Long title: Human demographic history has amplified the effects background selection across the genome

Short title: Background selection and demography in humans

Raul Torres[1], Zachary A. Szpiech[2], and Ryan D. Hernandez[2,3,4,5*]

[1] Biomedical Sciences Graduate Program, University of California San Francisco, San Francisco, CA, United States of America

[2] Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, United States of America

[3] Institute for Human Genetics, University of California San Francisco, San Francisco, CA, United States of America

[4] Institute for Computational Health Sciences, University of California San Francisco, San Francisco, CA, United States of America

[5] Quantitative Biosciences Institute, University of California San Francisco, San Francisco, CA, United States of America

* Corresponding author

E-mail: ryan.hernandez@ucsf.edu (RDH)

Abstract

Natural populations often grow, shrink, and migrate over time. Demographic processes such as these can impact genome-wide levels of genetic diversity. In addition, genetic variation in functional regions of the genome can be altered by natural selection, which drives adaptive mutations to higher frequency or purges deleterious ones. Such selective processes impact not only the sites directly under selection but also nearby neutral variation through genetic linkage. While there is extensive literature examining the impact of linked selection (i.e., genetic hitchhiking in the context of positive selection and background selection in the context of deleterious variants) at demographic equilibrium, less is known about how non-equilibrium demographic processes impact patterns of linked selection. Utilizing a global sample of human whole-genome sequences from the Thousand Genomes Project and extensive simulations, we investigate how non-equilibrium demographic processes magnify and dampen the consequences of background selection (BGS) across the human genome. We observe that, compared to Africans, non-African populations have experienced larger proportional decreases in neutral genetic diversity in regions of strong BGS. We replicate these findings in admixed populations by showing that non-African ancestral components of the genome have been impacted more severely in regions of strong BGS. We attribute these differences to the strong population bottlenecks that non-Africans experienced as they migrated out of Africa and throughout the globe. Furthermore, we observe a strong correlation between F_{ST} and BGS, suggesting a stronger rate of genetic drift in regions of strong BGS. Forward simulations of human demographic history and BGS support these observations. Our results show that non-equilibrium demography significantly alters the conse-

quences of BGS and support the need for more work investigating the dynamic process of multiple evolutionary forces operating in concert.

Author summary

Patterns of genetic diversity within a species are impacted at broad and fine scales by population size changes ("demography") and natural selection. From both population genetics theory and observation of genomic sequence data, it is known that population size changes can impact genome-wide average neutral genetic diversity. Additionally, natural selection can impact neutral genetic diversity regionally across the genome through the process of linked selection. During this process, natural selection acting on adaptive or deleterious variants in the genome will also impact diversity at linked neutral sites. However, less is known about the dynamic changes to diversity that occur in regions impacted by linked selection when a population undergoes a size change. We characterize these dynamic changes using human sequence data, focusing on regions of the genome experiencing linked selection that is caused by deleterious variation (called "background selection"). We find that the population size changes experienced by humans have shaped the consequences of linked selection in the human genome. In particular, population contractions, such as those experienced by non-Africans, have disproportionately decreased neutral diversity in regions of the genome experiencing strong BGS, resulting in large differences between African and non-African populations.

Introduction

Genetic diversity in a species is determined through the complex interplay of mutation, demography, genetic drift, and natural selection. These evolutionary forces operate in concert to shape patterns of diversity at both the local scale and genome-wide scale. For example, in recombining species, levels of genetic diversity are distributed heterogeneously across the genome as peaks and valleys that are often correlated with recombination rate and generated by past or ongoing events of natural selection [1]. But at the genome-wide scale, average levels of genetic diversity are primarily impacted by population size changes, yielding signatures of diversity that are a function of a population's demographic history [2]. These patterns of diversity may also yield information for inferring past events of natural selection and population history, giving valuable insight into how populations have evolved over time [3–8]. With recent advancements in sequencing technology yielding whole-genome data from thousands of individuals from species with complex evolutionary histories [9,10], formal inquiry into the interplay of demography and natural selection and testing of whether demographic effects act uniformly across the genome as a function of natural selection is now possible.

In the past decade, population genetic studies have shed light on the pervasiveness of dynamic population histories in shaping overall levels of genetic diversity across different biological species. For example, multiple populations have experienced major population bottlenecks that have resulted in decreased levels of genome-wide diversity. Evidence for population bottlenecks exists in domesticated species such as cattle [11], dogs [12], and rice [13], and in natural populations such as *Drosophila melanogaster* [14–16], rhesus macaque [17], and humans [18,19]. Notably, population bottlenecks

leave inordinately long lasting signatures of decreased diversity, which may be depressed long after a population has recovered to its ancestral size [20].

Locally (i.e., regionally) across the genome, the action of natural selection can also lead to distinct signatures of decreased genetic diversity (although some forms of selection, such as balancing selection, can increase genetic diversity [21]). For example, mutations with functional effects may be removed from the population due to purifying selection or fix due to positive selection, thereby resulting in the elimination of genetic diversity at the site. But while sites under direct natural selection in the genome represent only a small fraction of all sites genome-wide, the action of natural selection on these selected sites can have far-reaching effects across neutral sites in the genome due to linkage. Under positive selection, genetic hitchhiking [22] causes variants lying on the same haplotype as the selected allele to rise to high frequency during the selection process. Conversely, under purifying selection, background selection [23] causes linked neutral variants to decrease in frequency or to be removed from the population. Both of these processes of linked selection result in decreased neutral genetic diversity around the selected site. Recombination can decouple neutral sites from selected sites in both cases and neutral diversity tends to increase toward its neutral expectation as genetic distance from selected sites increases [24].

Evidence for genetic hitchhiking and background selection has been obtained from the genomes of several species, including *Drosophila melanogaster* [25–30], wild and domesticated rice [31,32], nematode [33,34], humans [3,6,35–39], and others (see [1] for a review). While the relative contributions of genetic hitchhiking and background selection to shaping patterns of human genomic diversity have been actively debated

[37,40–42], the data strongly support a large role for background selection in shaping genome-wide patterns of neutral genetic variation [38,39]. Indeed, recent arguments have been made in favor of background selection being treated as the null model when investigating the impact of linked selection across recombining genomes [1,29,42–44]. To build on this movement, our investigation will focus on aspects of background selection (BGS) in humans.

The impact of BGS has been quantified across the human genome [6], which suggests that neutral genetic diversity in the human autosomes has been reduced 19-26% because of BGS. However, it is not obvious whether populations that have experienced different demographic histories, such as African and non-African humans, should exhibit similar effects in regions of BGS. Early work resulted in the expression $\pi \approx 4f_0N_e\mu$ [23], which suggests that the expected level of diversity with background selection would be proportional to the neutral expectation (with proportionality constant f_0 being a function of the rates of deleterious mutation and recombination). While demographic forces will impact N_e, this classic model implies that the relative reduction in diversity due to BGS may be insensitive to different demographic histories. Recent work has demonstrated that this assumption may hold if selection is strong enough (or populations are large enough) such that mutation-selection balance is maintained [45,46]. However, humans and several other natural populations likely do not meet such criteria, and are rarely at demographic equilibrium. Therefore, other evolutionary forces, such as genetic drift, may perturb genetic diversity in regions of BGS that are unaccounted for in these models.

While little is known about the impact of demography on BGS' effects, recent studies have suggested that alleles directly under natural selection experience non-linear dynamics in the context of non-equilibrium demography. In general, the equilibrium frequency of an allele is dependent on its fitness effect, with deleterious alleles having lower equilibrium frequencies than neutral alleles. After a population size change, deleterious alleles tend to change frequency faster than neutral alleles, allowing them to reach their new equilibrium frequency at a faster rate [47,48]. This can result in relative differences in deleterious allele frequencies when comparing populations with different demographic histories. Such effects are especially apparent in populations suffering bottlenecks [49] and have been tested and observed between different human populations [50–52]. We hypothesized that these non-equilibrium dynamics could also amplify the effects of background selection, with classic models being poorly suited to describe such impacts in non-equilibrium populations. In the case of population bottlenecks, regions under BGS may respond more rapidly than neutral regions to such demographic events, if the rate of removal of genetic diversity in regions under BGS is also higher.

To investigate the impact of non-equilibrium dynamics in regions of BGS, we measure patterns of average pairwise neutral genetic diversity (π), or neutral heterozygosity if the population was admixed, as a function of the strength of BGS within a global set of human populations from phase 3 of the Thousand Genomes Project (TGP) [9]. We focus particularly on the ratio of neutral diversity in regions of strong BGS to regions of weak BGS, which we term "relative diversity." We also investigate how genetic differentiation between TGP populations (as measured by F_{ST}) is shaped by BGS strength. We find substantial differences in relative diversity between populations, which we at-

138 tribute to their non-equilibrium demographics. We confirm that the interplay of demogra-
139 phy and BGS can explain the differences of relative diversity across human populations
140 with simulations incorporating a parametric demographic model of human history [7]
141 with and without a model of BGS. Specifically, our simulations of non-African human
142 demography show that the out-of-Africa population bottleneck and European-Asian split
143 population bottleneck each coincide with a distinct drop in relative diversity. Further-
144 more, we demonstrate that back migration from Europeans and Asians into Africa re-
145 introduces sufficient deleterious variation to impact BGS strength, leading to decreased
146 relative diversity in Africans. These results demonstrate the strong impact that changing
147 demography has on perturbing levels of diversity in regions under BGS and have impli-
148 cations for population genetic studies seeking to characterize linked selection across
149 any species or population that is not at demographic equilibrium.
150

151 ## Results

152 ### Differential impact of BGS across human populations

153 We measured mean pairwise genetic diversity (π) among the 20 non-admixed
154 populations from the phase 3 TGP data set, consisting of 5 populations each from 4
155 continental groups: Africa (AFR), Europe (EUR), South Asia (SASN), and East Asia
156 (EASN) (Table S11 in Supporting information for population labels and grouping). After
157 a set of stringent filters were applied to all 20 populations to identify a high-quality set of
158 putatively neutral sites in the genome (see Materials and Methods), sites were divided
159 into quantile bins based on their BGS coefficient (termed "*B*"), which represents the in-
160 ferred strength of BGS from McVicker et al. [6] (see Materials and Methods). For our

161 initial set of analyses, we focused on the bins corresponding to the 1% of sites under
162 strongest BGS (or lowest *B*) and the 1% of sites under weakest BGS (or highest *B*).
163 Mean diversity (normalized by divergence with rhesus macaque) within these bins for
164 each population is shown in Figs 1A-B. As expected, normalized diversity was highest
165 in African populations and lowest in East Asian populations across both 1% *B* quantile
166 bins.

167 To obtain estimates of the reduction in diversity in regions of strong BGS for each
168 population, we calculated a statistic called "relative diversity" (analogous to π/π_0 in the
169 BGS literature; [23,53]), which we defined as the ratio of normalized diversity in the
170 strongest 1% BGS bin to normalized diversity in the weakest 1% BGS bin. Fig 1C
171 shows that relative diversity was lower in non-African populations (0.348-0.365 for non-
172 Africans, 0.396-0.408 for Africans), suggesting higher rates of genetic diversity reduc-
173 tion in strong BGS regions in non-African populations compared to African populations.
174 Furthermore, the greatest reduction was observed among East Asian populations (CDX,
175 JPT, and CHS populations).

176 To characterize these effects across a broader distribution of BGS strengths, we
177 grouped populations together according to their continental group (i.e., African, Europe-
178 an, South Asian, and East Asian, see Table S11 in Supporting information for a detailed
179 description) and estimated relative diversity at neutral sites for each of the continental
180 groups in bins corresponding to the bottom 1%, 5%, 10%, and 25% quantiles of *B* (note
181 these partitions were not disjoint). As expected, relative diversity increased for all conti-
182 nental groups as the bins became more inclusive (Fig 2B), reflecting a reduced impact
183 on the reduction of diversity due to BGS. We also observed that non-African continental

184 groups consistently had a lower relative diversity compared to African groups, demon-
185 strating that the patterns we observed in the most extreme BGS regions also held for
186 weaker BGS regions. Interestingly, we observed a consistent trend of rank order for rel-
187 ative diversity between the different continental groups for each quantile bin, with the
188 East Asian group experiencing the greatest reduction of relative diversity, followed by
189 the South Asian, European, and African groups. This result suggested a stronger effect
190 for demography on the diversity-reducing effect of BGS for those populations experienc-
191 ing the strongest bottlenecks. However, the observed differences in relative diversity
192 between non-African and African continental groups became less pronounced as the
193 bins became more inclusive (Fig 2B). These effects remained even after we controlled
194 for the effects of GC-biased gene conversion and recombination hotspots (Figs S2 and
195 S4) or if we did not normalize diversity by divergence (Figs S3 and S5).
196

197 **Recent admixture has not altered the impact of BGS**

198 We then investigated whether the effects of BGS have remained consistent
199 across human populations that have experienced recent admixture. To do so, we tested
200 for the effect of BGS on normalized and relative diversity in the 6 admixed TGP popula-
201 tions (ASW, ACB, CLM, MXL, PEL, and PUR). We first used the local ancestry tracks
202 inferred by the phase 3 TGP consortium to divide up admixed samples into genomic
203 segments that are homozygous for a specific local ancestry (i.e., African, European, or
204 Native American). These segments were then collated across all admixed samples (see
205 Materials and Methods), irrespective of their source population. We then calculated
206 normalized and relative heterozygosity across the collated segments as a function of

207 the BGS quantile bins described in the previous section. For comparison, heterozygosi-
208 ty was also calculated for each TGP continental group. Across all BGS quantile bins,
209 normalized diversity (heterozygosity/divergence) in African and European ancestry
210 segments closely matched the values observed in their non-admixed counterparts (Fig
211 3A). However, normalized diversity was significantly lower in the Native American an-
212 cestry segments of admixed individuals than in the East Asian continental group (Fig
213 3A). This was expected given the more recent divergence of Native American popula-
214 tions and the strong population bottleneck they experienced migrating into the Americas
215 [18,54,55].

216 Patterns of relative diversity for each of the ancestries also largely recapitulated
217 the patterns observed in their continental group counterparts across all BGS quantile
218 bins, with the largest reductions in relative diversity occurring for the Native American
219 and European ancestral segments (Fig 3B). These patterns were similar to the broader
220 analyses of the 20 non-admixed populations described above, with a consistent rank
221 order of decreasing relative diversity observed for African, European, and Native Ameri-
222 can ancestral segments. However, for the strongest 1% BGS quantile bin, relative di-
223 versity in Native American ancestry segments was observed to be greater than for the
224 European continental group or European local ancestry segments, which was incon-
225 sistent with the other BGS quantile bins.

226

227 **BGS has shaped patterns of population differentiation**

228 To test if greater genetic drift in regions of strong BGS has contributed to the ob-
229 served greater losses of relative diversity in non-African populations and to understand

how BGS has impacted local patterns of population differentiation, we measured F_{ST} as a function of B. It has been shown that estimates of F_{ST} are affected by SNP ascertainment as well as the choice of computation method [56].To limit such biases, we followed the procedures recommended by Bhatia et al. [56] (see Materials and Methods) and only analyzed SNPs that were polymorphic in an outgroup human population (the Khoe-San). After filtering, a total of 3,497,105 sites were left for estimating F_{ST}. Sites were divided into 2% quantile bins based on the genome-wide distribution of B, and F_{ST} was calculated within each bin for all population pairs in which both populations were from different continental groups (150 total pairwise comparisons). We performed simple linear regression using B as an explanatory variable and F_{ST} as our dependent variable with the linear model $F_{ST} = \beta_0 + \beta_1 B + \varepsilon$. We found that B was significantly associated with an increase in local F_{ST} across the genome (Fig 4A, Table 1). When considering all 150 population comparisons (referred to as the "Global" estimate in Table 1), B explained 26.9% of the change in F_{ST} across the most extreme B values (i.e., when comparing weak BGS [B = 1] to the strong BGS [B = 0]). To test if the linear model was bi ased by outliers or bins with high influence, we also performed robust linear regression but still observed a significant contribution of B to F_{ST}, with a 28.3% change between the highest and lowest B value (Table S6 in Supporting information).

Table 1. Regression coefficient estimates for linear regression of F_{ST} on 2% quantile bins of *B*.

	AFR vs. EASN	AFR vs. EUR	AFR vs. SASN	EUR vs. SASN	EUR vs. EASN	SASN vs. EASN	Global
β_0 ± SEM (p-value)	0.2044 ± 0.0039 (< 1e-04)	0.1716 ± 0.0031 (< 1e-04)	0.1596 ± 0.0029 (< 1e-04)	0.0455 ± 0.0011 (< 1e-04)	0.1216 ± 0.0029 (< 1e-04)	0.0903 ± 0.0023 (< 1e-04)	0.1322 ± 0.0019 (< 1e-04)
β_1 ± SEM (p-value)	-0.0434 ± 0.0046 (< 1e-04)	-0.0358 ± 0.0037 (< 1e-04)	-0.0355 ± 0.0034 (< 1e-04)	-0.0098 ± 0.0013 (< 1e-04)	-0.0173 ± 0.0035 (< 1e-04)	-0.0261 ± 0.0027 (< 1e-04)	-0.0280 ± 0.0022 (< 1e-04)
r ± SEM	-0.8363 ± 0.0295	-0.7441 ± 0.0362	-0.7794 ± 0.0332	-0.3847 ± 0.0414	-0.6220 ± 0.0785	-0.5968 ± 0.0348	-0.1292 ± 0.0098

The first two rows give the regression coefficients for the linear model $F_{ST} = \beta_0 + \beta_1 B + \varepsilon$, where *B* represents the mean background selection coefficient for the bin being tested and F_{ST} is the estimated F_{ST} for all population comparisons within a particular pair of continental groups (given in the column header). The final column, "Global", gives the regression coefficients for the linear model applied to all pairwise population comparisons (150 total). The correlation coefficient, *r*, between *B* and F_{ST} for each comparison is shown in the bottom row. Standard errors of the mean (SEM) for β_0, β_1, and *r* were calculated from 1,000 bootstrap iterations (see Materials and Methods). P-values are derived from a two-sided t-test of the t-value for the corresponding regression coefficient.

Earlier studies using SNP array data have shown that F_{ST} and recombination rate are correlated in humans [57]. We could only partially replicate these findings when we conducted linear regression of F_{ST} as a function of recombination rate (ρ) (measured in 2% recombination rate quantile bins) with the linear model $F_{ST} = \beta_0 + \beta_1\rho + \varepsilon$. We observed that recombination rate only significantly predicts a change in F_{ST} across the genome for comparisons between South Asian and East Asian populations (Fig 4B, Table 2). This result remained unchanged when performing robust linear regression for the model (Table S7 in Supporting information).

Table 2. Regression coefficient estimates for linear regression of F_{ST} on 2% quantile bins of recombination rate.

	AFR vs. EASN	AFR vs. EUR	AFR vs. SASN	EUR vs. SASN	EUR vs. EASN	SASN vs. EASN	Global
β_0 ± SEM (p-value)	0.1688 ± 0.0007 (< 1e-04)	0.1422 ± 0.0006 (< 1e-04)	0.1305 ± 0.0006 (< 1e-04)	0.0373 ± 0.0002 (< 1e-04)	0.1070 ± 0.0006 (< 1e-04)	0.0688 ± 0.0004 (< 1e-04)	0.1091 ± 0.0003 (< 1e-04)
β_1 ± SEM (p-value)	-0.0009 ± 0.0026 (0.7073)	0.0005 ± 0.0022 (0.8454)	0.0005 ± 0.0021 (0.8196)	-0.0015 ± 0.0007 (0.3906)	0.0005 ± 0.0021 (0.7002)	-0.0050 ± 0.0014 (0.0363)	-0.0010 ± 0.0012 (0.8842)
r ± SEM	-0.0106 ± 0.0287	0.0055 ± 0.0257	0.0065 ± 0.0253	-0.0243 ± 0.0119	0.0109 ± 0.0379	-0.0592 ± 0.0159	-0.0017 ± 0.0021

The first two rows give the regression coefficients for the linear model $F_{ST} = \beta_0 + \beta_1\rho + \varepsilon$, where ρ represents the mean recombination rate for the bin being tested and F_{ST} is the estimated F_{ST} for all population comparisons within a particular pair of continental groups (given in the column header). The final column, "Global", gives the regression coefficients for the linear model applied to all pairwise population comparisons (150 total). When performing the regression, ρ was first scaled to between 0 and 1, such that 1 represents the maximum observed recombination rate (126.88 cM/Mb) and 0 represents the minimum observed recombination rate (0.0 cM/Mb). The correlation coefficient, r, between ρ and F_{ST} for each comparison is shown in the bottom row. Standard errors of the mean (SEM) for β_0, β_1, and r were calculated from 1,000 bootstrap iterations (see Materials and Methods). P-values are derived from a two-sided t-test of the t-value for the corresponding regression coefficient.

Since the correlation between F_{ST} and recombination rate was previously documented as being strongest in coding regions [57], where the impact of BGS and other modes of linked selection is also expected to be strongest, we sought to disentangle the roles of BGS and recombination rates as explanatory variables for F_{ST} by using multiple linear regression. This served as a test for any additional effect recombination rate may have on F_{ST} that B failed to capture. To do so, we first split the genome into 2% recombination rate quantile bins and further subdivided each of these bins into 4% B quantile bins (50×25 = 1,250 bins total). We then measured F_{ST} within each bin. We also partitioned sites in the reverse order (2% B bins followed by 4% recombination rate bins)

306 and repeated all analyses. Our choice in total number of bins resulted in a minimum of
307 320 SNPs per bin for estimating F_{ST} between any two populations, which should be suf-
308 ficient to avoid errors when estimating F_{ST} across multiple loci [58]. After performing
309 multiple linear regression of F_{ST} on B, recombination rate (ρ), and an interaction term
310 between the two ($B\rho$) with the linear model $F_{ST} = \beta_0 + \beta_1 B + \beta_2 \rho + \beta_3 B\rho + \varepsilon$, we ob-
311 served that B was a statistically significant predictor ($p < 1e\text{-}04$) for F_{ST} across all popu-
312 lation comparisons regardless of how we partitioned sites (Table S8 in Supporting in-
313 formation). This result remained unchanged when performing robust regression. In con-
314 trast, recombination rate exhibited sporadic significance as an explanatory variable for
315 F_{ST} across population comparisons and was dependent upon how sites were partitioned
316 (i.e., whether we first partitioned by B or by recombination rate) (Table S8 in Supporting
317 information). Furthermore, strong differences between the two binning schemes were
318 observed for the magnitude of the recombination rate regression coefficient for certain
319 population comparisons (e.g., African vs. East Asian and South Asian vs. East Asian),
320 while the coefficients for B were consistent across binning schemes. The direction in
321 which recombination rate explained F_{ST} was also inconsistent across different popula-
322 tion comparisons, with European vs. South Asian and European vs. East Asian compar-
323 isons showing a significant positive change in F_{ST} as a function of increasing recombi-
324 nation rate. This result was contrary to an expectation of decreasing F_{ST} as a function of
325 increasing recombination rate [57]. We also failed to observe consistent effects from the
326 interaction term for B and recombination rate on F_{ST} across population comparisons or
327 binning schemes (Table S8 in Supporting information). Performing robust regression on
328 the model did not change these results. However, in contrast to recombination rate,

329 when the model was performed utilizing all TGP populations (i.e., the "Global" estimate),
330 the interaction term was significant in explaining F_{ST} across both types of binning
331 schemes.

332 To aid in visualizing the results of our multidimensional linear model, we plotted
333 F_{ST} for each population comparison as a function of recombination rate (across 4%
334 quantile bins) while conditioning on *B* (Fig 5A). We also plotted points in the reciprocal
335 direction, with F_{ST} being plotted as a function of *B* while conditioning on recombination
336 rate (Fig 5B). These data points were derived from the same points used as input for
337 the multiple linear regression model described above. The results for F_{ST} between Afri-
338 can and South Asian populations showed that *B* separated different levels of F_{ST} across
339 most recombination rate bins (Fig 5A, Table S9 in Supporting information). Furthermore,
340 regardless of how *B* was conditioned on recombination rate, it still exhibited a strong
341 trend of increasing F_{ST} as the strength of BGS increased (Fig 5B, Table S10 in Support-
342 ing information). These patterns were imperfect though, and statistical significance was
343 not always attained, especially for comparisons between non-African populations (Fig
344 S6, Table S10 in Supporting information). However, greater separation in F_{ST} was gen-
345 erally achieved when conditioning recombination rate on *B* and the slope was always
346 negative when plotting F_{ST} across *B*, regardless of which recombination rate percentile
347 bin *B* was conditioned on. These results suggested that *B* is a dominant predictor for
348 F_{ST}, while recombination rate plays a minor role.

349

350 **Demographic inference in putatively neutral regions of the genome**

351 One consequence of BGS driving patterns of neutral variation within and be-

352 tween human populations is that demographic inference could be substantially biased
353 [59,60]. To assess the degree of bias, we inferred a 13-parameter demographic model
354 of African, European, and East Asian demography using only putatively neutral regions
355 of the genome under the weakest effects of BGS ($B \geq 0.994$) from a subset of TGP indi-
356 viduals with high coverage whole genome sequence data (see Materials and Methods).
357 Our demographic model followed that of Gutenkunst et al. [7], with an ancient human
358 expansion in Africa and a single out-of-Africa bottleneck followed by European- and
359 East Asian-specific bottlenecks, as well as exponential growth in both non-African popu-
360 lations and migration between all populations. To make comparisons to previous studies
361 that have used sequence data from coding regions or genes [7,61,62], which may be
362 under strong BGS, we also inferred demographic parameters using coding four-fold de-
363 generate synonymous sites. Our inferred parameters for human demography were strik-
364 ingly different between the two sets of sequence data (Fig S1, Table S1 in Supporting
365 information). Notably, inferred effective population size parameters were larger for con-
366 temporary population sizes when using four-fold degenerate synonymous sites versus
367 weak BGS neutral regions, with N_e inferred to be 22%, 23%, and 29% larger for AFR,
368 EUR, and EASN populations, respectively. This is despite the fact that the ancestral N_e
369 was inferred to be lower for four-fold degenerate synonymous sites (N_e = 18,449 and
370 17,118, for weak BGS neutral regions and four-fold degenerate sites, respectively). This
371 result may stem from the expected decrease in N_e going into the past in regions of
372 BGS, which can lead to inflated estimates of recent population growth [60] and has also
373 been shown in simulation studies of synonymous sites under BGS [59].
374

375 **Simulations confirm that demographic effects can impact background selection**

376 Using the demographic parameters inferred from weak BGS neutral regions, we
377 simulated the joint effects of demography and BGS on neutral diversity for African, Eu-
378 ropean, and East Asian populations (see Materials and Methods). We also ran an iden-
379 tical set of simulations without BGS in order to generate an expectation of the effect of
380 demography only on neutral diversity. To measure the relative impact of BGS for each
381 population, we then took the ratio of neutral diversity from BGS simulations (π) and neu-
382 tral diversity from simulations without BGS (π_0) to calculate relative diversity (π/π_0). As
383 expected, we found that BGS reduced relative diversity ($\pi/\pi_0 < 1$) for all three popula-
384 tions in our simulations. However, non-African populations experienced a proportionally
385 larger decrease in π/π_0 compared to the African population (π/π_0 = 0.43, 0.42, 0.41 in
386 AFR, EUR, and EASN respectively). To understand how this dynamic process occurs,
387 we sampled all simulated populations every 100 generations through time to observe
388 the effect of population size change on π/π_0 (Fig 6). We observed that there is a distinct
389 drop in π/π_0 at each population bottleneck experienced by non-Africans, with East
390 Asians suffering the largest drop overall, followed by Europeans. Interestingly, Africans
391 also experienced a large drop in π/π_0 (but less than non-Africans) even though they did
392 not experience any bottlenecks. This was attributable to migration between non-Africans
393 and Africans and this pattern disappeared when we ran simulations using an identical
394 demographic model with BGS but without migration between populations (Fig S7). This
395 finding highlights an evolutionary role that deleterious alleles can play when they are
396 transferred across populations through migration (see Discussion).

397 Our simulations were based on the functional density found in the region of the

398 human genome with the strongest effect of BGS (chr3: 48,600,000-50,600,000), where
399 20.46% of sites were either coding or conserved non-coding (see Materials and Meth-
400 ods). Thus, the fraction of the genome experiencing deleterious mutation in our simula-
401 tions of strong BGS (which we define as *U*) was 0.2046. The patterns we observed in
402 these simulations likely represent an upper bound on the strength of BGS in the human
403 genome. We therefore lowered the impact of BGS by reducing *U* (see Materials and
404 Methods). When *U* is decreased 2-4 fold, we continued to observe a stepwise decrease
405 in π/π_0 similar to our simulations of *U* = 0.2046, with the specific rank order of African,
406 followed by European, and then East Asian populations (Fig S8). As expected, π/π_0 in-
407 creased overall for all populations as the fraction of sites that were simulated as delete-
408 rious was decreased (π/π_0 = 0.641 vs. 0.802, 0.62 vs. 0.777, and 0.611 vs. 0.777 for
409 AFR, EUR, and EASN when *U* = 0.1023 and *U* = 0.05115, respectively).
410

411 ## Discussion

412 In our analyses of thousands of genomes from globally distributed human popu-
413 lations, we have confirmed that neutral variation in humans is governed largely by the
414 processes of both linked selection and demography. While this observation is not unex-
415 pected, we have characterized for the first time the dynamic consequences of non-
416 equilibrium demographic processes in regions of BGS. We find that demography's ef-
417 fect in these regions do not result in the same relative magnitude of decreased neutral
418 diversity across populations. Instead, relative diversity in regions of BGS is highly de-
419 pendent on a population's demographic history, with bottlenecks playing a particularly
420 strong role. For example, the strongest overall decreases in relative diversity occur in

421 the East Asian continental group, which is expected if they also suffered the strongest
422 population bottlenecks. To remove any possible biases that would influence our results,
423 we controlled for functional effects of mutations, variability in mutation along the ge-
424 nome, potential sequencing artifacts, GC-based gene conversion, and the potential mu-
425 tagenic effects of recombination hotspots. None of these factors qualitatively affected
426 our results.

427 We do recognize that one caveat of our controls is the fact that divergence itself
428 is not independent of BGS [63], and this may present biases when using divergence to
429 control for variation in mutation rate along the genome. This is because the rate of coa-
430 lescence in the ancestral population of two groups will be faster in regions of strong
431 BGS compared to regions of weak BGS due to the lower N_e of the former, leading to a
432 decrease in overall divergence in those regions. To limit the contribution of such biases
433 in ancestral N_e to divergence, we use rhesus macaque since it is more distantly related
434 to humans than other primate species such as orangutan or chimpanzee (human-
435 rhesus divergence: 29.6 MYA; [64]). However, biases in divergence due to BGS can be
436 still be observed between species as distantly related as human and mouse [63] (hu-
437 man-mouse divergence: 75 MYA; [65]). Despite this, the apparent correlation between
438 BGS and divergence should not qualitatively affect our results of relative diversity. Ra-
439 ther, it is likely that differences in normalized diversity between different BGS bins are
440 greater than what we observe here. Normalizing by the lower divergence that is charac-
441 teristic of strong BGS bins and the higher divergence that is characteristic of weak BGS
442 bins should make any differences between the two smaller, not greater. In fact, for our
443 calculations of relative diversity in which we skip the normalization step, the differences

444 in diversity between the strongest 1% and weakest 1% BGS bins are greater and give a
445 lower ratio of relative diversity (π/π_0 for AFR is 0.373 without the divergence step and
446 0.402 with the divergence step). A similar pattern is also observed for other continental
447 groups (compare Fig 2 and Fig S5). More importantly though, we should not expect the
448 potential biases of our divergence step to contribute to the differences in relative diversi-
449 ty between each of the continental groups since biases in divergence across the ge-
450 nome should impact all human populations equally.

451 We also note that the estimates of *B* by McVicker et al. [6] may be biased by
452 model assumptions concerning mutation rates and the specific sites subject to purifying
453 selection, with the exact values of *B* also unlikely to be precisely inferred. However, a
454 similar study by Comeron et al. [29] that investigated BGS in *Drosophila* and utilized the
455 same model of BGS as McVicker et al. found that biases presented by model assump-
456 tions or mis-inference on the exact value of *B* do not significantly change the overall
457 rank order for the inferred strength of BGS across the genome. Thus we, expect
458 McVicker et al.'s inference of *B* to provide good separation between the weakest and
459 strongest regions of BGS in the human genome, with model misspecification unlikely to
460 change our empirical results.

461 The increased population differentiation we observe in regions of strong BGS are
462 consistent with the increased action of genetic drift occurring in genomic regions with
463 characteristically low N_e, which has been predicted theoretically [66,67] and observed in
464 previous studies [57,68,69]. Here, we replicate this pattern by measuring F_{ST} as a func-
465 tion of *B* across a global set of populations, providing a rich description of global human
466 population differentiation across the genome. Furthermore, unlike previous studies of

467 F_{ST} in humans that have been plagued by biases due to SNP ascertainment or allele
468 frequency [56], we make use of SNP data from publically available KhoeSan genomes
469 to select a set of truly informative SNPs for estimating F_{ST}. Our results show that popu-
470 lation differentiation can be dramatically influenced by BGS, with estimates of F_{ST} at
471 neutral sites in the weakest 2% BGS bin and strongest 2% BGS bin in the genome dif-
472 fering by 23.6%-70.1% when we compare all pairs of continental groups (Fig 4A). When
473 accounting for differences using recombination rate alone, pairs of continental groups
474 showed little change at all (Fig 4B), thus demonstrating the added value of using explicit
475 models of linked selection when resolving properties of neutral variation across the ge-
476 nome.

477 While we focus specifically on the effects of BGS on generating local patterns of
478 diversity across humans, there is still a possibility that the effects of positive selection,
479 through recurrent-hitchhiking, soft-sweeps, and classic selective sweeps, are contrib-
480 uting to these patterns as well. Models taking into account both the effects of purifying
481 selection and positive selection will be necessary to provide a full description of the im-
482 pact of linked selection across the human genome and its behavior as a function of dif-
483 ferent population demographic histories [42]. Development of mathematical models
484 characterizing the joint effects of both BGS and genetic hitchhiking [24] as well as their
485 recent application to different species [30,44], gives promise towards this goal. Recent-
486 ly, a joint model of classic selective sweeps and BGS was applied to *Drosophila* and
487 predicted that BGS has had a ~1.6 to 2.5-fold greater effect on neutral genetic diversity
488 than classic selective sweeps [30]. We should expect this magnitude to be even greater
489 for humans, since classic selective sweeps were found to be rare in recent human evo-

490 lution [38] and adaptive substitutions in the human genome are much less frequent
491 than *Drosophila* [5,70,71]. Despite this, we still attempt to alleviate any contribution from
492 other modes of linked selection by masking out regions of hard-sweeps or soft-sweeps
493 using a novel H12-like method [72] that uses an integrated haplotype homozygosity
494 framework [36] to scan for population specific signatures of these events in the genome
495 ([73]; S2 Appendix). While there may still be some residual contribution of these events
496 to our estimates of π/π_0, if such signatures do exist, they are likely to localize where
497 BGS is also expected to act. Thus, it is possible that our empirical results describe the
498 consequences of non-equilibrium demography in the broader context of both positive
499 and purifying modes of linked selection. Indeed, our simulations of BGS fail to capture
500 the complete effects of linked selection on reducing π/π_0 in different human populations
501 (compare Figs 2B and 6C) and the additional contribution of hitchhiking, which we did
502 not simulate, may explain this discrepancy.

503 Non-equilibrium demography has also been been recently investigated with
504 regard to its impact on patterns of deleterious variation across human populations. Initial
505 work comparing non-synonymous and synonymous sites in European-Americans and
506 African-Americans found that, while African-Americans have a greater number of
507 segregating non-synoymous and synonymous sites than European-Americans,
508 European-Americans have an overall greater amount of non-synonymous and
509 synonymous homozygous derived alleles [50]. Other work showed similar findings
510 [62,74], with Henn et al. [75] demonstrating that the number of deleterious derived
511 alleles per genome increased with distance from Africa. In addition, similar results in
512 human founder populations [76,77], *Arabadopsis* [78], and domesticated species such

513 as dogs [12] and sunflowers [79] further demonstrate the pervasive impact that
514 demography has on influencing the relative amount of deleterious variation across a
515 variety of populations and species. Since BGS itself is a function of deleterious
516 variation, it is not surprising that we also witness differences in π/π_0 across human
517 populations that have experienced different demographic histories. These effects are
518 likley ubiquitous across other species as well. However, there has been recent
519 contention about whether the previously described patterns of increased deleterious
520 variation are driven by a decrease in the efficacy of natural selection or are solely
521 artifacts of the response of deleterious variation to demographic change [51,52,80–82].
522 Following a strong population bottleneck, a transient increase in the relative ratio of
523 deleterious variation to neutral variation is expected, that is independent of the efficacy
524 of selection [47,48]. Since our own investigation focuses solely on neutral variants, we
525 can only invoke differences in population demography and drift on causing the
526 differences between populations, supporting the conjecture that demographic processes
527 may transiently dominate patterns of diversity at sites that may be under the direct
528 influence (or in the case of BGS, indirect influence) of purifying selection.

529 Recently, Koch et al. [49] investigated the temporal dynamics of human
530 demography on selected sites and observed that after a population contraction,
531 heterozygosity at selected sites can undershoot its expected value at equilibrium as
532 low-frequency variants are lost at a quicker rate before the recovery of intermediate
533 frequency variants can occur. In the context of BGS, which skews the site frequency
534 spectrum of linked neutral mutations towards rare variants [23,83], we also expect a
535 transient decrease in diversity as low-frequency variants are lost quickly during a

536 population contraction. This effect may be driving the results of the greater losses of
537 relative diversity in non-African populations, but further work is needed to better
538 understand the impact of population contractions on perturbing the site-frequency
539 spectrum of regions under BGS. Koch et al. demonstrate that this effect is only
540 temporary and that long-term diversity at selected sites approaches greater values once
541 equilibrium is reached. We stress that these temporal effects also impact the patterns of
542 neutral diversity caused by BGS and likely explain the differences of π/π_0 that we
543 observed in humans. We predict that as evolution continues forward through time, the
544 magnitude and rank order of relative diversity among populations may change.
545 However, this will be highly dependent on the demographic effects that different
546 populations experience. Although more flexible modeling frameworks that can account
547 for BGS under changing population size have been recently developed [45,46], they as-
548 sume that selection is strong enough that the impact of demography does not perturb
549 mutation-selection balance, and they also ignore the additional impact that demography
550 itself has on perturbing neutral variants.

551 A greater contemporary N_e in non-Africans could theoretically result in a greater
552 efficacy of purifying selection and, consequently, a stronger impact of BGS. However, it
553 is very unlikely that this is occuring or driving the observed patterns of relative diversity
554 that we see. The greater contemporary population size of non-Africans has transpired
555 only in the very recent past, with accelerated growth in Europeans occurring within the
556 last few hundred generations [62,84–87]. Thus, most population specific mutations in
557 non-Africans have arisen only during this recent population expansion and are both
558 young and very rare [88,89]. It is doubtful that these population-specific variants have

559 had enough time to exercise a large effect on the absolute strength of BGS itself.
560 Rather, our simulations indicate that the response of π in regions under BGS is driven
561 by population contractions, with reductions in π/π_0 occurring concomitantly with the out-
562 of-Africa bottleneck and European-East Asian split bottleneck events (Fig 6) and contin-
563 uing even after the European and East Asian expansion events. Interestingly though,
564 our simulations reveal an additional factor that can influence the impact of BGS within
565 populations – migration between populations. We observe that the exchange of delete-
566 rious variants between populations can impact linked selection and lead to decreases in
567 π/π_0 in populations of constant size, such as the recent history inferred for Africans (Fig
568 6B). In particular, a decrease in π/π_0 for Africans coincides with decreases in π/π_0 for
569 non-Africans. This trend disappears, though, once migration is removed (Fig S7). This
570 result suggests that the migration of nearly neutral or mildly deleterious alleles in Euro-
571 peans/East Asians into Africans results in those variants being more strongly selected in
572 Africans due to their historically larger N_e, resulting in stronger BGS and lower π. How-
573 ever, more work is needed to definitively test this. Conversely, migration of African al-
574 leles into Europeans/East Asians has no noticeable effect on π in our simulations with
575 BGS (compare solid lines in Fig 6B and Fig S7B). In simulations where U is lower than
576 our base level of 0.2046, we also witness differences in π/π_0 between simulations with
577 and without migration. Some of these involve rank order changes in π/π_0 between Afri-
578 cans, Europeans, and East Asians (Fig S8, U = 0.02046 with migration vs. no migra-
579 tion), which is likely a consequence of the increased stochasticity that migration be-
580 tween populations introduces, as simulations without migration restored the consistent
581 patterns of π/π_0. These results demonstrate the complexities that migration introduces to

582 patterns of diversity in regions of linked selection and to the dynamic changes of π/π_0
583 over time.

584 While we describe here the differential effects of non-equilibrium demography on
585 neutral diversity in regions under strong and weak BGS, it is worth mentioning that dif-
586 ferences in the reduction of neutral diversity in the genome between different popula-
587 tions have also been investigated at the level of entire chromosomes. In particular,
588 analyses of neutral diversity comparing autosomes to non-autosomes (i.e., sex chromo-
589 somes and the mitochondrial genome [mtDNA]) have been conducted. Interestingly,
590 these studies have shown that population contractions have impacted the relative re-
591 duction of neutral diversity between non-autosomes and autosomes in a similar fashion
592 to what we have observed between regions of strong BGS and weak BGS, with the
593 greatest losses occurring in bottlenecked populations. This was demonstrated in both
594 humans and *Drosophila* [90,91] and later modeled [92], showing that stronger genetic
595 drift due to the lower N_e of non-autosomes causes diversity to be lost more quickly in
596 response to population size reductions. Recent work in humans has confirmed such
597 predictions by showing that relative losses of neutral diversity in the non-autosomes are
598 greatest for non-Africans [93–95]. These studies, plus others [96], have also shown that
599 there is strong evidence for a more dominant effect of linked selection on the sex chro-
600 mosomes relative to the autosomes in humans.

601 Since linked selection is a pervasive force in shaping patterns of diversity across
602 the genomes in a range of biological species [1], it has been provided as an argument
603 for why neutral diversity and estimates of N_e are relatively constrained across species in
604 spite of the large variance in census population sizes that exist [44,97]. However, since

605 population bottlenecks are common among species and have an inordinate influence on
606 N_e [20], demography has also been argued as a major culprit for constrained diversity
607 [2,97–99]. Yet, as we show in humans, it is likely that patterns of neutral diversity are in
608 fact jointly impacted by both of these forces, magnifying one another to deplete levels of
609 diversity beyond what is expected by either one independently. In our work, we also
610 identify a potentially substantial role for migration from smaller populations that harbor
611 more strongly deleterious alleles on patterns of linked neutral diversity in large popula-
612 tions. Together, these combined effects may help provide additional clues for the puz-
613 zling lack of disparity in genetic diversity among different species [100].

614 Our results give caution that not properly accounting for the impacts of demogra-
615 phy on patterns driven by linked selection may lead to spurious rejections of certain
616 models of linked selection in favor of others. For example, when comparing humans to
617 other primates, a recent study by Nam et al. [101] found a greater relative decrease of
618 neutral genetic diversity in regions close to genes (where linked selection is greatest)
619 versus regions far away from genes in non-human primate species. The authors sug-
620 gest that these patterns are best described by the action of selective sweeps since non-
621 human primate populations usually exhibit a larger N_e and thus should experience a
622 greater influx of adaptive alleles per unit time. However, certain orangutan, gorilla, and
623 chimpanzee species have also experienced large population contractions in their recent
624 history that are comparable in time and potentially more extreme in scale to what hu-
625 mans have also experienced [102–105]. Thus, the consequences of strong demograph-
626 ic change to regions of BGS can provide an alternative explanation for why these pri-
627 mate species exhibit a larger relative reduction in neutral diversity near genes, despite

628 their larger historical N_e. While Nam et al. also test for BGS using simulations with de-
629 mography, their model assumed a more limited population contraction (50% population
630 reduction) over a shorter time scale than the demographic model we test.

631 Finally, our results also have implications for human medical genetics research,
632 since selection may be acting on functional regions contributing to disease susceptibil-
633 ity. Since different populations will have experienced different demographic histories,
634 the action of linked selection may result in disparate patterns of genetic variation (with
635 elevated levels of drift) near causal loci. Recent work has already demonstrated that
636 BGS's consequence of lowering diversity impacts power for disease association tests
637 [106]. Our results indicate that this impact may be even further exacerbated by demog-
638 raphy in bottlenecked populations, leading to potentially larger discrepancies in power
639 between different populations. Overall, this should encourage further scrutiny for tests
640 and SNP panels optimized for one population since they may not be easily translatable
641 to other populations. It should also further motivate investigators to simultaneously ac-
642 count for demography and linked selection when performing tests to uncover disease
643 variants within the genome [88,106,107].

644

645 ## Materials and methods

646 ### Data

647 2,504 samples from 26 populations in phase 3 of the Thousand Genomes Project
648 (TGP) [9] were downloaded from ftp://ftp.ncbi.nlm.nih.gov/1000genomes/. vcftools
649 (v0.1.12a) [108] and custom python scripts were used to gather all bi-allelic SNP sites
650 from the autosomes of the entire sample set.

651 A subset of TGP samples that were sequenced to high coverage (~45X) by
652 Complete Genomics (CG) were downloaded from
653 ftp://ftp.ncbi.nlm.nih.gov/1000genomes/. After filtering out related individuals via pedi-
654 gree analyses, we analyzed 53 YRI, 64 CEU, and 62 CHS samples (Table S2). The
655 cgatools (v1.8.0) listvariants program was first used to gather all SNPs from the 179
656 samples using their CG ASM "Variations Files" (CG format version 2.2). Within each
657 population, the number of reference and alternate allele counts for each SNP was then
658 calculated using the cgatools testvariants program and custom python scripts. Only al-
659 lele counts across high quality sites (i.e., those classified as VQHIGH variant quality by
660 CG) were included. Low quality sites (i.e., those with VQLOW variant quality) were
661 treated as missing data. Only autosomes were kept. Non-bi-allelic SNPs and sites vio-
662 lating Hardy-Weinberg equilibrium (HWE) (p-value < 0.05 with a Bonferroni correction
663 for multiple SNP testing) were also removed.

664 We collected 13 whole-genome sequenced KhoeSan samples (sequence-
665 coverage: 2.5-50X, see Table S3 in Supporting information) from 3 studies [75,109,110]
666 and used the processed vcf files from each of those respective studies to gather all bi-
667 allelic polymorphic SNPs (i.e., the union of variants across all vcf files). SNPs were only
668 retained if they were polymorphic within the 13 samples (i.e., sites called as alternate
669 only within the sample set were ignored).

670

671 **Filtering and ascertainment scheme**

672 Positions in the genome were annotated for background selection by using the
673 background selection coefficient, *B*, which was inferred by McVicker et al. [6] and down-

674 loaded from http://www.phrap.org/othersoftware.html. *B* was inferred by applying a clas-
675 sical model of BGS [53], which treats its effects as a simple reduction in N_e at neutral
676 sites as a function of their recombination distance from conserved loci, the strength of
677 purifying selection at those conserved loci, and the deleterious mutation rate. *B* can be
678 interpreted as the reduced fraction of neutral genetic diversity at a particular site along
679 the genome that is caused by BGS, with a value of 0 indicating a near complete remov-
680 al of neutral genetic diversity due to BGS and a *B* value of 1 indicating little to no effect
681 of BGS on neutral genetic diversity ($B = \pi/\pi_0 = N_e/N_0$). Positions for *B* were lifted over
682 from hg18 to hg19 using the UCSC liftOver tool. Sites that failed to uniquely map from
683 hg18 to hg19 or failed to uniquely map in the reciprocal direction were excluded. Sites
684 lacking a *B* value were also ignored. We focused our analyses on those regions of the
685 genome within the top 1%, 5%, 10%, and 25% of the genome-wide distribution of *B* and
686 within the bottom 1% of the genome-wide distribution of *B*. These quantiles correspond
687 to the *B* values 0.095, 0.317, 0.463, 0.691, and 0.994, respectively.

688 A set of 13 filters (referred to as the "13-filter set") were used to limit errors from
689 sequencing and misalignments with rhesus macaque and to remove regions potentially
690 under the direct effects of natural selection and putative selective sweeps (we ignore
691 the linked selection effects of background selection). These filters were applied to all
692 samples in phase 3 TGP (all filters are in build hg19) for all sets of analyses (see Table
693 S4 in Supporting information for the total number of Mb that passed the described filters
694 below for each particular *B* quantile):

695 1. Coverage/exome: For phase 3 data, regions of the genome that were part of
696 the high coverage exome were excluded (see

697 ftp://ftp.ncbi.nlm.nih.gov/1000genomes/ftp/technical/reference/exome_pull_do
698 wn_targets/20130108.exome.targets.bed.README). This was done to limit
699 biases due to differing levels of coverage across the genome and to remove
700 likely functional sites within the exome.

701 2. phyloP: Sites with phyloP [111] scores > 1.2 or < -1.2 were removed to limit
702 the effects of natural selection due to conservation or accelerated evolution.
703 Scores were downloaded from
704 http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phyloP46way/.

705 3. phastCons: Regions in the UCSC conservation 46-way track (table:
706 phastCons46wayPlacental) [112] were removed to limit the effects of natural
707 selection due to conservation.

708 4. CpG: CpG islands in the UCSC CpG islands track were removed because of
709 their potential role in gene regulation and/or being conserved.

710 5. ENCODE blacklist: Regions with high signal artifacts from next-generation
711 sequencing experiments discovered during the ENCODE project [113] were
712 removed.

713 6. Accessible genome mask: Regions not accessible to next-generation se-
714 quencing using short reads, according to the phase 3 TGP “strict” criteria,
715 were removed (downloaded from
716 ftp://ftp.ncbi.nlm.nih.gov/1000genomes/ftp/release/20130502/supporting/acce
717 ssible_genome_masks/StrictMask/).

718 7. Simple repeats: Regions in the UCSC simple repeats track were removed
719 due to potential misalignments with outgroups and/or being under natural se-

720 lection.

721 8. Gaps/centromeres/telomeres: Regions in the UCSC gap track were removed,
722 including centromeres and telomeres.

723 9. Segmental duplications: Regions in the UCSC segmental dups track [114]
724 were removed to limit potential effects of natural selection and/or misalign-
725 ments with rhesus macaque.

726 10. Transposons: Active transposons (HERVK retrotransposons, the AluY sub-
727 family of Alu elements, SVA elements, and L1Ta/L1pre-Ta LINEs) in the hu-
728 man genome were removed.

729 11. Recent positive selection: Regions inferred to be under hard and soft selec-
730 tive sweeps (using iHS and iHH12 regions from selscan v1.2.0 [73]; S2 Ap-
731 pendix) within each phase 3 population were removed.

732 12. Non-coding transcripts: Non-coding transcripts from the UCSC genes track
733 were removed to limit potential effects of natural selection.

734 13. Synteny: Regions that did not share conserved synteny with rhesus macaque
735 (rheMac2) from UCSC syntenic net filtering were removed (downloaded from
736 http://hgdownload.soe.ucsc.edu/goldenPath/hg19/vsRheMac2/syntenicNet/).

737 Additionally, an extra set of filters was applied, but only for those estimates of diversity
738 that controlled for GC-biased gene conversion and recombination hotspots:

739 14. GC-biased gene conversion (gBGC): Regions in UCSC phastBias track [115]
740 from UCSC genome browser were removed to limit regions inferred to be un-
741 der strong GC-biased gene conversion.

742 15. Recombination hotspots: All sites within 1.5 kb (i.e., 3 kb windows) of sites

743 with recombination rates ≥ 10 cM/Mb in the 1000G OMNI genetic maps for
744 non-admixed populations (downloaded from ftp://ftp-
745 trace.ncbi.nih.gov/1000genomes/ftp/technical/working/20130507_omni_reco
746 mbination_rates/) and the HapMap II genetic map [116] were removed. 1.5 kb
747 flanking regions surrounding the center of hotspots identified by Ref. [117]
748 (downloaded from
749 http://science.sciencemag.org/content/sci/suppl/2014/11/12/346.6211.125644
750 2.DC1/1256442_DatafileS1.txt) were also removed, except for the cases in
751 which the entire hotspot site was greater than 3 kb in length (in which case
752 just the hotspot was removed).

753 To generate a set of four-fold degenerate synonymous sites, all polymorphic
754 sites that we retained from the high-coverage Complete Genomic samples were anno-
755 tated using the program ANNOVAR [118] with Gencode V19 annotations. ANNOVAR
756 and Gencode V19 annotations were also used to gather an autosome-wide set of four-
757 fold degenerate sites, resulting in 5,188,972 total sites.

758

759 **Demographic inference**

760 The inference tool dadi (v1.6.3) [7] was used to fit, via maximum likelihood, the 3-
761 population 13-parameter demographic model of Gutenkunst et al. [7] to the 179 YRI,
762 CEU, and CHS samples from the high coverage CG dataset of TGP. This sample set
763 consisted of 53 YRI (African), 64 CEU (European), and 62 CHS (East Asian) samples.
764 The demographic model incorporates an ancient human expansion in Africa and a sin-
765 gle out-of-Africa bottleneck followed by European- and East Asian-specific bottlenecks,

766 as well as exponential growth in both non-African populations and migration between
767 populations. During the inference procedure, each population was projected down to
768 106 chromosomes, corresponding to the maximum number of chromosomes available
769 in the CG YRI population. Sites were polarized with chimpanzee to identify putative an-
770 cestral/derived alleles using the chain and netted alignments of hg19 with panTro4
771 (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/vsPanTro4/axtNet/), and the correc-
772 tion for ancestral misidentification [119] option in dadi was used. The 13-filter set de-
773 scribed previously was applied to the CG data set, and an additional filter keeping only
774 the weakest ~1% of autosomal sites under background selection ($B \geq 0.994$) was also
775 applied in order to mitigate potential biases in inference due to BGS [59,60] or other
776 forms of linked selection [120]. After site filtering and correction for ancestral misidentifi-
777 cation, a total of 110,582 segregating sites were utilized by dadi for the inference proce-
778 dure. For optimization, grid points of 120, 130, and 140 were used, and 15 independent
779 optimization runs were conducted from different initial parameter points to ensure con-
780 vergence upon a global optimum. An effective sequence length (L) of 7.15 Mb was cal-
781 culated from the input sequence data after accounting for the fraction of total sites re-
782 moved due to filtering. In addition to the 13-filter set, this filtering included sites violating
783 HWE, sites without B value information, sites that did not have at least 106 sampled
784 chromosomes in each population, sites with more than two alleles, sites that did not
785 have tri-nucleotide information for the correction for ancestral misidentification step, and
786 sites treated as missing data. For calculating the reference effective population size, a
787 mutation rate (μ) of 1.66 x 10^{-8} (inferred from Ref. [121]) was used. Using the optimized
788 θ from dadi after parameter fitting, the equation $\theta = 4N_e\mu L$ was solved for N_e to generate

789 the reference effective population size, from which all other population N_e's were calcu-
790 lated. This same procedure was also used to infer demographic parameters from four-
791 fold degenerate synonymous sites across the same set of samples. After site filtering
792 (note that *B* and the 13-filter set were not included in the filtering step for four-fold de-
793 generate synonymous sites), 41,260 segregating sites were utilized by dadi for the in-
794 ference procedure, and an effective sequence length of 2.37 Mb was used for calculat-
795 ing the reference effective population size.
796
797 **Simulations**
798 Forward simulations incorporating the results from the demographic inference
799 procedure described above and a model of background selection were conducted using
800 SFS_CODE [122]. For the model of background selection, the recombination rate, ρ,
801 and the fraction of the genome experiencing deleterious mutation, *U*, was calculated us-
802 ing the 2 Mb region of chr3: 48,600,000-50,600,000, which has been subject to the
803 strongest amount of BGS in the human genome (mean B = 0.002). A population-scaled
804 recombination rate (ρ) of 6.0443 x 10^{-5} was calculated for this region using the HapMap
805 II GRCh37 genetic map [116]. For generating *U*, the number of non-coding "functional"
806 sites in this region was first calculated by taking the union of all phastCons sites and
807 phyloP sites with scores > 1.2 (indicating conservation) that did not intersect with any
808 coding exons. This amount totaled to 270,348 base pairs. Additionally, the number of
809 coding sites was calculated by summing all coding exons within this region from GEN-
810 CODE v19, which totaled to 138,923 base pairs. From these totals, a *U* of 0.2046 was
811 generated.

812 The background selection model was simulated using a middle 30 kb neutral re-
813 gion flanked by two 1 Mb regions under purifying selection. From the calculated *U* de-
814 scribed above, 20.46% of sites in the two 1 Mb flanking regions were simulated as be-
815 ing deleterious. Two distributions of fitness effects were used for the deleterious sites,
816 with 66.06% of deleterious sites using the gamma distribution of fitness effects inferred
817 across conserved non-coding regions by Ref. [123] (β = 0.0415, α = 0.00515625) and
818 33.94% of deleterious sites using the gamma distribution of fitness effects inferred
819 across coding regions by Ref. [5] (β = 0.184, α = 0.00040244). The relative number of
820 sites receiving each distribution of fitness effects in our simulations was determined by
821 the relative number of non-coding "functional" sites and coding exons described above.
822 Gamma distribution parameters are scaled to the ancestral population size of the de-
823 mographic models used in Refs. [5,123]. To simulate varying levels of background se-
824 lection strength, different total fractions of our calculated *U* were used (5%, 10%, 25%,
825 50%, and 100% of 0.2046) but with the same relative percentage of non-coding and
826 coding sites just described. To simulate only the effects of demography without back-
827 ground selection (i.e., *U* = 0), only the 30 kb neutral region was simulated. 2,000 inde-
828 pendent simulations were conducted for each particular *U* (12,000 total). Simulations
829 output population genetic information every 100 generations and also at each genera-
830 tion experiencing a population size change (22,117 total generations were simulated),
831 for which mean pairwise nucleotide diversity (π) was calculated across the 2,000 simu-
832 lations.
833
834 **Calculating recombination rate across the genome**

835 We used recombination rate estimates from the HapMap II GRCh37 genetic
836 map. To annotate sites in phase 3 that were not in HapMap II, recombination rates were
837 interpolated to the midway point between the preceding and following positions in Hap-
838 Map II. If the difference between successive HapMap II positions was greater than
839 18,848 base pairs (the first standard deviation for the distribution of distances between
840 positions in HapMap II), then the recombination rate was only extended out 9,424 base
841 pairs beyond the focal position. Positions beyond this distance were then ignored during
842 analysis in which the recombination rate was used. Recombination rate quantiles were
843 calculated using the genome-wide distribution of recombination rates (i.e., the distribu-
844 tion of recombination rates across all sites, including those that are not polymorphic in
845 the data set) resulting from the procedure described above.
846
847
848 **Population-specific calculations of diversity**
849 Mean pairwise genetic diversity (π) was calculated as a function of the *B* quantile
850 bins described in "Filtering and ascertainment scheme" for each of the 20 non-admixed
851 populations in phase 3 TGP and across 4 broad populations that grouped the 20 non-
852 admixed populations together by continent (African, European, South Asian, and East
853 Asian, see Table S11 in Supporting information). Additionally, only regions of the ge-
854 nome passing the 13-filter set were used in the diversity calculations (see Table S4 in
855 Supporting information for total number of Mb used in diversity calculations for each *B*
856 quantile). For estimates of diversity controlling for gBGC or recombination hotspots, the
857 additional corresponding filters described in "Filtering and ascertainment scheme" were

858 also used. Only 100 kb regions of the genome with at least 10 kb of divergence infor-
859 mation with Rhesus macaque were used in the diversity calculations (see "Normaliza-
860 tion of diversity and divergence calculations with Rhesus macaque" below).
861
862 **Ancestry specific calculations of diversity**
863 To calculate genetic diversity as a function of local ancestry across the admixed
864 phase 3 population samples, we used the ancestry deconvolution results generated by
865 the 1000 Genomes Project Consortium (see
866 ftp://ftp.ncbi.nlm.nih.gov/1000genomes/ftp/technical/working/20140818_ancestry_decon
867 volution/README_20140721_phase3_ancestry_deconvolution). Briefly, the local-
868 ancestral inference tool, RFMix [124], was run across the ACB, ASW, CLM, MXL, PEL,
869 and PUR phase 3 TGP samples. For the reference panel, 50 unrelated shapeit2 [125]
870 trio-phased YRI and CEU samples each (from phase 3 TGP) and 43 shapeit2 popula-
871 tion-phased Native American samples (from Ref. [126]) were used. We utilized local an-
872 cestry tracks that were inferred by RFMix using "trio-phased" mode.
873 Admixed samples were then parsed for all genomic segments homozygous for
874 each particular ancestry (i.e., African, European, or Native American). These homozy-
875 gous segments were also filtered according to the 13-filter set described in "Filtering
876 and ascertainment scheme." Heterozygosity was calculated across admixed samples
877 for each set of homozygous ancestries and *B* quantile bins described previously. Sam-
878 ples were included in the analyses only if the total length of their genome that passed all
879 filters for the particular ancestry and *B* quantile bin was greater than 1 Mb. Additionally,
880 only 100 kb segments that had at least 10 kb of divergence information with Rhesus

881 macaque were used (see below). Per-site heterozygosity estimates for each ancestry
882 and *B* quantile set were averaged across all admixed samples, regardless of their TGP
883 population of origin. See Table S5 in Supporting information for total number of Mb used
884 in the analyses. Additionally, heterozygosity was also calculated across the 4 continen-
885 tal groups using the same 13-filter set and as a function of the same *B* quantile bins.
886

887 **Normalization of diversity and divergence calculations with Rhesus macaque**

888 To calculate human divergence with Rhesus macaque, we downloaded the
889 syntenic net alignments between hg19 and rheMac2 that were generated by blastz from
890 http://hgdownload.cse.ucsc.edu/goldenpath/hg19/vsRheMac2/syntenicNet/. We binned
891 the human genome into non-overlapping 100 kb bins and calculated divergence within
892 each bin by taking the proportion of base pair differences between human and Rhesus
893 macaque. Gaps between human and Rhesus macaque, positions lacking alignment in-
894 formation, and positions that did not pass the 13-filter set described in "Filtering and
895 ascertainment scheme" were ignored in the divergence estimate. Additionally, a sepa-
896 rate set of divergence estimates were also made using the additional set of filtering cri-
897 teria that removed those regions under gBGC or in recombination hotspots and were
898 used for normalizing diversity in those measurements that controlled for gBGC and
899 hotspots.

900 When normalizing diversity by divergence, only 100 kb bins that had at least 10
901 kb of divergence information were used (21,100 bins total for 13-filter set, 20,935 bins
902 total for the 13-filter set plus the additional gBGC and hotspot filters). Bins with less than
903 10 kb of divergence information were ignored. To make estimates comparable, in those

904 measurements of diversity that did not normalize by divergence, diversity was still calcu-
905 lated using the same set of 100 kb bins that had at least 10 kb for estimating diver-
906 gence.
907
908 **Calculations of population differentiation (F_{ST}) and linear regression**
909 F_{ST} calculations were performed as a function of *B* and recombination rate be-
910 tween every pair of non-admixed phase 3 TGP populations not belonging to the same
911 continental group (150 pairs total). We followed the recommendations in Bhatia et al.
912 [56] to limit biases in F_{ST} due to 1) type of estimator used, 2) averaging over SNPs, and
913 3) SNP ascertainment. Specifically, we 1) used the Hudson-based F_{ST} estimator [127],
914 2) used a ratio of averages for combining F_{ST} estimated across different SNPs, and 3)
915 ascertained SNPs based on being polymorphic in an outgroup (i.e., the KhoeSan). For
916 ascertaining SNPs in the KhoeSan, we also performed filtering according to the filtering
917 scheme described under "Filtering and ascertainment scheme." For a position to be
918 considered polymorphic in the KhoeSan, at least one alternate allele and one reference
919 allele had to be called across the 13 genomes we utilized (see "Data"). These criteria
920 left 3,497,105 total sites in the genome in the phase 3 dataset for F_{ST} to be estimated
921 across.
922 F_{ST} was calculated across either 2% quantile bins of *B* or 2% quantile bins of
923 recombination rate in order to perform simple linear regression, with either *B* or
924 recombination rate acting as an explanatory variable using the linear model $F_{ST} = \beta_0 +$
925 $\beta_1 X + \varepsilon$ (where X represents either *B* or recombination rate [ρ]). This was done for all
926 pairwise comparisons of populations between a specific pair of continental groups (25

927 pairs total) or across all pairwise comparisons using all continental groups (150 pairs
928 total). Recombination rate was scaled to be between 0 and 1 (the minimum and maxi-
929 mum observed recombination rate was 0.0 cM/Mb and 126.88 cM/Mb, respectively) to
930 aid in the comparison of the regression coefficient with *B*. Additionally, the mean of the
931 bounds defining each quantile bin was used when defining the explanatory variables for
932 the regression. Linear regression, robust linear regression, and simple correlation were
933 performed using the lm(), rlm(), and cor() functions, respectively, in the R programming
934 language (www. r-project.org). To generate standard errors of the mean, this same pro-
935 cedure was performed on F_{ST} results generated from each of 1,000 bootstrapped
936 iterations of the data.

937 F_{ST} was also calculated across bins defined by both *B* and recombination rate
938 (ρ). For this step, bins where first defined into 2% quantile bins of either variable. Each
939 2% bin was further defined into 4% quantile bins of the other variable, generating 1,250
940 separate bins in which F_{ST} was estimated for both *B* and recombination rate
941 simultaneosly. Multiple linear regression and robust regression was performed for the
942 model $F_{ST} = \beta_0 + \beta_1 B + \beta_2 \rho + \beta_3 B\rho + \varepsilon$ with the lm() and rlm() functions in R. As with the
943 simple linear regression step, recombination rate was scaled to be between 0 and 1 and
944 the mean of the bounds defining each quantile bin was used when defining the explana-
945 tory variables.

946

947 **Bootstrapping**

948 **Diversity Estimates.** To control for the structure of linkage disequilibrium and correla-
949 tion between SNPs along the genome, we partitioned the human genome into non-

950 overlapping 100 kb bins (these bins were identical to the 100 kb bins used for estimat-
951 ing divergence) and calculated mean pairwise diversity (π) or heterozygosity within each
952 bin. We also normalized the diversity estimates by divergence within each bin. We then
953 bootstrapped individual genomes by sampling, with replacement, the 100 kb bins until
954 the number of sampled bins equaled the number of bins used for calculating the diversi-
955 ty point estimates (i.e., 21,100 bins or 20,935 bins total, depending on whether filters for
956 gBGC and hotspots were applied). 1,000 total bootstrap iterations were completed and
957 standard errors of the mean were calculated by taking the standard deviation from the
958 resulting bootstrap distribution.

959 ***F_{ST}*.** For bootstrapping F_{ST}, the human genome was partitioned into non-overlapping
960 100 kb bins and were sampled with replacement until 28,823 bins were selected (the
961 total number of non-overlapping 100 kb bins in the human autosomes). F_{ST} was then
962 calculated genome-wide for the bootstrapped genome as a function of either *B* and/or
963 recombination rate for every pairwise comparison of non-admixed phase 3 TGP popula-
964 tions not belonging to the same continental group. 1,000 total bootstrap iterations were
965 completed and standard errors of the mean were calculated by taking the standard de-
966 viation from the F_{ST} distribution calculated from all 1,000 iterations.

967

968 **Acknowledgements**

969 We thank Lawrence Uricchio, Dominic Tong, Melissa Spear, and Nicolas Strauli for
970 helpful comments on the manuscript. The computations in this paper were run on the
971 QB3 Shared Cluster at the University of California, San Francisco.

972

Author contributions

Conceived and designed the experiments: RT SAZ RDH.

Performed the experiments: RT SAZ.

Analyzed the data: RT SAZ.

Wrote the paper: RT RDH.

Wrote S2 Appendix: SAZ.

References

1. Cutter AD, Payseur BA. Genomic signatures of selection at linked sites: unifying the disparity among species. Nat Rev Genet. Nature Publishing Group; 2013;14: 262–74. doi:10.1038/nrg3425
2. Ellegren H, Galtier N. Determinants of genetic diversity. Nat Rev Genet. 2016;17: 422–433. doi:10.1038/nrg.2016.58
3. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. Nature Publishing Group; 2002;419: 832–837. doi:10.1038/nature01027.1.
4. Williamson SH, Hernandez R, Fledel-Alon a, Zhu L, Nielsen R, Bustamante CD. Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc Natl Acad Sci U S A. National Academy of Sciences; 2005;102: 7882–7887. doi:10.1073/pnas.0502300102
5. Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, et al. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. 2008;4: e1000083. doi:10.1371/journal.pgen.1000083
6. McVicker G, Gordon D, Davis C, Green P. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. 2009;5: e1000471. doi:10.1371/journal.pgen.1000471
7. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 2009;5. doi:10.1371/journal.pgen.1000695
8. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011;475: 493–496. doi:10.1038/nature10231
9. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015;526: 68–74. doi:10.1038/nature15393
10. Lack JB, Lange JD, Tang AD, Corbett-Detig RB, Pool JE. A thousand fly genomes: An expanded drosophila genome nexus. Mol Biol Evol. Oxford

1011 University Press; 2016;33: 3308–3313. doi:10.1093/molbev/msw195
1012 11. Gibbs RA, Taylor JF, Van Tassell CP, Barendse W, Eversole KA, Gill CA, et al.
1013 Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle
1014 Breeds [Internet]. NIH Public Access; 2009 Apr 24 pp. 528–532.
1015 doi:10.1126/science.1167936
1016 12. Marsden CD, Ortega-Del Vecchyo D, O'Brien DP, Taylor JF, Ramirez O, Vilà C,
1017 et al. Bottlenecks and selective sweeps during domestication have increased
1018 deleterious genetic variation in dogs. Proc Natl Acad Sci. 2016;113: 152–157.
1019 doi:10.1073/pnas.1512501113
1020 13. Caicedo AL, Williamson SH, Hernandez RD, Boyko A, Fledel-Alon A, York TL, et
1021 al. Genome-wide patterns of nucleotide polymorphism in domesticated rice. PLoS
1022 Genet. Sinauer Associates; 2007;3: 1745–1756.
1023 doi:10.1371/journal.pgen.0030163
1024 14. Begun DJ, Aquadro CF. African and North American populations of Drosophila
1025 melanogaster are very different at the DNA level. Nature. 1993;365: 548–550.
1026 doi:10.1038/365548a0
1027 15. Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P. Multilocus patterns of
1028 nucleotide variability and the demographic and selection history of Drosophila
1029 melanogaster populations. Genome Res. Cold Spring Harbor Laboratory Press;
1030 2005;15: 790–799. doi:10.1101/gr.3541005
1031 16. Ometto L, Glinka S, De Lorenzo D, Stephan W. Inferring the effects of
1032 demography and selection on Drosophila melanogaster populations from a
1033 chromosome-wide scan of DNA variation. Mol Biol Evol. Nature Publishing Group;
1034 2005;22: 2119–2130. doi:10.1093/molbev/msi207
1035 17. Hernandez RD, Hubisz MJ, Wheeler DA, Smith DG, Ferguson B, Rogers J, et al.
1036 Demographic histories and patterns of linkage disequilibrium in Chinese and
1037 Indian rhesus macaques. Science. 2007;316: 240–3.
1038 doi:10.1126/science.1140462
1039 18. Ramachandran S, Deshpande O, Roseman CC, Rosenberg N a, Feldman MW,
1040 Cavalli-Sforza LL. Support from the relationship of genetic and geographic
1041 distance in human populations for a serial founder effect originating in Africa. Proc
1042 Natl Acad Sci U S A. 2005;102: 15942–15947. doi:10.1073/pnas.0507611102
1043 19. Henn BM, Cavalli-Sforza LL, Feldman MW. The great human expansion. Proc
1044 Natl Acad Sci U S A. 2012;109: 17758–64. doi:10.1073/pnas.1212380109
1045 20. Charlesworth B. Fundamental concepts in genetics: effective population size and
1046 patterns of molecular evolution and variation. Nat Rev Genet. 2009;10: 195–205.
1047 doi:10.1038/nrg2526
1048 21. Charlesworth D. Balancing selection and its effects on sequences in nearby
1049 genome regions. PLoS Genet. Benjamin/Cummings; 2006;2: 379–384.
1050 doi:10.1371/journal.pgen.0020064
1051 22. Maynar Smith J, Haigh J. The hitch-hiking effect of a favourable gene. Genet Res.
1052 1974;23: 23–35. doi:10.1017/S0016672308009579
1053 23. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations
1054 on neutral molecular variation. Genetics. 1993;134: 1289–303. Available:
1055 http://www.ncbi.nlm.nih.gov/pubmed/8375663
1056 24. Kim Y, Stephan W. Joint effects of genetic hitchhiking and background selection

1057 on neutral variation. Genetics. 2000;155: 1415–1427. doi:drosophila-
1058 melanogaster; deleterious mutations; recombination; polymorphism; rates;
1059 population; evolution; variance; linkage; locus
1060 25. Begun DJ, Aquadro CF. Levels of naturally occurring DNA polymorphism
1061 correlate with recombination rates in D. melanogaster. Nature. Nature Publishing
1062 Group; 1992;356: 519–520. doi:10.1038/356519a0
1063 26. Charlesworth B. Background selection and patterns of genetic diversity in
1064 Drosophila melanogaster. Genet Res. 1996;68: 131–149.
1065 doi:10.1017/S0016672300034029
1066 27. Andolfatto P. Hitchhiking effects of recurrent beneficial amino acid substitutions in
1067 the Drosophila melanogaster genome. Genome Res. Cold Spring Harbor
1068 Laboratory Press; 2007;17: 1755–1762. doi:10.1101/gr.6691007
1069 28. Sella G, Petrov DA, Przeworski M, Andolfatto P. Pervasive natural selection in the
1070 Drosophila genome? [Internet]. Nachman MW, editor. PLoS Genetics. Plenum;
1071 2009. p. e1000495. doi:10.1371/journal.pgen.1000495
1072 29. Comeron JM. Background Selection as Baseline for Nucleotide Variation across
1073 the Drosophila Genome. Begun DJ, editor. PLoS Genet. 2014;10: e1004434.
1074 doi:10.1371/journal.pgen.1004434
1075 30. Elyashiv E, Sattath S, Hu TT, Strutsovsky A, McVicker G, Andolfatto P, et al. A
1076 Genomic Map of the Effects of Linked Selection in Drosophila. Barton NH, editor.
1077 PLoS Genet. Public Library of Science; 2016;12: e1006130.
1078 doi:10.1371/journal.pgen.1006130
1079 31. Flowers JM, Molina J, Rubinstein S, Huang P, Schaal BA, Purugganan MD.
1080 Natural selection in gene-dense regions shapes the genomic pattern of
1081 polymorphism in wild and domesticated rice. Mol Biol Evol. Oxford University
1082 Press; 2012;29: 675–687. doi:10.1093/molbev/msr225
1083 32. Xu X, Liu X, Ge S, Jensen JJDJJDJ, Hu F, Li X, et al. Resequencing 50
1084 accessions of cultivated and wild rice yields markers for identifying agronomically
1085 important genes. Nat Biotechnol. 2012;30: 105–11. doi:10.1038/nbt.2050
1086 33. Andersen EC, Gerke JP, Shapiro JA, Crissman JR, Ghosh R, Bloom JS, et al.
1087 Chromosome-scale selective sweeps shape Caenorhabditis elegans genomic
1088 diversity. Nat Genet. 2012;44: 285–90. doi:10.1038/ng.1050
1089 34. Cutter AD, Payseur BA. Selection at linked sites in the partial selfer
1090 Caenorhabditis elegans. Mol Biol Evol. 2003;20: 665–673.
1091 doi:10.1093/molbev/msg072
1092 35. Reed FA, Akey JM, Aquadro CF. Fitting background-selection predictions to
1093 levels of nucleotide variation and divergence along the human autosomes.
1094 Genome Res. 2005;15: 1211–1221. doi:10.1101/gr.3413205
1095 36. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection
1096 in the human genome. Hurst L, editor. PLoS Biol. Public Library of Science;
1097 2006;4: 0446–0458. doi:10.1371/journal.pbio.0040072
1098 37. Cai JJ, Macpherson JM, Sella G, Petrov DA. Pervasive hitchhiking at coding and
1099 regulatory sites in humans. PLoS Genet. 2009;5: e1000336.
1100 doi:10.1371/journal.pgen.1000336
1101 38. Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, et al.
1102 Classic Selective Sweeps Were Rare in Recent Human Evolution. Science (80-).

2011;331: 920–924. doi:10.1126/science.1198878
39. Lohmueller KE, Albrechtsen A, Li Y, Kim SY, Korneliussen T, Vinckenbosch N, et al. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. PLoS Genet. 2011;7: e1002326. doi:10.1371/journal.pgen.1002326
40. Alves I, Šrámková Hanulová A, Foll M, Excoffier L. Genomic data reveal a complex making of humans [Internet]. Schierup MH, editor. PLoS Genetics. Public Library of Science; 2012. p. e1002837. doi:10.1371/journal.pgen.1002837
41. Granka JM, Henn BM, Gignoux CR, Kidd JM, Bustamante CD, Feldman MW. Limited evidence for classic selective sweeps in African populations. Genetics. 2012;192: 1049–1064. doi:10.1534/genetics.112.144071
42. Enard D, Messer PW, Petrov DA. Genome-wide signals of positive selection in human evolution. Genome Res. Cold Spring Harbor Laboratory Press; 2014;24: 885–895. doi:10.1101/gr.164822.113
43. Bank C, Ewing GB, Ferrer-Admettla A, Foll M, Jensen JD. Thinking too positive? Revisiting current methods of population genetic selection inference. Trends Genet. Elsevier Ltd; 2014;30: 540–546. doi:10.1016/j.tig.2014.09.010
44. Corbett-Detig RB, Hartl DL, Sackton TB. Natural Selection Constrains Neutral Diversity across A Wide Range of Species. Barton NH, editor. PLoS Biol. Springer; 2015;13: e1002112. doi:10.1371/journal.pbio.1002112
45. Zeng K. A coalescent model of background selection with recombination, demography and variation in selection coefficients. Heredity (Edinb). Nature Publishing Group; 2013;110: 363–71. doi:10.1038/hdy.2012.102
46. Nicolaisen LE, Desai MM. Distortions in Genealogies due to Purifying Selection and Recombination. Genetics. 2013;195: 221–230. doi:10.1534/genetics.113.152983
47. Pennings PS, Kryazhimskiy S, Wakeley J. Loss and Recovery of Genetic Diversity in Adapting Populations of HIV. Fraser C, editor. PLoS Genet. Public Library of Science; 2014;10: e1004000. doi:10.1371/journal.pgen.1004000
48. Brandvain Y, Wright SI. The Limits of Natural Selection in a Nonequilibrium World. Trends Genet. Elsevier Ltd; 2016;32: 201–210. doi:10.1016/j.tig.2016.01.004
49. Koch E, Novembre J. A Temporal Perspective on the Interplay of Demography and Selection on Deleterious Variation in Humans. G3 (Bethesda). 2017;7: g3.117.039651. doi:10.1534/g3.117.039651
50. Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, Hubisz MJ, et al. Proportionally more deleterious genetic variation in European than in African populations. Nature. 2008;451: 994–997. doi:10.1038/nature06611
51. Simons YB, Turchin MC, Pritchard JK, Sella G. The deleterious mutation load is insensitive to recent population history. Nat Genet. 2014;46: 220–4. doi:10.1038/ng.2896
52. Do R, Balick D, Li H, Adzhubei I, Sunyaev S, Reich D. No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. Nat Genet. 2015;47: 126–131. doi:10.1038/ng.3186
53. Nordborg M, Charlesworth B, Charlesworth D. The effect of recombination on background selection. Genet Res. 1996;67: 159–74. doi:10.1017/S0016672300033619

1149 54. Moreno-Estrada A, Gignoux CR, Fernández-López JC, Zakharia F, Sikora M,
1150 Contreras A V, et al. The genetics of Mexico recapitulates Native American
1151 substructure and affects biomedical traits. Science (80-). 2014;344: 1280–5.
1152 doi:10.1126/science.1251688
1153 55. Hey J. On the number of new world founders: A population genetic portrait of the
1154 peopling of the Americas. Clark AG, editor. PLoS Biol. Springer; 2005;3: 0965–
1155 0975. doi:10.1371/journal.pbio.0030193
1156 56. Bhatia G, Patterson N, Sankararaman S, Price AL. Estimating and interpreting
1157 FST: The impact of rare variants. Genome Res. 2013;23: 1514–1521.
1158 doi:10.1101/gr.154831.113
1159 57. Keinan A, Reich D. Human population differentiation is strongly correlated with
1160 local recombination rate. Begun DJ, editor. PLoS Genet. Public Library of
1161 Science; 2010;6: e1000886. doi:10.1371/journal.pgen.1000886
1162 58. Willing EM, Dreyer C, van Oosterhout C. Estimates of genetic differentiation
1163 measured by fst do not necessarily require large sample sizes when using many
1164 snp markers. PLoS One. 2012;7: 1–7. doi:10.1371/journal.pone.0042649
1165 59. Messer PW, Petrov DA. Frequent adaptation and the McDonald-Kreitman test.
1166 Proc Natl Acad Sci U S A. 2013;110: 8615–20. doi:10.1073/pnas.1220835110
1167 60. Ewing GB, Jensen JD. The consequences of not accounting for background
1168 selection in demographic inference. Mol Ecol. Wiley-Blackwell; 2016;25: 135–141.
1169 doi:10.1111/mec.13390
1170 61. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, et al.
1171 Demographic history and rare allele sharing among human populations. Proc Natl
1172 Acad Sci. 2011;108: 11983–11988. doi:10.1073/pnas.1019276108
1173 62. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al.
1174 Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing
1175 of Human Exomes. Science (80-). 2012;337: 64–69.
1176 doi:10.1126/science.1219240
1177 63. Phung TN, Huber CD, Lohmueller KE. Determining the Effect of Natural Selection
1178 on Linked Neutral Divergence across Species. Akey JM, editor. PLoS Genet. W.
1179 H. Freeman; 2016;12: e1006199. doi:10.1371/journal.pgen.1006199
1180 64. Burgess R, Yang Z. Estimation of hominoid ancestral population sizes under
1181 Bayesian coalescent models incorporating mutation rate variation and sequencing
1182 errors. Mol Biol Evol. Oxford University Press; 2008;25: 1979–1994.
1183 doi:10.1093/molbev/msn148
1184 65. Mouse Genome Sequencing Consortium MGS, Waterston RH, Lindblad-Toh K,
1185 Birney E, Rogers J, Abril JF, et al. Initial sequencing and comparative analysis of
1186 the mouse genome. Nature. 2002;420: 520–62. doi:10.1038/nature01262
1187 66. Charlesworth B, Nordborg M, Charlesworth D. The effects of local selection,
1188 balanced polymorphism and background selection on equilibrium patterns of
1189 genetic diversity in subdivided populations. Genet Res. 1997;70: 155–174.
1190 doi:10.1017/S0016672397002954
1191 67. Hu XS, He F. Background selection and population differentiation. J Theor Biol.
1192 2005;235: 207–219. doi:10.1016/j.jtbi.2005.01.004
1193 68. Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Absher D, et al. The role of
1194 geography in human adaptation. Schierup MH, editor. PLoS Genet. R Foundation

1195 for Statistical Computing; 2009;5: e1000500. doi:10.1371/journal.pgen.1000500
1196 69. Bhatia G, Patterson N, Pasaniuc B, Zaitlen N, Genovese G, Pollack S. Genome-
1197 Wide Comparison of African-Ancestry Populations from CARe and Other Cohorts
1198 Reveals Signals of Natural Selection. 2011;89: 368–381.
1199 doi:10.1016/j.ajhg.2011.07.025.
1200 70. Eyre-Walker A, Keightley PD. Estimating the rate of adaptive molecular evolution
1201 in the presence of slightly deleterious mutations and population size change. Mol
1202 Biol Evol. 2009;26: 2097–2108. doi:10.1093/molbev/msp119
1203 71. Galtier N. Adaptive Protein Evolution in Animals and the Effective Population Size
1204 Hypothesis. Schierup MH, editor. PLoS Genet. Oxford University Press; 2016;12:
1205 e1005774. doi:10.1371/journal.pgen.1005774
1206 72. Garud NR, Messer PW, Buzbas EO, Petrov DA. Recent selective sweeps in North
1207 American Drosophila melanogaster show signatures of soft sweeps. Copenhaver
1208 GP, editor. PLoS Genet. Wiley; 2015;11: e1005004.
1209 doi:10.1371/journal.pgen.1005004
1210 73. Szpiech ZA, Hernandez RD. Selscan: An efficient multithreaded program to
1211 perform EHH-based scans for positive selection. Mol Biol Evol. 2014;31: 2824–
1212 2827. doi:10.1093/molbev/msu211
1213 74. Fu W, Gittelman RM, Bamshad MJ, Akey JM. Characteristics of neutral and
1214 deleterious protein-coding variation among individuals and populations. Am J
1215 Hum Genet. Springer-Verlag, New York; 2014;95: 421–436.
1216 doi:10.1016/j.ajhg.2014.09.006
1217 75. Henn BM, Botigué LR, Peischl S, Dupanloup I, Lipatov M, Maples BK, et al.
1218 Distance from sub-Saharan Africa predicts mutational load in diverse human
1219 genomes. Proc Natl Acad Sci. National Academy of Sciences; 2015;113:
1220 201510805. doi:10.1073/pnas.1510805112
1221 76. Casals F, Hodgkinson A, Hussin J, Idaghdour Y, Bruat V, de Maillard T, et al.
1222 Whole-Exome Sequencing Reveals a Rapid Change in the Frequency of Rare
1223 Functional Variants in a Founding Population of Humans. Williams SM, editor.
1224 PLoS Genet. Public Library of Science; 2013;9: e1003815.
1225 doi:10.1371/journal.pgen.1003815
1226 77. Lim ET, Würtz P, Havulinna AS, Palta P, Tukiainen T, Rehnström K, et al.
1227 Distribution and Medical Impact of Loss-of-Function Variants in the Finnish
1228 Founder Population. Cutler D, editor. PLoS Genet. Public Library of Science;
1229 2014;10: e1004494. doi:10.1371/journal.pgen.1004494
1230 78. Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, et al. Whole-
1231 genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet.
1232 2011;43: 956–963. doi:10.1038/ng.911
1233 79. Renaut S, Rieseberg LH. The accumulation of deleterious mutations as a
1234 consequence of domestication and improvement in sunflowers and other
1235 compositae crops. Mol Biol Evol. Oxford University Press; 2015;32: 2273–2283.
1236 doi:10.1093/molbev/msv106
1237 80. Balick DJ, Do R, Cassa CA, Reich D, Sunyaev SR. Dominance of Deleterious
1238 Alleles Controls the Response to a Population Bottleneck. Coop G, editor. PLoS
1239 Genet. Cambridge University Press; 2015;11: e1005436.
1240 doi:10.1371/journal.pgen.1005436

1241 81. Simons YB, Sella G. The impact of recent population history on the deleterious
1242 mutation load in humans and close evolutionary relatives [Internet]. Current
1243 Opinion in Genetics and Development. 2016. pp. 150–158.
1244 doi:10.1016/j.gde.2016.09.006
1245 82. Gravel S. When is selection effective? Genetics. 2016;203: 451–462.
1246 doi:10.1534/genetics.115.184630
1247 83. Charlesworth B. The Effects of Deleterious Mutations on Evolution at Linked
1248 Sites. Genetics. 2012;190: 5–22. doi:10.1534/genetics.111.134288
1249 84. Coventry A, Bull-Otterson LM, Liu X, Clark AG, Maxwell TJ, Crosby J, et al. Deep
1250 resequencing reveals excess rare recent variants consistent with explosive
1251 population growth. Nat Commun. Nature Publishing Group; 2010;1: 131.
1252 doi:10.1038/ncomms1130
1253 85. Keinan A, Clark AG. Recent explosive human population growth has resulted in
1254 an excess of rare genetic variants. Science (80-). 2012;336: 740–3.
1255 doi:10.1126/science.1217283
1256 86. Nelson MR, Wegmann D, Ehm MG, Kessner D, St. Jean P, Verzilli C, et al. An
1257 Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in
1258 14,002 People. Science (80-). 2012;337: 100–104. doi:10.1126/science.1217876
1259 87. Gazave E, Ma L, Chang D, Coventry A, Gao F, Muzny D, et al. Neutral genomic
1260 regions refine models of recent rapid human population growth. Proc Natl Acad
1261 Sci U S A. 2014;111: 757–62. doi:10.1073/pnas.1310398110
1262 88. Maher MC, Uricchio LH, Torgerson DG, Hernandez RD. Population Genetics of
1263 Rare Variants and Complex Diseases. Hum Hered. NIH Public Access; 2012;74:
1264 118–128. doi:10.1159/000346826
1265 89. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of
1266 6,515 exomes reveals the recent origin of most human protein-coding variants.
1267 Nature. NIH Public Access; 2013;493: 216–220. doi:10.1038/nature11690
1268 90. Fay JC, Wu CI. A human population bottleneck can account for the discordance
1269 between patterns of mitochondrial versus nuclear DNA variation. Mol Biol Evol.
1270 1999;16: 1003–1005. doi:10.1093/oxfordjournals.molbev.a026175
1271 91. Wall JD, Andolfatto P, Przeworski M. Testing Models of Selection and
1272 Demography in Drosophila simulans. Genetics. 2002;162. Available:
1273 http://www.genetics.org/content/162/1/203
1274 92. Pool JE, Nielsen R. Population size changes reshape genomic patterns of
1275 diversity. Evolution (N Y). 2007;61: 3001–3006. doi:10.1111/j.1558-
1276 5646.2007.00238.x
1277 93. Gottipati S, Arbiza L, Siepel A, Clark AG, Keinan A. Analyses of X-linked and
1278 autosomal genetic variation in population-scale whole genome sequencing. Nat
1279 Genet. NIH Public Access; 2011;43: 741–743. doi:10.1038/ng.877
1280 94. Arbiza L, Gottipati S, Siepel A, Keinan A. Contrasting X-linked and autosomal
1281 diversity across 14 human populations. Am J Hum Genet. 2014;94: 827–844.
1282 doi:10.1016/j.ajhg.2014.04.011
1283 95. Wilson Sayres MA, Lohmueller KE, Nielsen R. Natural Selection Reduced
1284 Diversity on Human Y Chromosomes. PLoS Genet. 2014;10.
1285 doi:10.1371/journal.pgen.1004064
1286 96. Hammer MF, Woerner AE, Mendez FL, Watkins JC, Cox MP, Wall JD. The ratio

of human X chromosome to autosome diversity is positively correlated with genetic distance from genes. Nat Genet. 2010;42: 830–1. doi:10.1038/ng.651

97. Leffler EM, Bullaughey K, Matute DR, Meyer WK, Ségurel L, Venkat A, et al. Revisiting an Old Riddle: What Determines Genetic Diversity Levels within Species? PLoS Biol. Public Library of Science; 2012;10: e1001388. doi:10.1371/journal.pbio.1001388

98. Vucetich JA, Waite TA, Nunney L. Fluctuating Population Size and the Ratio of Effective to Census Population Size Published by : Society for the Study of Evolution Stable URL : http://www.jstor.org/stable/2411022. Source Evol. 2008;51: 2017–2021. doi:10.2307/2411022

99. Coop G. Does linked selection explain the narrow range of genetic diversity across species? [Internet]. bioRxiv. 2016. doi:10.1101/042598

100. Lewontin RC. The Genetic Basis of Evolutionary Change [Internet]. The Genetic Basis of Evolutionary Change. New York and London: Columbia University Press; 1974. doi:10.1136/ard.2010.140574

101. Nam K, Munch K, Mailund T, Nater A, Greminger MP, Krützen M, et al. Evidence that the rate of strong selective sweeps increases with population size in the great apes. Proc Natl Acad Sci. National Academy of Sciences; 2016;114: 1–6. doi:10.1073/pnas.1605660114

102. Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, et al. Great ape genetic diversity and population history. Nature. Nature Research; 2013;499: 471–475. doi:10.1038/nature12228

103. McManus KF, Kelley JL, Song S, Veeramah KR, Woerner AE, Stevison LS, et al. Inference of gorilla demographic and selective history from whole-genome sequence data. Mol Biol Evol. Cambridge University Press, Cambridge; 2015;32: 600–612. doi:10.1093/molbev/msu394

104. Nater A, Greminger MP, Arora N, Van Schaik CP, Goossens B, Singleton I, et al. Reconstructing the demographic history of orang-utans using Approximate Bayesian Computation. Mol Ecol. 2015;24: 310–327. doi:10.1111/mec.13027

105. Kuhlwilm M, de Manuel M, Nater A, Greminger MP, Kr?tzen M, Marques-Bonet T. Evolution and demography of the great apes. Curr Opin Genet Dev. 2016;41: 124–129. doi:10.1016/j.gde.2016.09.005

106. Uricchio LH, Torres R, Witte JS, Hernandez RD. Population genetic simulations of complex phenotypes with implications for rare variant association tests. Genet Epidemiol. 2015;39: 35–44. doi:10.1002/gepi.21866

107. Uricchio LH, Zaitlen NA, Ye CJ, Witte JS, Hernandez RD. Selection and explosive growth alter genetic architecture and hamper the detection of causal rare variants. Genome Res. Cold Spring Harbor Laboratory Press; 2016;26: 863–873. doi:10.1101/gr.202440.115

108. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27: 2156–2158. doi:10.1093/bioinformatics/btr330

109. Kidd J, Sharpton T, Bobo D, Norman P, Martin A, Carpenter M, et al. Exome capture from saliva produces high quality genomic and metagenomic data. BMC Genomics. BioMed Central; 2014;15: 262. doi:10.1186/1471-2164-15-262

110. Kim HL, Ratan A, Perry GH, Montenegro A, Miller W, Schuster SC. Khoisan

1333 hunter-gatherers have been the largest population throughout most of modern-
1334 human demographic history. Nat Commun. 2014;5: 5692.
1335 doi:10.1038/ncomms6692
1336 111. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral
1337 substitution rates on mammalian phylogenies. Genome Res. Cold Spring Harbor
1338 Laboratory Press; 2010;20: 110–121. doi:10.1101/gr.097857.109
1339 112. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al.
1340 Evolutionarily conserved elements in vertebrate, insect, worm, and yeast
1341 genomes. Genome Res. 2005;15: 1034–1050. doi:10.1101/gr.3715005
1342 113. Consortium ENCODEP, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C,
1343 et al. An integrated encyclopedia of DNA elements in the human genome. Nature.
1344 2012;489: 57–74. doi:10.1038/nature11247
1345 114. Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. Segmental duplications:
1346 organization and impact within the current human genome project assembly.
1347 Genome Res. 2001;11: 1005–17. doi:10.1101/gr.187101
1348 115. Capra JA, Hubisz MJ, Kostka D, Pollard KS, Siepel A. A Model-Based Analysis of
1349 GC-Biased Gene Conversion in the Human and Chimpanzee Genomes. Coop G,
1350 editor. PLoS Genet. Public Library of Science; 2013;9: e1003684.
1351 doi:10.1371/journal.pgen.1003684
1352 116. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, et al. A second
1353 generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449:
1354 851–61. doi:10.1038/nature06258
1355 117. Pratto F, Brick K, Khil P, Smagulova F, Petukhova G V, Camerini-Otero RD. DNA
1356 recombination. Recombination initiation maps of individual human genomes.
1357 Science. 2014;346: 1256442. doi:10.1126/science.1256442
1358 118. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic
1359 variants from high-throughput sequencing data. Nucleic Acids Res. Oxford
1360 University Press; 2010;38: e164–e164. doi:10.1093/nar/gkq603
1361 119. Hernandez RD, Williamson SH, Zhu L, Bustamante CD. Context-dependent
1362 mutation rates may cause spurious signatures of a fixation bias favoring higher
1363 GC-content in humans. Mol Biol Evol. 2007;24: 2196–2202.
1364 doi:10.1093/molbev/msm149
1365 120. Schrider DR, Shanku AG, Kern AD. Effects of linked selective sweeps on
1366 demographic inference and model selection. Genetics. 2016;204: 1207–1223.
1367 doi:10.1534/genetics.116.190223
1368 121. Palamara PF, Francioli LC, Wilton PR, Genovese G, Gusev A, Finucane HK, et
1369 al. Leveraging Distant Relatedness to Quantify Human Mutation and Gene-
1370 Conversion Rates. Am J Hum Genet. 2015;97: 775–789.
1371 doi:10.1016/j.ajhg.2015.10.006
1372 122. Hernandez RD. A flexible forward simulator for populations subject to selection
1373 and demography. Bioinformatics. 2008;24: 2786–2787.
1374 doi:10.1093/bioinformatics/btn522
1375 123. Torgerson DG, Boyko AR, Hernandez RD, Indap A, Hu X, White TJ, et al.
1376 Evolutionary processes acting on candidate cis-regulatory regions in humans
1377 inferred from patterns of polymorphism and divergence. PLoS Genet. 2009;5.
1378 doi:10.1371/journal.pgen.1000592

124. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. Am J Hum Genet. The American Society of Human Genetics; 2013;93: 278–288. doi:10.1016/j.ajhg.2013.06.020
125. O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, et al. A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. PLoS Genet. Public Library of Science; 2014;10: e1004234. doi:10.1371/journal.pgen.1004234
126. Mao X, Bigham AW, Mei R, Gutierrez G, Weiss KM, Brutsaert TD, et al. A genomewide admixture mapping panel for Hispanic/Latino populations. Am J Hum Genet. 2007;80: 1171–1178. doi:10.1086/518564
127. Hudson RR, Slatkin M, Maddison WP. Estimation of levels of gene flow from DNA sequence data. Genetics. 1992;589: 583–589. Available: http://www.genetics.org/content/132/2/583.short

Supporting information

S1 Supporting information.

Fig S1. Inference models inferred from TGP CG weak BGS neutral regions and coding four-fold degenerate sites. Solid lines are the inference results from running dadi on 53 YRI (African), 64 CEU (European), and 62 CHS (East Asian) TGP CG samples (projected down to 106 chromosomes during inference procedure) across neutral regions in the weakest 1% BGS bin ($B \geq 0.994$). Broken lines represent the inference results using the same CG samples, but with sequence data only from coding four-fold degenerate synonymous sites**.**

Fig S2. Diversity for TGP non-admixed populations while controlling for GC-biased gene conversion and recombination hotspots. (A) Normalized diversity (π/divergence) measured across the strongest 1% BGS bin. (B) Normalized diversity measured across the weakest 1% BGS bin. (C) Relative diversity: the ratio of normalized diversity for the strongest 1% BGS bin to normalized diversity for the weakest 1%

1409 BGS bin (π/π_0). Error bars represent ±1 SEM calculated from 1,000 bootstrapped da-
1410 tasets.

1411 **Fig S3. Diversity for TGP non-admixed populations without normalizing by diver-**
1412 **gence with Rhesus macaque.** (A) Diversity (π) measured across the strongest 1%
1413 BGS bin. (B) Diversity measured across the weakest 1% BGS bin. (C) Relative diversi-
1414 ty: the ratio of diversity for the strongest 1% BGS bin to diversity for the weakest 1%
1415 BGS bin (π/π_0). Error bars represent ±1 SEM calculated from 1,000 bootstrapped da-
1416 tasets.

1417 **Fig S4. Diversity for TGP continental groups while controlling for GC-biased gene**
1418 **conversion and recombination hotspots.** (A) Normalized diversity (π/divergence)
1419 measured across the strongest 1%, 5%, 10% and 25% BGS bins and the weakest 1%
1420 BGS bin (as classified by *B*). (B) Relative diversity (π/π_0) for the strongest 1%, 5%, 10%,
1421 and 25% BGS bins. Error bars represent ±1 SEM calculated from 1,000 bootstrapped
1422 datasets.

1423 **Fig S5. Diversity for TGP continental groups without normalizing by divergence**
1424 **with Rhesus macaque.** (A) Diversity (π) measured across the strongest 1%, 5%, 10%
1425 and 25% BGS bins and the weakest 1% BGS bin (as classified by *B*). (B) Relative di-
1426 versity (π/π_0) for the strongest 1%, 5%, 10%, and 25% BGS bins. Error bars represent
1427 ±1 SEM calculated from 1,000 bootstrapped datasets.

1428 **Fig S6. F_{ST} measured across joint bins of *B* and recombination rate for different**
1429 **TGP continental groups**. The left panels of Figures S6 A-E show F_{ST} measured as a
1430 function of 25 4% quantile recombination rate bins conditional on three 2% quantile *B*
1431 bins (note log scale of x-axis for recombination rate). The right panels of Figures S6 A-E

show F_{ST} measured as a function of 25 4% quantile *B* bins conditional on three 2% quantile recombination rate bins. The following continental group comparisons are shown for each plot: (A) African vs. European, (B) African vs. East Asian, (C) European vs. South Asian, (D) European vs. East Asian, (E) South Asian vs. East Asian. Smaller transparent points and lines show the F_{ST} estimates and corresponding lines of best fit (using linear regression) for each of the pairwise population comparisons within a particular pair of continental groups (25 comparisons total). Larger opaque points are mean F_{ST} estimates across all pairwise comparisons within a particular pair of continental groups (with bold lines showing their corresponding lines of best fit).

Fig S7. Simulations of diversity and relative diversity under BGS using a human demographic model without migration. (A) Inferred demographic model from Complete Genomics TGP data. The demographic model used for the simulations in Figure S7 are identical to those used for Figure 6, except that migration parameters between all populations are set to 0. (B) Simulated diversity at neutral sites across populations as a function of time under our inferred demographic model without BGS (π_0 - dashed colored lines) and with BGS (π - solid colored lines). (C) Relative diversity (π/π_0) measured by taking the ratio of diversity with BGS (π) to diversity without BGS (π_0) at each time point. Note that the x-axes in all three figures are on the same scale. Time is scaled using a human generation time of 25 years per generation. Simulation data was sampled every 100 generations.

Fig S8. Simulations of diversity and relative diversity under BGS using various values of *U*. Values of *U* are provided in the header for each set of plots. Left column plots show results of simulations under a demographic model with migration between all

1455 human populations. Right column plots show results of simulations under a demograph-
1456 ic model with no migration. Colored lines represent different populations though time
1457 and are identical to those in Figure 6 and Figure S7. The demographic model used is
1458 also identical to that in Figure 6 (for simulations with migration) and Figure S7 (for simu-
1459 lations without migration). Simulation data was sampled every 100 generations.
1460
1461 **S2 Appendix. Soft sweep detection and implementation in selscan v1.2.0.**
1462
1463 **Main Figure Legends**
1464 **Fig 1. Normalized diversity and relative diversity for non-admixed populations of**
1465 **the Thousand Genomes Project (TGP)**.
1466 (A) Normalized diversity (π/divergence) measured across the strongest 1% background
1467 selection (BGS) bin. (B) Normalized diversity measured across the weakest 1% BGS
1468 bin. (C) Relative diversity: the ratio of normalized diversity for the strongest 1% BGS bin
1469 to normalized diversity for the weakest 1% BGS bin (π/π_0). BGS bins were classified by
1470 *B*. TGP population labels are indicated below each bar (see Table S11 in Supporting
1471 information for population label descriptions), with African populations colored by gold
1472 shades, European populations colored by blue shades, South Asian populations colored
1473 by violet shades, and East Asian populations colored by green shades. Error bars rep-
1474 resent ±1 SEM calculated from 1,000 bootstrapped datasets.
1475 **Fig 2. Normalized and relative diversity for Thousand Genomes Project (TGP)**
1476 **continental groups**.

1477 (A) Normalized diversity (π/divergence) measured across the strongest 1%, 5%, 10%
1478 and 25% background selection (BGS) bins and the weakest 1% BGS bin (as classified
1479 by *B*). (B) Relative diversity: the ratio of normalized diversity for each strong BGS bin in
1480 (A) to normalized diversity for the weakest 1% BGS bin (π/π_0). Error bars represent ±1
1481 SEM calculated from 1,000 bootstrapped datasets.

1482 **Fig 3. Comparing patterns of diversity between local ancestry segments of ad-**
1483 **mixed samples and continental groups**.

1484 (A) Normalized diversity (heterozygosity/divergence) and (B) Relative diversity: the ratio
1485 of normalized diversity for each strong BGS bin in (A) to normalized diversity for the
1486 weakest 1% BGS bin. Local ancestry segments include African, European, and Native
1487 American ancestries. Continental groups include African, European, and East Asian
1488 populations. Error bars represent ±1 SEM calculated from 1,000 bootstrapped datasets.

1489 **Fig 4. F_{ST} is correlated with *B* but not recombination rate**.

1490 (A) F_{ST} measured across 2% quantile bins of *B*. (B) F_{ST} measured across 2% quantile
1491 recombination rate bins. The right panel of Figure 4B displays a narrower range of re-
1492 combination rates to show detail. Smaller transparent points and lines show the esti-
1493 mates and corresponding lines of best fit (using linear regression) for F_{ST} between every
1494 pairwise population comparison for a particular pair of continental groups (25 pairwise
1495 comparisons each). Larger opaque points and lines are mean F_{ST} estimates and lines of
1496 best fit across all Thousand Genomes Project (TGP) population comparisons between a
1497 particular pair of continental groups. Error bars represent ±1 SEM calculated from 1,000
1498 bootstrapped datasets.

1499 **Fig 5. F_{ST} between African (AFR) and South Asian (SASN) populations jointly**
1500 **across *B* and recombination rate**.
1501 (A) F_{ST} as a function of 25 recombination rate bins (4% quantile bins) conditional on
1502 three different 2% quantile *B* bins (note log scale of x-axis for recombination rate). (B)
1503 F_{ST} as a function of 25 *B* bins (4% quantile bins) conditional on three different 2% quan-
1504 tile recombination rate bins. Smaller transparent points and lines show the F_{ST} esti-
1505 mates and corresponding lines of best fit (using linear regression) for each of the pair-
1506 wise comparisons of AFR vs. SASN Thousand Genomes Project (TGP) populations (25
1507 comparisons total). Larger opaque points are mean F_{ST} estimates across all pairwise
1508 comparisons of AFR vs. SASN TGP populations (with bold lines showing their corre-
1509 sponding lines of best fit).
1510 **Fig 6. Simulations confirm that demographic events shape the impact of back-**
1511 **ground selection (BGS).**
1512 (A) Inferred demographic model from Complete Genomics Thousand Genomes Project
1513 (TGP) data showing population size changes for Africans (AFR), Europeans (EUR), and
1514 East Asians (EASN) as a function of time that was used for the simulations of BGS. (B)
1515 Simulated diversity at neutral sites across populations as a function of time under our
1516 inferred demographic model without BGS (π_0 - dashed colored lines) and with BGS (π -
1517 solid colored lines). (C) Relative diversity (π/π_0) measured by taking the ratio of diversity
1518 with BGS (π) to diversity without BGS (π_0) at each time point. Note that the x-axes in all
1519 three figures are on the same scale. Time is scaled using a human generation time of
1520 25 years per generation. Simulation data was sampled every 100 generations.
1521

Figure 1

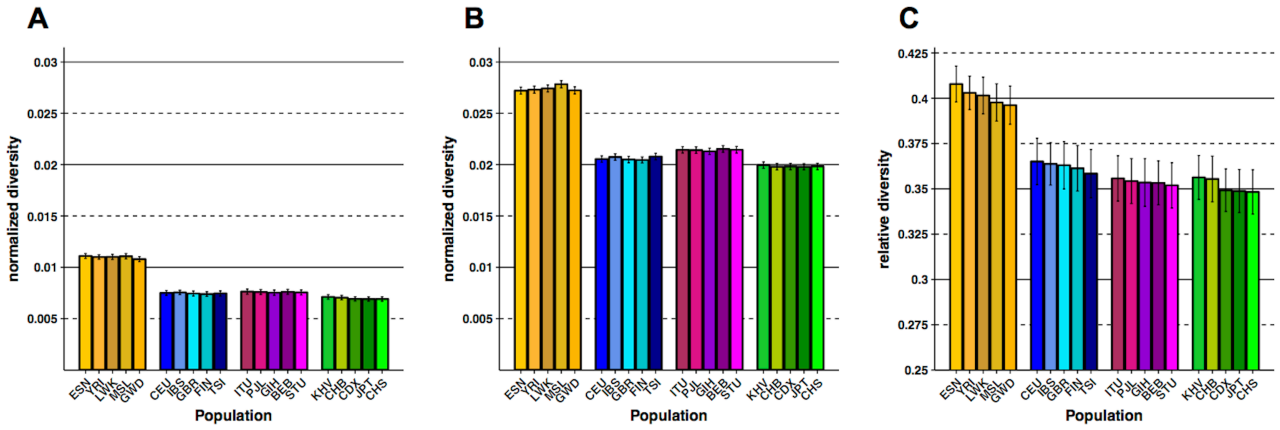

A
B
C
normalized diversity
relative diversity
Population
ESN YRI LWK MSL GWD CEU IBS GBR FIN TSI ITU PJL GIH BEB STU KHV CHB CDX JPT CHS

Figure 2

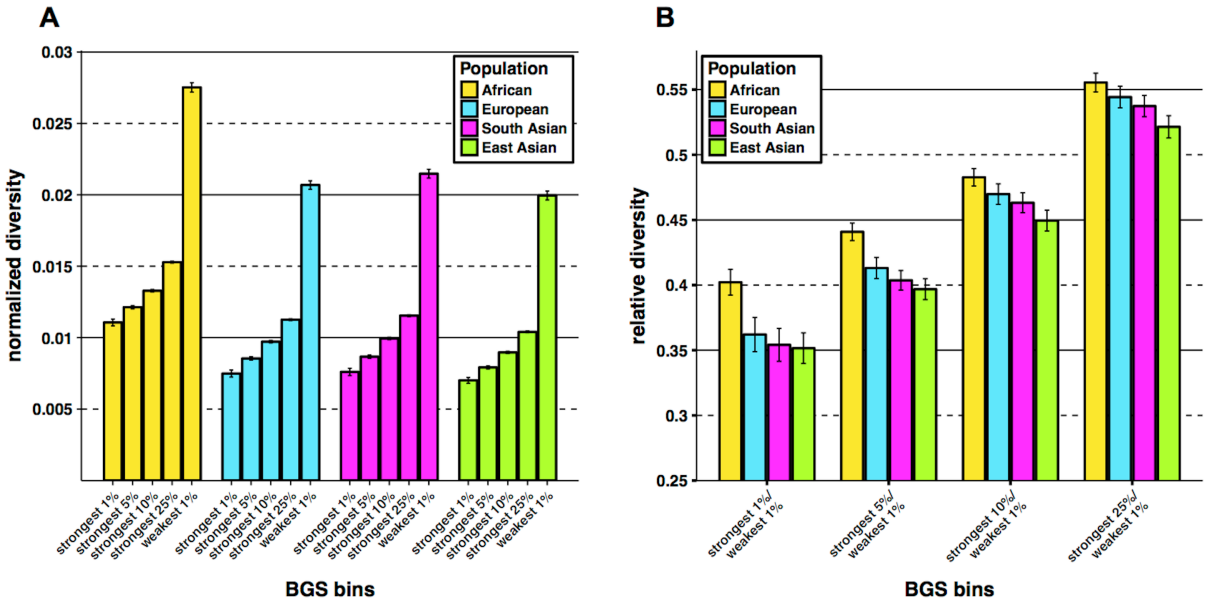

A
B
Population
African
European
South Asian
East Asian
normalized diversity
relative diversity
0.03
0.025
0.02
0.015
0.01
0.005
0.55
0.5
0.45
0.4
0.35
0.3
0.25
strongest 1%
strongest 5%
strongest 10%
strongest 25%
weakest 1%
BGS bins

Figure 3

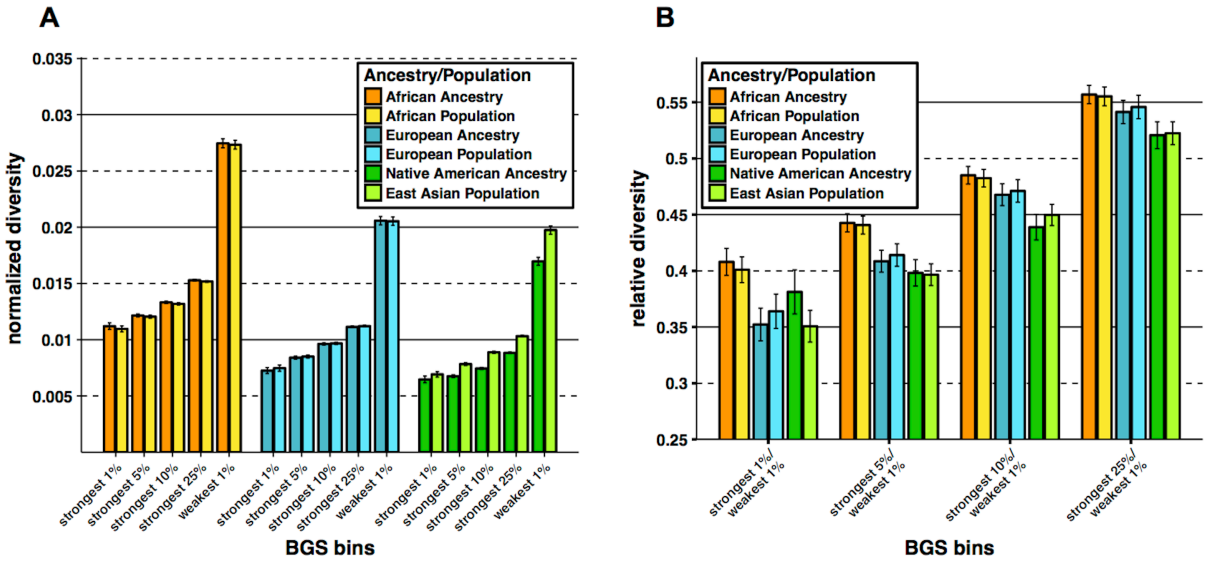

A
normalized diversity
0.035
0.03
0.025
0.02
0.015
0.01
0.005
Ancestry/Population
African Ancestry
African Population
European Ancestry
European Population
Native American Ancestry
East Asian Population
strongest 1%
strongest 5%
strongest 10%
strongest 25%
weakest 1%
BGS bins
B
relative diversity
0.55
0.5
0.45
0.4
0.35
0.3
0.25
Ancestry/Population
African Ancestry
African Population
European Ancestry
European Population
Native American Ancestry
East Asian Population
strongest 1%/ weakest 1%
strongest 5%/ weakest 1%
strongest 10%/ weakest 1%
strongest 25%/ weakest 1%
BGS bins

Figure 4

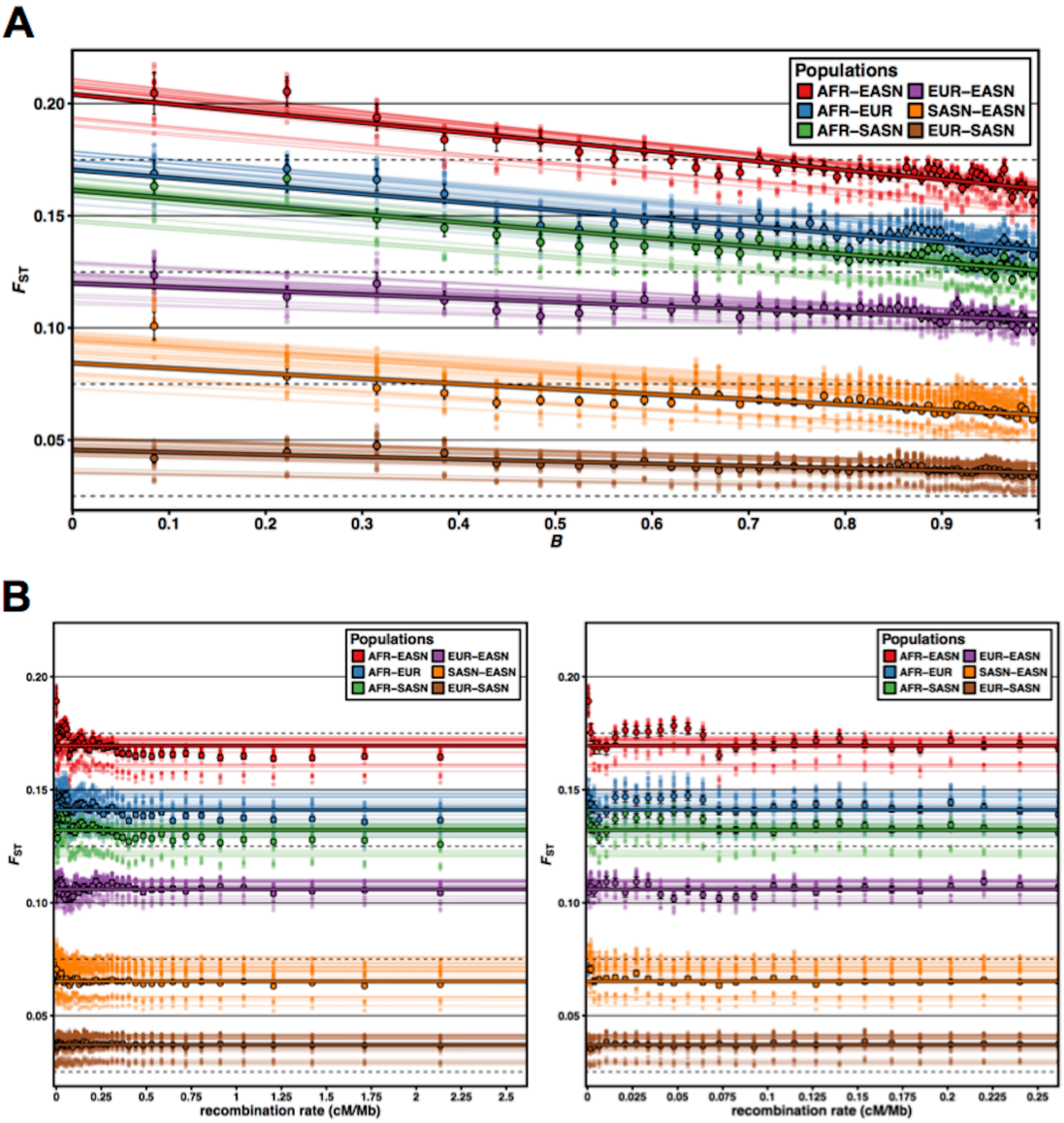

A
Populations
AFR–EASN
EUR–EASN
AFR–EUR
SASN–EASN
AFR–SASN
EUR–SASN
F_{ST}
B
recombination rate (cM/Mb)

Figure 5

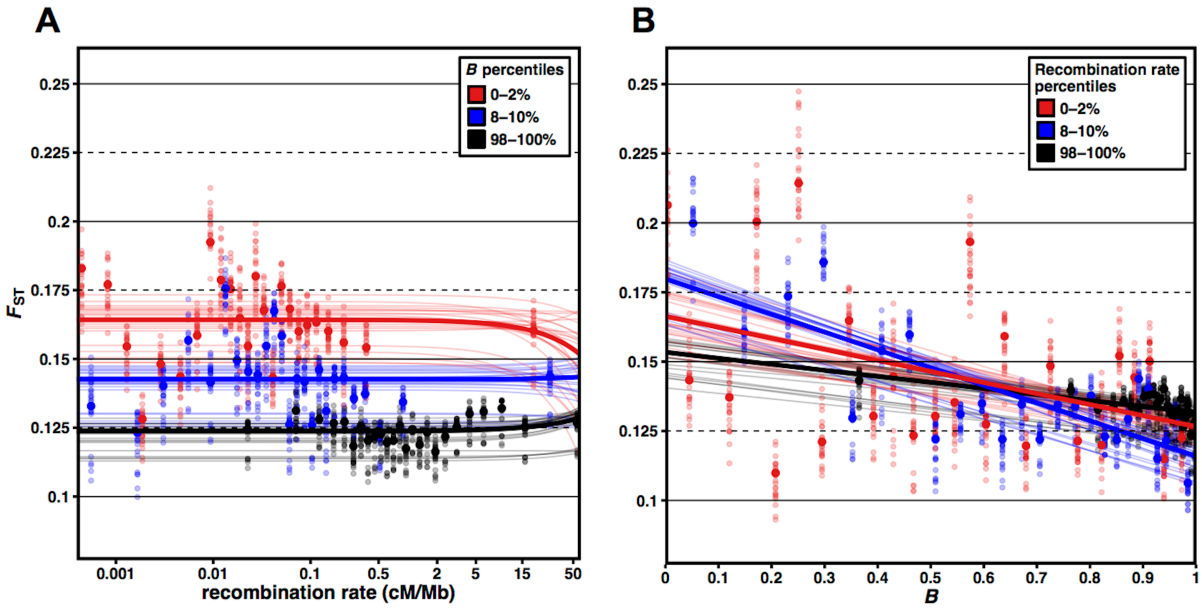
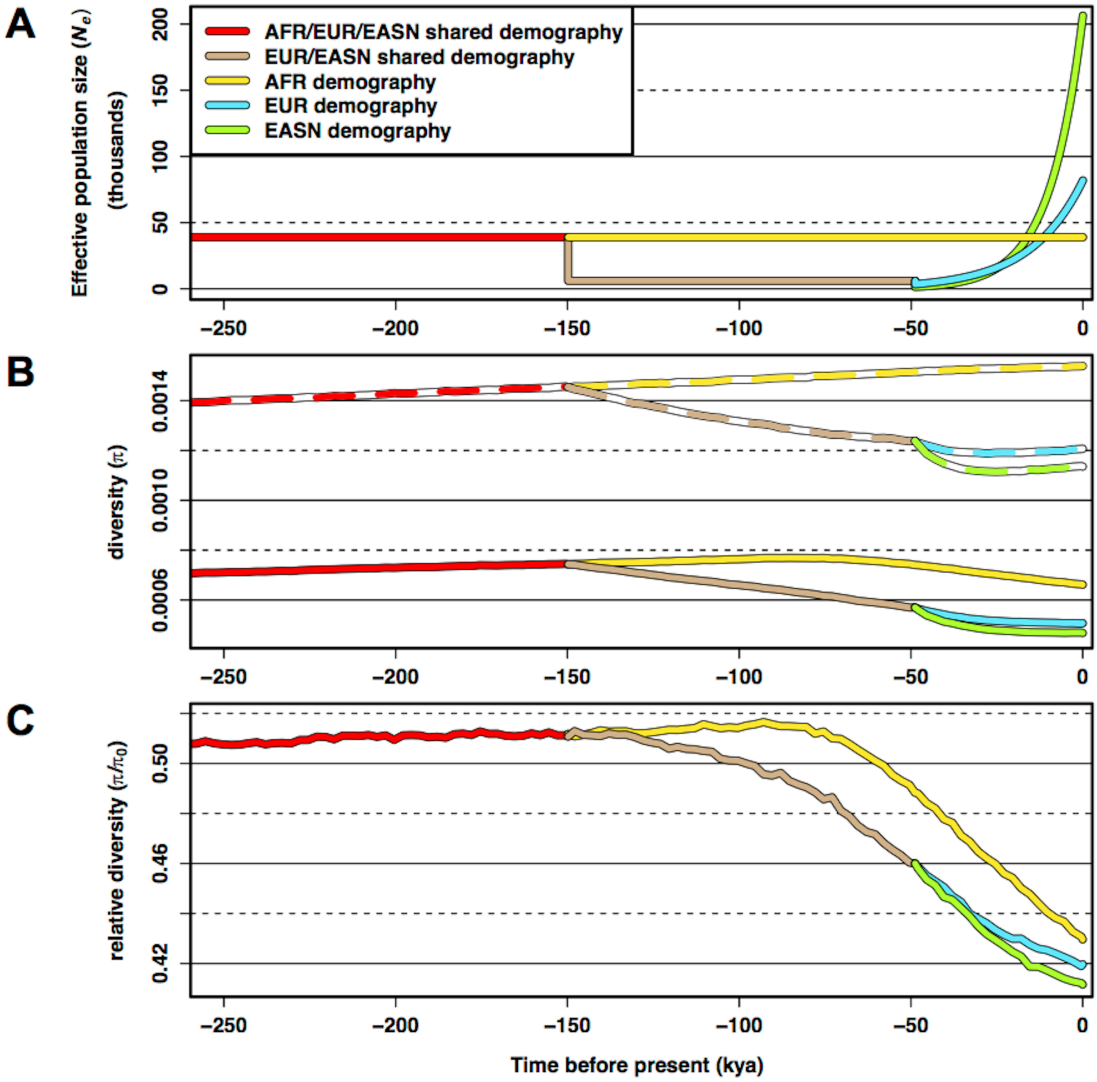

A
B
B percentiles
0–2%
8–10%
98–100%
Recombination rate percentiles
0–2%
8–10%
98–100%
F_{ST}
0.25
0.225
0.2
0.175
0.15
0.125
0.1
0.001
0.01
0.1
0.5
2
5
15
50
recombination rate (cM/Mb)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
B

Figure 6


A
B
C
AFR/EUR/EASN shared demography
EUR/EASN shared demography
AFR demography
EUR demography
EASN demography
Effective population size (N_e) (thousands)
0
50
100
150
200
diversity (π)
0.0006
0.0010
0.0014
relative diversity (π/π_0)
0.42
0.46
0.50
−250
−200
−150
−100
−50
0
Time before present (kya)

S1 Supporting information

Table S1

Parameters	$B \geq 0.994$	four-fold degenerate
$N_{Ancestral}$	18,449	17,118
N_{AFR}	38,874	47,537
N_{Bott}	5,946	6,408
N_{EUR0}	3,413	4,331
N_{EUR}	81,901	100,614
N_{EASN0}	1,317	1,678
N_{EASN}	206,804	266,616
$T_{AFR}+T_{Bott}+T_{EUR_EASN}$ (kya)	552,939	413,337
$T_{Bott}+T_{EUR_EASN}$ (kya)	149,813	198,603
T_{EUR_EASN} (kya)	48,822	69,584
r_{EUR} (%)	0.163	0.113
r_{EASN} (%)	0.259	0.182
$m_{AFR\text{-}Bott}$ ($\times 10^{-5}$)	7.83	7.02
$m_{AFR\text{-}EUR}$ ($\times 10^{-5}$)	0.51	0.47
$m_{AFR\text{-}EASN}$ ($\times 10^{-5}$)	0.13	0.18
$m_{EUR\text{-}EASN}$ ($\times 10^{-5}$)	0.98	1.14

Table S1. Inferred parameters from running dadi on TGP CG data across neutral regions in the weakest 1% BGS bin ($B \geq 0.994$) and across four-fold degenerate sites. The demographic model inferred is the Out-of-Africa demographic model of Gutenkunst et al. 2009. Time parameters, *T*, assume a generation time of 25 years per generation. Growth rates, *r*, and migration rates, *m*, are per generation. Parameters with subscript, "*Bott*", represent parameters inferred for the ancestral European and East Asian out-of-Africa bottleneck population. Time parameters with subscript "*EUR_EASN*" represent the European-East Asian population split.

Table S3

Henn et al. 2015 samples		
SampleID	**Number of Sites**	**Mean Depth**
HGDP00991	2,207,845	6.96118
HGDP00987	2,229,426	7.19132
HGDP01036	2,373,023	11.6072
HGDP00992	2,452,509	12.1913
HGDP01029	2,415,792	12.3526
HGDP01032	2,407,400	12.8113
Kidd et al. 2014 samples		
SampleID	**Number of Sites**	**Mean Depth**
SA1000A	547,527	2.56481
SA1025A	2,136,905	9.1239
Kim et al. 2014 samples		
SampleID	**Number of Sites**	**Mean Depth**
KB2	2,756,225	27.5951
NB1	2,599,220	28.0148
MD8	2,777,871	38.4532
NB8	2,778,198	40.1789
KB1	2,757,336	50.5629

Table S3. Number of polymorphic sites and mean depth coverage of 13 Khoisan samples used for SNP ascertainment in calculations of F_{ST}.

Table S4

	top 1% *B*	top 5% *B*	top 10% *B*	top 25% *B*	bottom 1% *B*
filters	7.59	40.42	87.86	246.59	13.1
filters + gBGC and hotspots removal	7.26	38.68	83.75	231.71	7.94

Table S4. Total number of Mb in the human genome passing the set of 13 filters described in Materials and Methods that were used for calculating pairwise genetic diversity (π) for each quantile of *B*. The bottom row is the total number Mb when including the set of filters to remove regions sensitive to GC-biased gene conversion (gBGC) or sites in recombination hotspots. Additionally, these totals only include those 100 kb regions that had a minimum of 10 kb of divergence information for Rhesus macaque (see Materials and Methods).

Table S5

Ancestry	**top 1% *B***	**top 5% *B***	**top 10% *B***	**top 25% *B***	**bottom 1% *B***
African	841.97	4471.54	9720.15	27333.95	1447.04
European	815.74	4296.69	9293.04	26034.57	1366.26
Native American	497.29	2603.12	5640.13	15776.71	834.46

Table S5. Total number of Mb of homozygous ancestry that passed all filters and were used in the analyses of admixed samples in the 6 admixed TGP populations (ACB, ASW, CLM, MXL, PEL, PUR) for each quantile of *B*. Additionally, these totals only include those 100 kb regions that had a minimum of 10 kb of divergence information for Rhesus macaque (see Materials and Methods).

Figure S1

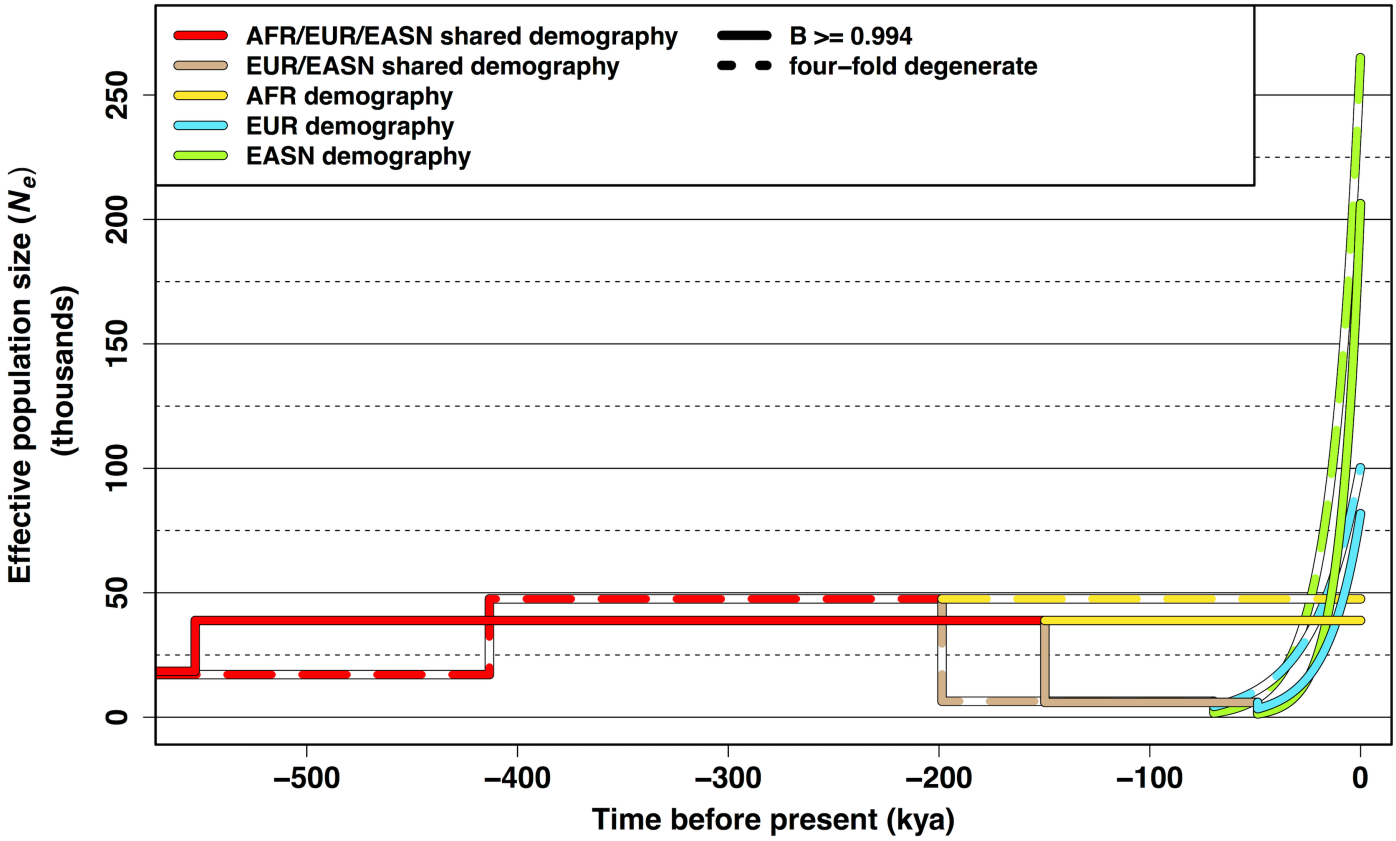

AFR/EUR/EASN shared demography
EUR/EASN shared demography
AFR demography
EUR demography
EASN demography
B >= 0.994
four–fold degenerate
Effective population size (N_e)
(thousands)
0
50
100
150
200
250
–500
–400
–300
–200
–100
0
Time before present (kya)

Figure S2

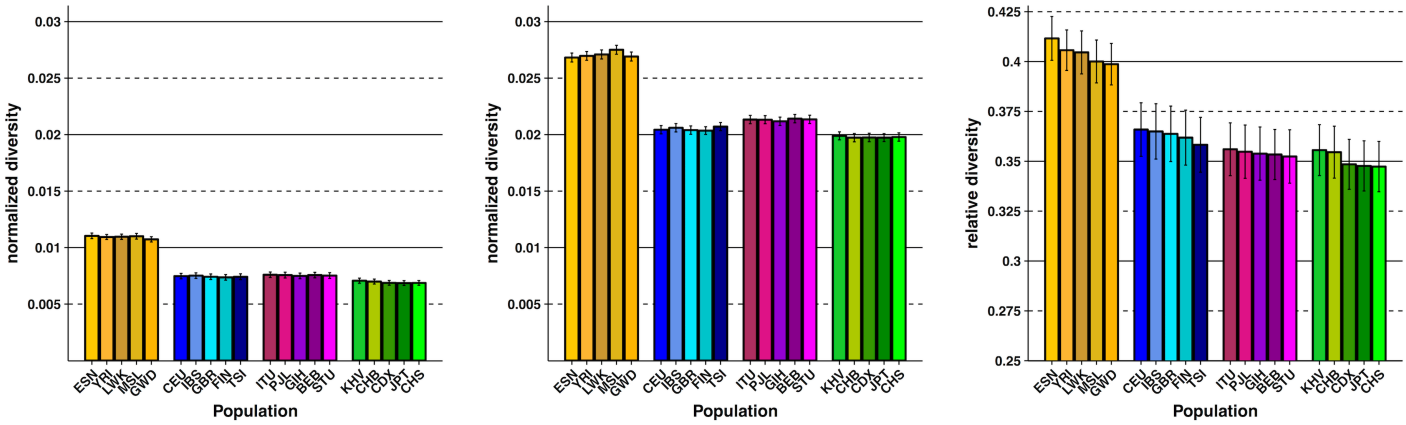

normalized diversity
0.03
0.025
0.02
0.015
0.01
0.005
ESN YRI LWK MSL GWD CEU IBS GBR FIN TSI ITU PJL GIH BEB STU KHV CHB CDX JPT CHS
Population
normalized diversity
0.03
0.025
0.02
0.015
0.01
0.005
ESN YRI LWK MSL GWD CEU IBS GBR FIN TSI ITU PJL GIH BEB STU KHV CHB CDX JPT CHS
Population
relative diversity
0.425
0.4
0.375
0.35
0.325
0.3
0.275
0.25
ESN YRI LWK MSL GWD CEU IBS GBR FIN TSI ITU PJL GIH BEB STU KHV CHB CDX JPT CHS
Population

Figure S3

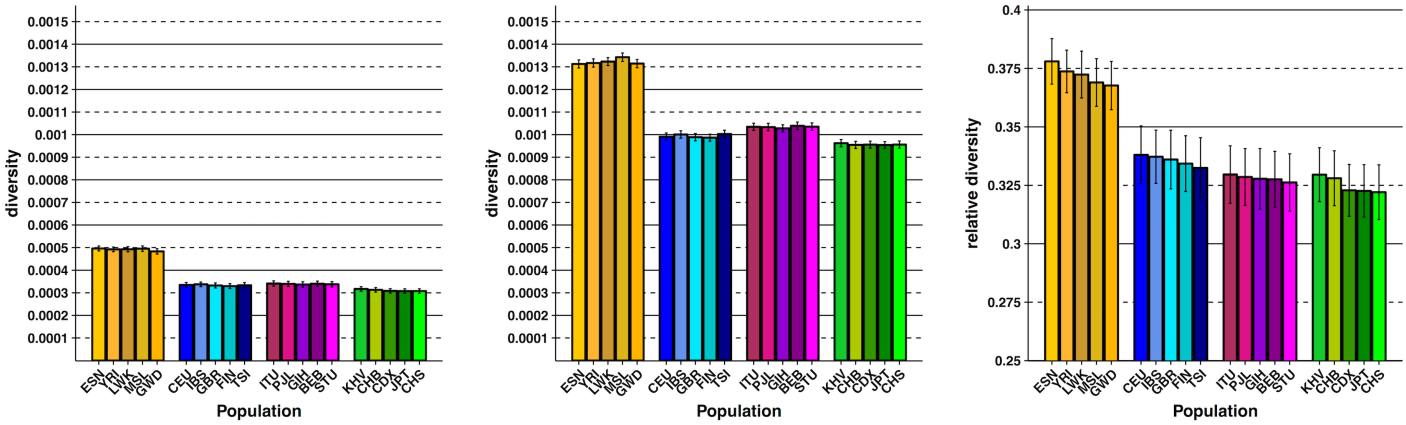

diversity
0.0015
0.0014
0.0013
0.0012
0.0011
0.001
0.0009
0.0008
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
ESN YRI LWK MSL GWD CEU IBS GBR FIN TSI ITU PJL GIH BEB STU KHV CHB CDX JPT CHS
Population
diversity
0.0015
0.0014
0.0013
0.0012
0.0011
0.001
0.0009
0.0008
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
ESN YRI LWK MSL GWD CEU IBS GBR FIN TSI ITU PJL GIH BEB STU KHV CHB CDX JPT CHS
Population
relative diversity
0.4
0.375
0.35
0.325
0.3
0.275
0.25
ESN YRI LWK MSL GWD CEU IBS GBR FIN TSI ITU PJL GIH BEB STU KHV CHB CDX JPT CHS
Population

Figure S4

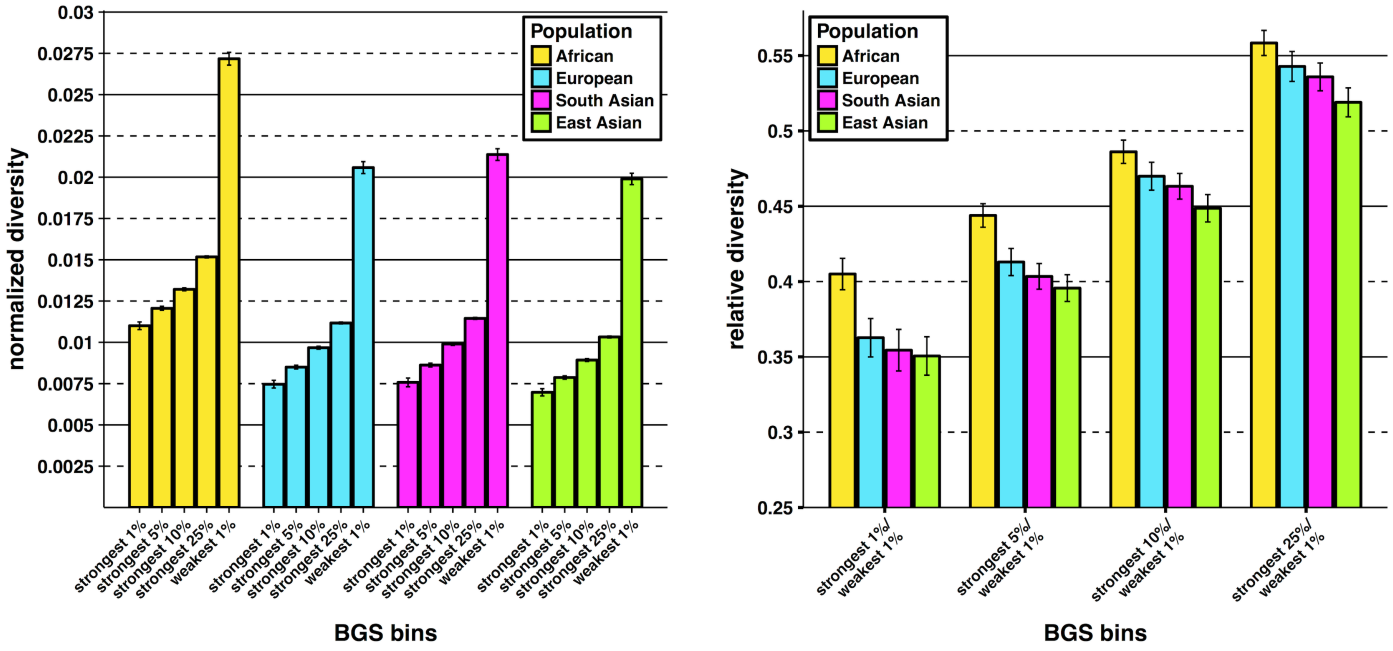

Population
African
European
South Asian
East Asian
normalized diversity
0.03
0.0275
0.025
0.0225
0.02
0.0175
0.015
0.0125
0.01
0.0075
0.005
0.0025
strongest 1%
strongest 5%
strongest 10%
strongest 25%
weakest 1%
BGS bins
relative diversity
0.55
0.5
0.45
0.4
0.35
0.3
0.25
strongest 1%/ weakest 1%
strongest 5%/ weakest 1%
strongest 10%/ weakest 1%
strongest 25%/ weakest 1%
BGS bins

Figure S5

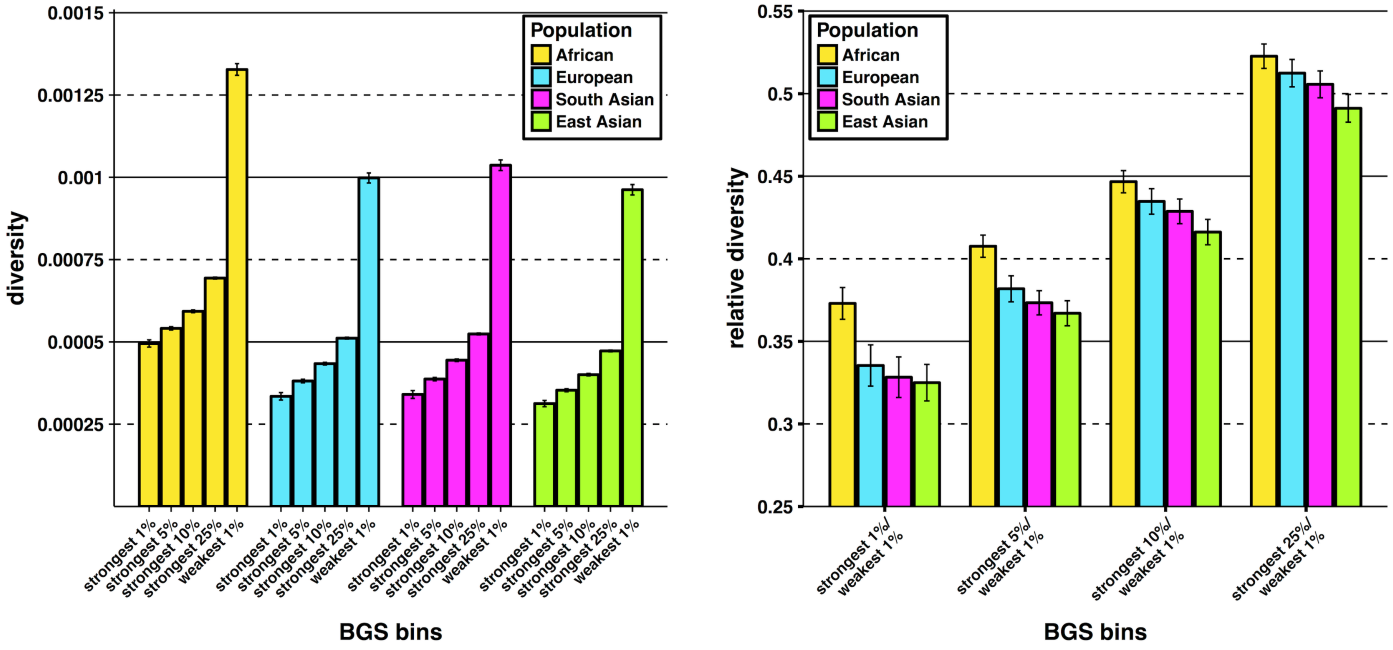

Population
African
European
South Asian
East Asian
diversity
0.0015
0.00125
0.001
0.00075
0.0005
0.00025
strongest 1%
strongest 5%
strongest 10%
strongest 25%
weakest 1%
BGS bins
relative diversity
0.55
0.5
0.45
0.4
0.35
0.3
0.25
strongest 1%/ weakest 1%
strongest 5%/ weakest 1%
strongest 10%/ weakest 1%
strongest 25%/ weakest 1%
BGS bins

Figure S6

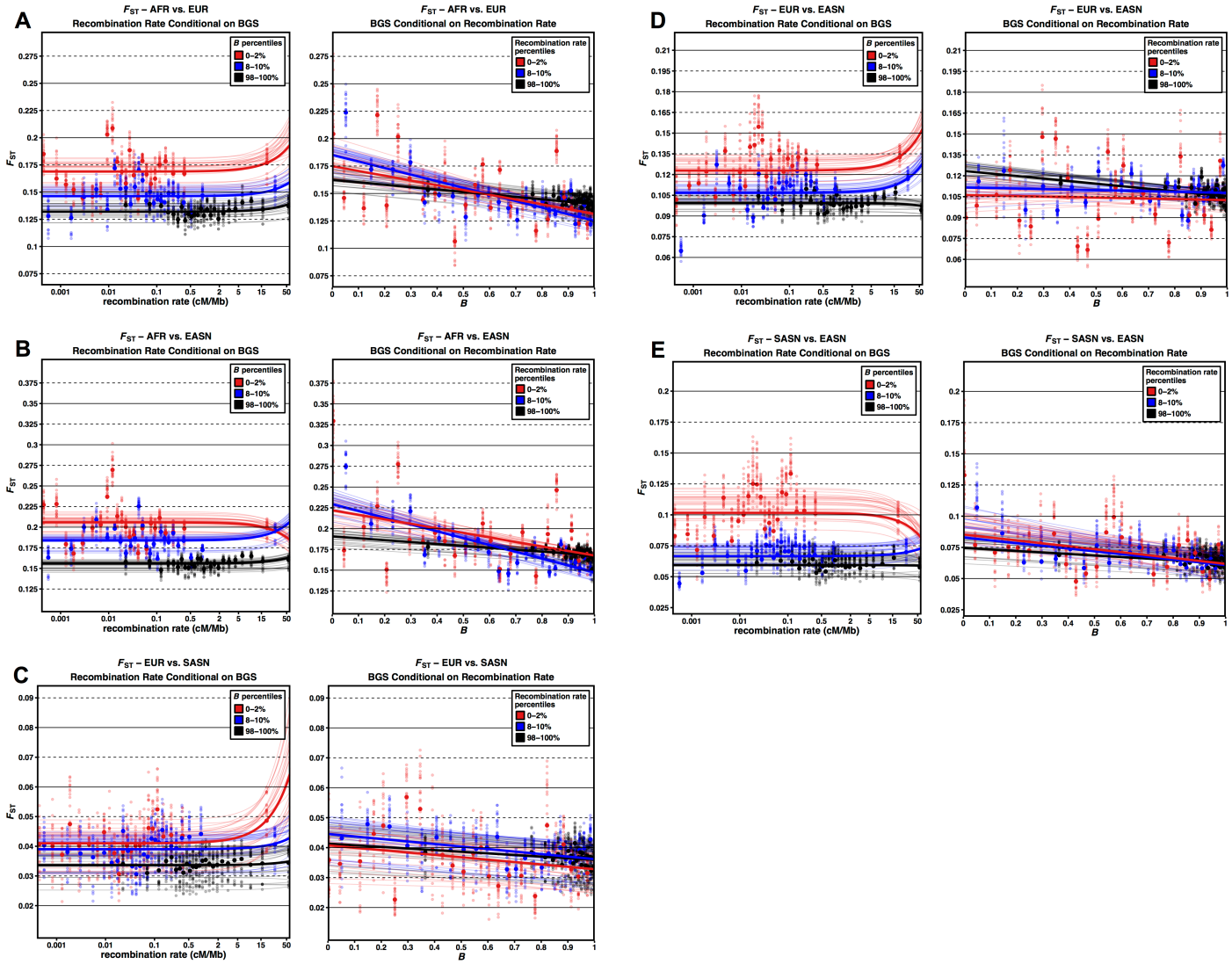

A
F_{ST} – AFR vs. EUR
Recombination Rate Conditional on BGS
F_{ST} – AFR vs. EUR
BGS Conditional on Recombination Rate
B
F_{ST} – AFR vs. EASN
Recombination Rate Conditional on BGS
F_{ST} – AFR vs. EASN
BGS Conditional on Recombination Rate
C
F_{ST} – EUR vs. SASN
Recombination Rate Conditional on BGS
F_{ST} – EUR vs. SASN
BGS Conditional on Recombination Rate
D
F_{ST} – EUR vs. EASN
Recombination Rate Conditional on BGS
F_{ST} – EUR vs. EASN
BGS Conditional on Recombination Rate
E
F_{ST} – SASN vs. EASN
Recombination Rate Conditional on BGS
F_{ST} – SASN vs. EASN
BGS Conditional on Recombination Rate
B percentiles
0–2%
8–10%
98–100%
Recombination rate percentiles
0–2%
8–10%
98–100%
F_{ST}
recombination rate (cM/Mb)
B

Figure S7

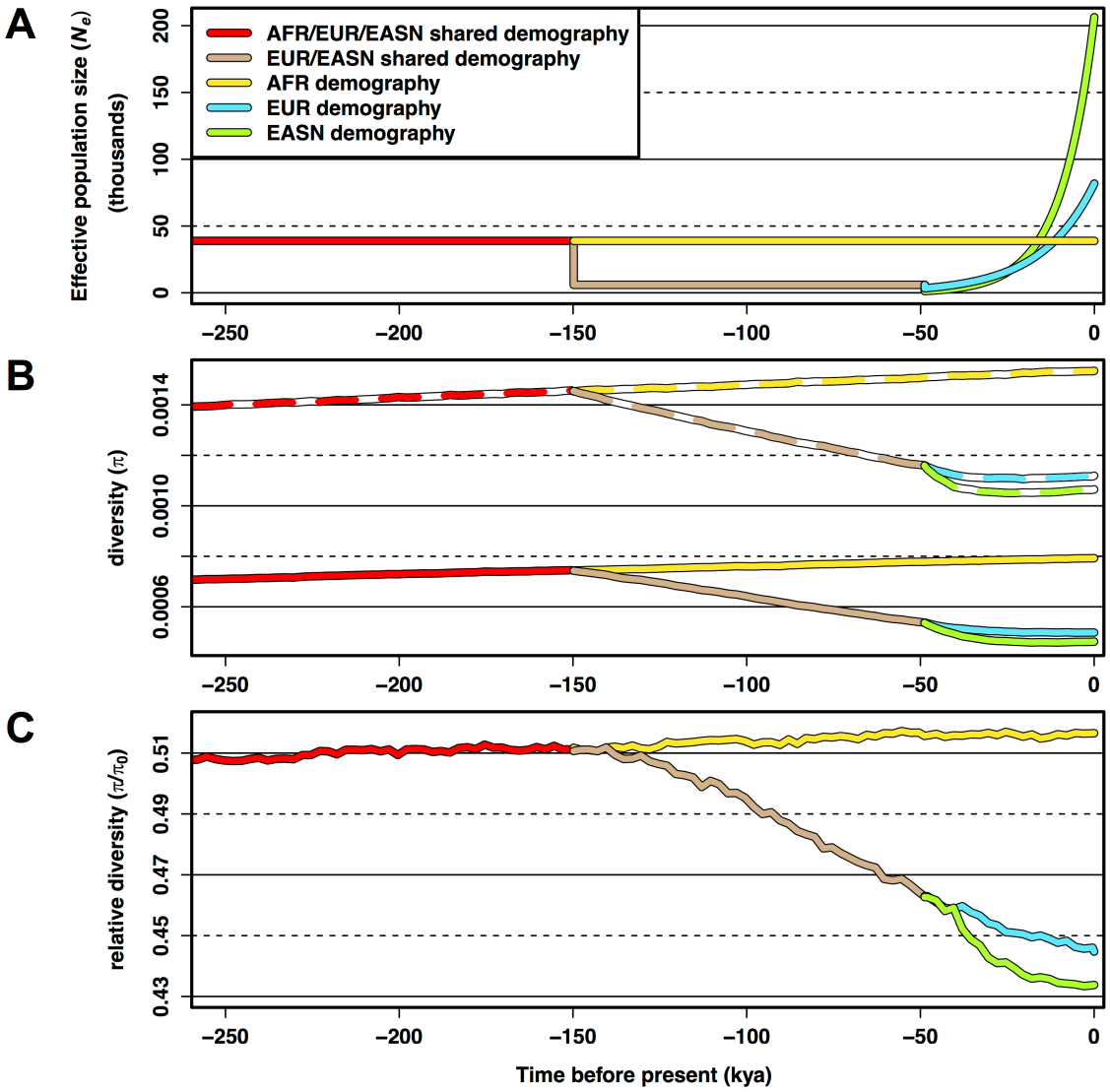
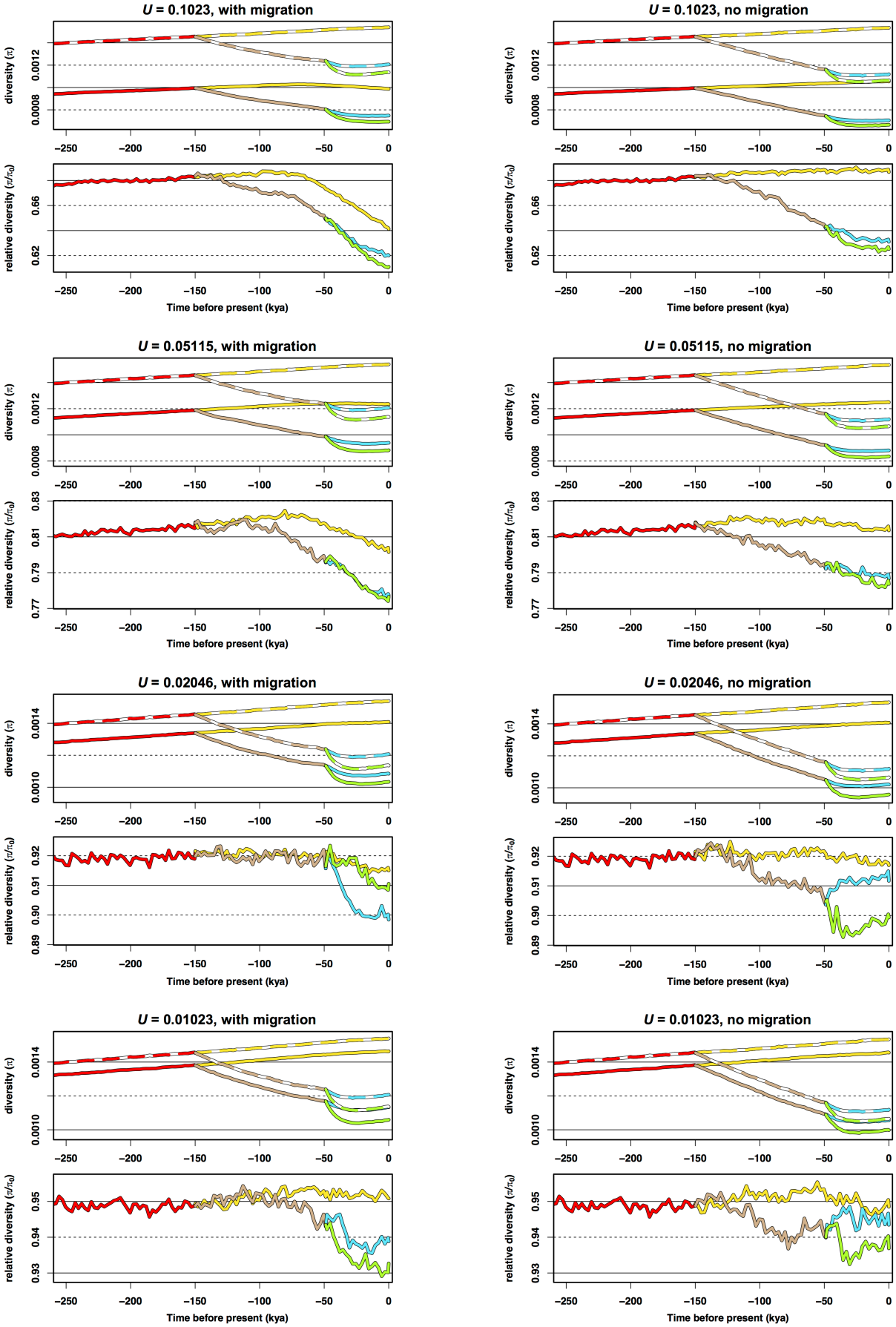

A
B
C
AFR/EUR/EASN shared demography
EUR/EASN shared demography
AFR demography
EUR demography
EASN demography
Effective population size (N_e)
(thousands)
0
50
100
150
200
diversity (π)
0.0006
0.0010
0.0014
relative diversity (π/π_0)
0.43
0.45
0.47
0.49
0.51
−250
−200
−150
−100
−50
0
Time before present (kya)

Figure S8


U = 0.1023, with migration
U = 0.1023, no migration
U = 0.05115, with migration
U = 0.05115, no migration
U = 0.02046, with migration
U = 0.02046, no migration
U = 0.01023, with migration
U = 0.01023, no migration
diversity (π)
relative diversity (π/π0)
Time before present (kya)
−250
−200
−150
−100
−50
0

S2 Appendix. Soft sweep detection and implementation in selscan v1.2.0.

Detecting soft sweeps

Under the model of a soft sweep, there is an increased chance of multiple distinct haplotypes sweeping to high frequency in a population. Garud et al. [1] developed a window-based statistic ($H12$) with good power to detect this process, and here we adapt $H12$ into an integrated haplotype homozygosity framework [2–4]. We call this new statistic $iHH12$. The general principle of these statistics is to combine the top two most frequent haplotypes into a single haplotype class to avoid the reduced power that iHS has when the adaptive allele segregates on more than one haplotype background. We calculate $iHH12$ as follows.

Following the notation of Szpiech and Hernandez [5] in a sample of n chromosomes we let $\mathcal{C}$ be the set of all possible distinct haplotypes at the locus x_0. $\mathcal{C}(x_i)$ is then the set of all possible distinct haplotypes extending from locus x_0 to locus x_i. Let h_i in $\mathcal{C}(x)$ be the ith most frequent haplotype. We then calculate $EHH12$ of the entire sample of haplotypes from x_0 to x_i as

$$EHH12(x_i) = \frac{\binom{n_{h_1}+n_{h_2}}{2}}{\binom{n}{2}} + \sum_{j>2}^{|\mathcal{C}(x_i)|} \frac{\binom{n_{h_j}}{2}}{\binom{n}{2}}$$

where n_{h_j} is the number of h_j haplotypes in the sample.

If $EHH12(x_i)$ is calculated repeatedly for several x_i moving farther away from x_0, we expect to observe more haplotypes and therefore we expect to observe lower haplotype homozygosity. However, the decay of homozygosity is slower in a region under selection [2–4]. Therefore, we integrate the decay of $EHH12$ as a function of genetic distance in order to summarize the pattern and make genome-wide comparisons. This integrated score is calculated as

$$iHH12 = \sum_{i=1}^{|\mathcal{D}|} \frac{1}{2}(EHH12(x_{i-1}) - EHH(x_i))g(x_{i-1}, x_i) + \sum_{i=1}^{|\mathcal{U}|} \frac{1}{2}(EHH12(x_{i-1}) - EHH(x_i))g(x_{i-1}, x_i)$$

where $g(x_{i-1}, x_i)$ is the genetic distance between markers x_{i-1} and x_i. $\mathcal{D}$ and $\mathcal{U}$ represent sets of markers downstream and upstream from x_0, respectively. In practice the curve is integrated until $EHH12 < 0.05$ on both sides of the focal locus. Finally, $iHH12$ is normalized genome-wide in order to account for the effects of demographic history on the distribution of haplotype homozygosity. We integrated this new statistical framework to detect soft-sweeps into `selscan` version 1.2.0 (https://github.com/szpiech/selscan) [5].

We evaluated the power of our $iHH12$ statistic implementation in `selscan` to detect hard and soft sweeps relative to iHS across a range of parameters. We simulated neutrally evolving sequences with `ms` [6] and non-neutrally evolving sequences with `mssel`, a modified version of `ms` also developed by R. Hudson

that conditions on an allele frequency trajectory. We simulated trajectories backwards in time under a selection on standing variation model with $s = 0.01$. Once an adaptive variant reached a set frequency backwards in time, the selection coefficient was set to $s = 0$ and was allowed drift neutrally until loss. We simulated 200 replicates across several sampling frequencies (0.7, 0.8, 0.9), several frequencies at which the variant become adaptive (0, 0.01, 0.02, 0.05, 0.10), and several demographic histories (Constant, African, European; [7]).

For both iHS and $iHH12$ scans, we normalized scores with respect to the neutral simulations and calculated the critical threshold for the most extreme 1% of scores. Using non-overlapping 100 kb windows across the genome, we calculated the fraction of scores in each window above this threshold. The top 1% of windows are identified as putatively under positive selection. This scheme controls the false positive rate to be no greater than 1%.

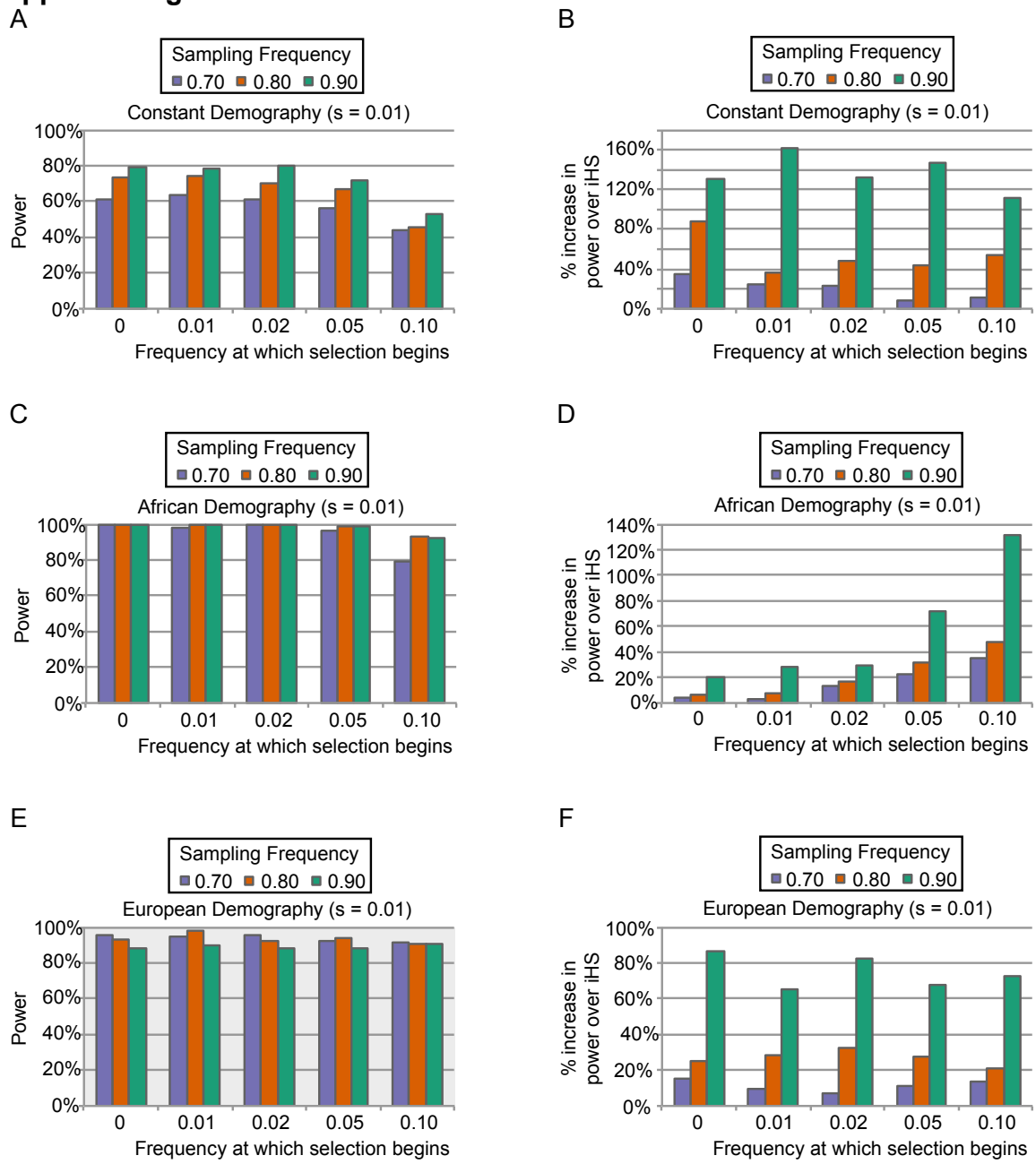
$iHH12$ has good power to detect hard and soft sweeps (Appendix Fig A1 A, C, E) and has improved power to identify both types of sweeps over iHS (Appendix Fig A1 B, D, F), particularly under realistic models of human demography.

Computing iHS and iHH12 scores in the Thousand Genomes Project (TGP)

We used `selscan` to compute both iHS and $iHH12$ scores for phase 3 TGP [8] phased whole genome sequences with a genetic map from HapMap3 [9]. Genetic map locations for sites not present in HapMap3 were linearly interpolated. The statistics were calculated for each population separately, and variants of frequency < 0.05 were filtered by `selscan`. All `selscan` runs used default parameters.

Using `selscan`'s companion program `norm`, for each population we normalized $iHH12$ scores genome-wide and normalized iHS scores in 1% frequency bins genome-wide. We identified the critical threshold representing the most extreme 1% of scores for each statistic. Then, to identify putative regions under selection, we partitioned the genome into non-overlapping 100 kb windows, and then we calculated the fraction of scores in each window above this threshold. The top 1% of windows were identified as putatively under positive selection. This scheme controlled the false positive rate to be no greater than 1%.

Appendix Figure A1



Appendix Fig A1. Power of $iHH12$ and comparison with iHS. Simulated power of $iHH12$ (A), (C), and (E) under varying parameters and comparison with iHS power (B), (D), and (F) in the same scenario. Panels (A) and (B) show results for a constant demography; panels (C) and (D) show results for an African demography; and panels (E) and (F) show results for a European demography. Non-constant demographies are from Gutenkunst et al. [7]. When the frequency at which selection begins is > 0, the sweep is considered soft. All simulations assume a selection coefficient of $s = 0.01$.

References

1. Garud NR, Messer PW, Buzbas EO, Petrov DA. Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps. Copenhaver GP, editor. PLoS Genet. Wiley; 2015;11: e1005004. doi:10.1371/journal.pgen.1005004
2. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. Nature Publishing Group; 2002;419: 832–837. doi:10.1038/nature01027.1.
3. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. Hurst L, editor. PLoS Biol. Public Library of Science; 2006;4: 0446–0458. doi:10.1371/journal.pbio.0040072
4. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. Nature. Nature Publishing Group; 2007;449: 913–8. doi:10.1038/nature06250
5. Szpiech ZA, Hernandez RD. Selscan: An efficient multithreaded program to perform EHH-based scans for positive selection. Mol Biol Evol. 2014;31: 2824–2827. doi:10.1093/molbev/msu211
6. Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. Oxford University Press; 2002;18: 337–338. doi:10.1093/bioinformatics/18.2.337
7. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 2009;5: e1000695. doi:10.1371/journal.pgen.1000695
8. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015;526: 68–74. doi:10.1038/nature15393
9. The International HapMap 3 Consortium, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467: 52–8. doi:10.1038/nature09298