of positive selection in human evolution. Genome Res. 24:885–895. doi: 10.1101/gr.164822.113.

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. doi: 10.1093/molbev/msp119.

Francioli LC et al. 2015. Genome-wide patterns and properties of de novo mutations in humans. Nat. Genet. 47:822–826. doi: 10.1038/ng.3292.

Galtier N. 2016. Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis. PLoS Genet. 12:e1005774. doi: 10.1371/journal.pgen.1005774.

Gillespie JH. 2000. Genetic Drift in an Infinite Population: The Pseudohitchhiking Model. Genetics. 155:909–919. http://www.genetics.org/content/155/2/909.long.

Glémin S et al. 2015. Quantification of GC-biased gene conversion in the human genome. Genome Res. 25. doi: 10.1101/gr.185488.114.

Glémin S. 2010. Surprising fitness consequences of GC-biased gene conversion: I. Mutation load and inbreeding depression. Genetics. 185:939–959. doi: 10.1534/genetics.110.116368.

Gordo I, Dionisio F. 2005. Nonequilibrium model for estimating parameters of deleterious mutations. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. 71:031907. doi: 10.1103/PhysRevE.71.031907.

Gossmann TI, Keightley PD, Eyre-Walker A. 2012. The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. Genome Biol. Evol. 4:658–667. doi: 10.1093/gbe/evs027.

Halldorsson BV et al. 2019. Characterizing mutagenic effects of recombination through a sequence-level genetic map. Science. 363. doi: 10.1126/science.aau1043.

Hartfield M, Otto SP, Keightley PD. 2010. The role of advantageous mutations in enhancing the evolution of a recombination modifier. Genetics. 184:1153–1164. doi:

10.1534/genetics.109.112920.

Hellmann I et al. 2005. Why do human diversity levels vary at a megabase scale? Genome Res. 15:1222–1231. doi: 10.1101/gr.3461105.

Hussin JG et al. 2015. Recombination affects accumulation of damaging and disease-associated mutations in human populations. Nat. Genet. 47. doi: 10.1038/ng.3216.

James J, Castellano D, Eyre-Walker A. 2017. DNA sequence diversity and the efficiency of natural selection in animal mitochondrial DNA. Heredity . 118:88–95. doi: 10.1038/hdy.2016.108.

Jónsson H et al. 2017. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. Nature. 549:519–522. doi: 10.1038/nature24018.

Kim BY, Huber CD, Lohmueller KE. 2017. Inference of the Distribution of Selection Coefficients for New Nonsynonymous Mutations Using Large Samples. Genetics. 206:345–361. doi: 10.1534/genetics.116.197145.

Kong A et al. 2002. A high-resolution recombination map of the human genome. Nat. Genet. 31:241–247. doi: 10.1038/ng917.

Kong A et al. 2010. Fine-scale recombination rate differences between sexes, populations and individuals. Nature. 467:1099–1103. doi: 10.1038/nature09525.

Laayouni H et al. 2011. Similarity in recombination rate estimates highly correlates with genetic differentiation in humans. PLoS One. 6:e17913. doi: 10.1371/journal.pone.0017913.

Lachance J, Tishkoff SA. 2014. Biased gene conversion skews allele frequencies in human populations, increasing the disease burden of recessive alleles. Am. J. Hum. Genet. 95:408–420. doi: 10.1016/j.ajhg.2014.09.008.

Lercher MJ, Hurst LD. 2002. Human SNP variability and mutation rate are higher in regions of high recombination. Trends Genet. 18:337–340. https://www.ncbi.nlm.nih.gov/pubmed/12127766.

McVean GAT et al. 2004. The fine-scale structure of recombination rate variation in the human genome. Science. 304:581–584. doi: 10.1126/science.1092500.

McVean GAT, Charlesworth B. 1999. A population genetic model for the evolution of synonymous codon usage: patterns and predictions. Genet. Res. 74:145–158. https://www.cambridge.org/core/journals/genetics-research/article/population-genetic-model-for-the-evolution-of-synonymous-codon-usage-patterns-and-predictions/785E506112E3C9505F6EFFF7D262EC68 (Accessed September 22, 2018).

McVicker G, Gordon D, Davis C, Green P. 2009. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. 5:e1000471. doi: 10.1371/journal.pgen.1000471.

Messer PW, Petrov D a. 2013. Frequent adaptation and the McDonald-Kreitman test. Proc. Natl. Acad. Sci. U. S. A. 110:8615–8620. doi: 10.1073/pnas.1220835110.

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. Science. 310:321–324. doi:

10.1126/science.1117196.

Myers S, Freeman C, Auton A, Donnelly P, McVean G. 2008. A common sequence motif associated with recombination hot spots and genome instability in humans. Nat. Genet. 40:1124–1129. doi: 10.1038/ng.213.

Nam K et al. 2017. Evidence that the rate of strong selective sweeps increases with population size in the great apes. Proc. Natl. Acad. Sci. U. S. A. 114:1613–1618. doi: 10.1073/pnas.1605660114.

Necşulea A et al. 2011. Meiotic recombination favors the spreading of deleterious mutations in human populations. Hum. Mutat. 32:198–206. doi: 10.1002/humu.21407.

Ohta T. 1992. The Nearly Neutral Theory of Molecular Evolution. Annu. Rev. Ecol. Syst. 23:263–286. doi: 10.1146/annurev.es.23.110192.001403.

Otto SP, Lenormand T. 2002. Resolving the paradox of sex and recombination. Nat. Rev. Genet. 3:252–261. doi: 10.1038/nrg761.

Payseur BA, Nachman MW. 2002. Gene density and human nucleotide polymorphism. Mol. Biol. Evol. 19:336–340. doi: 10.1093/oxfordjournals.molbev.a004086.

Pennings PS, Kryazhimskiy S, Wakeley J. 2014. Loss and recovery of genetic diversity in adapting populations of HIV. PLoS Genet. 10:e1004000. doi: 10.1371/journal.pgen.1004000.

Phung TN, Huber CD, Lohmueller KE. 2016. Determining the Effect of Natural Selection on Linked Neutral Divergence across Species. PLoS Genet. 12:e1006199. doi: 10.1371/journal.pgen.1006199.

Prado-Martinez J et al. 2013. Great ape genetic diversity and population history. Nature. 499:471–475. doi: 10.1038/nature12228.

Pratto F et al. 2014. Recombination initiation maps of individual human genomes. Science. 346:1256442. doi: 10.1126/science.1256442.

Serre D, Nadon R, Hudson TJ. 2005. Large-scale recombination rate patterns are conserved among human populations. Genome Res. 15:1547–1552. doi: 10.1101/gr.4211905.

Siepel A et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15:1034–1050. doi: 10.1101/gr.3715005.

Smith JM, Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genet. Res. 23:23–35. https://www.ncbi.nlm.nih.gov/pubmed/4407212.

Smith NGC, Eyre-Walker A. 2002. Adaptive protein evolution in Drosophila. Nature. 415:1022–1024. doi: 10.1038/4151022a.

Smith TCA, Arndt PF, Eyre-Walker A. 2018. Large scale variation in the rate of germ-line de novo mutation, base composition, divergence and diversity in humans. PLoS Genet. 14:e1007254. doi: 10.1371/journal.pgen.1007254.

Song L et al. 2011. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res. 21:1757–1767. doi:

10.1101/gr.121541.111.

Stevison LS et al. 2016. The Time Scale of Recombination Rate Evolution in Great Apes. Mol. Biol. Evol. 33:928–945. doi: 10.1093/molbev/msv331.

Terekhanova NV, Seplyarskiy VB, Soldatov RA, Bazykin GA. 2017. Evolution of Local Mutation Rate and Its Determinants. Mol. Biol. Evol. 34:1100–1109. doi: 10.1093/molbev/msx060.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7:256–276. https://www.ncbi.nlm.nih.gov/pubmed/1145509.

Webb AJ, Berg IL, Jeffreys A. 2008. Sperm cross-over activity in regions of the human genome showing extreme breakdown of marker association. Proc. Natl. Acad. Sci. U. S. A. 105:10471–10476. doi: 10.1073/pnas.0804933105.

Welch JJ, Eyre-Walker A, Waxman D. 2008. Divergence and polymorphism under the nearly neutral theory of molecular evolution. J. Mol. Evol. 67:418–426. doi: 10.1007/s00239-008-9146-9.

Wong WSW et al. 2016. New observations on maternal age effect on germline de novo mutations. Nat. Commun. 7:10486. doi: 10.1038/ncomms10486.

Zhen Y, Huber CD, Davies RW, Lohmueller KE. 2018. Stronger and higher proportion of beneficial amino acid changing mutations in humans compared to mice and flies. bioRxiv. 427583. doi: 10.1101/427583.