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 consequences 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 attribute to their non-equilibrium demographics. We confirm that the interplay of demography and BGS can explain the differences of relative diversity across human populations with simulations incorporating a parametric demographic model of human history [7] with and without a model of BGS. Specifically, our simulations of non-African human demography show that the out-of-Africa population bottleneck and European-Asian split population bottleneck each coincide with a distinct drop in relative diversity. Furthermore, we demonstrate that back migration from Europeans and Asians into Africa re-introduces sufficient deleterious variation to impact BGS strength, leading to decreased relative diversity in Africans. These results demonstrate the strong impact that changing demography has on perturbing levels of diversity in regions under BGS and have implications for population genetic studies seeking to characterize linked selection across any species or population that is not at demographic equilibrium.

Results

Differential impact of BGS across human populations

We measured mean pairwise genetic diversity (π) among the 20 non-admixed populations from the phase 3 TGP data set, consisting of 5 populations each from 4 continental groups: Africa (AFR), Europe (EUR), South Asia (SASN), and East Asia (EASN) (Table S11 in Supporting information for population labels and grouping). After a set of stringent filters were applied to all 20 populations to identify a high-quality set of putatively neutral sites in the genome (see Materials and Methods), sites were divided into quantile bins based on their BGS coefficient (termed "*B*"), which represents the inferred strength of BGS from McVicker et al. [6] (see Materials and Methods). For our initial set of analyses, we focused on the bins corresponding to the 1% of sites under strongest BGS (or lowest *B*) and the 1% of sites under weakest BGS (or highest *B*). Mean diversity (normalized by divergence with rhesus macaque) within these bins for each population is shown in Figs 1A-B. As expected, normalized diversity was highest in African populations and lowest in East Asian populations across both 1% *B* quantile bins.

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

To characterize these effects across a broader distribution of BGS strengths, we grouped populations together according to their continental group (i.e., African, European, South Asian, and East Asian, see Table S11 in Supporting information for a detailed description) and estimated relative diversity at neutral sites for each of the continental groups in bins corresponding to the bottom 1%, 5%, 10%, and 25% quantiles of *B* (note these partitions were not disjoint). As expected, relative diversity increased for all continental groups as the bins became more inclusive (Fig 2B), reflecting a reduced impact on the reduction of diversity due to BGS. We also observed that non-African continental groups consistently had a lower relative diversity compared to African groups, demonstrating that the patterns we observed in the most extreme BGS regions also held for weaker BGS regions. Interestingly, we observed a consistent trend of rank order for relative diversity between the different continental groups for each quantile bin, with the East Asian group experiencing the greatest reduction of relative diversity, followed by the South Asian, European, and African groups. This result suggested a stronger effect for demography on the diversity-reducing effect of BGS for those populations experiencing the strongest bottlenecks. However, the observed differences in relative diversity between non-African and African continental groups became less pronounced as the bins became more inclusive (Fig 2B). These effects remained even after we controlled for the effects of GC-biased gene conversion and recombination hotspots (Figs S2 and S4) or if we did not normalize diversity by divergence (Figs S3 and S5).

Recent admixture has not altered the impact of BGS

We then investigated whether the effects of BGS have remained consistent across human populations that have experienced recent admixture. To do so, we tested for the effect of BGS on normalized and relative diversity in the 6 admixed TGP populations (ASW, ACB, CLM, MXL, PEL, and PUR). We first used the local ancestry tracks inferred by the phase 3 TGP consortium to divide up admixed samples into genomic segments that are homozygous for a specific local ancestry (i.e., African, European, or Native American). These segments were then collated across all admixed samples (see Materials and Methods), irrespective of their source population. We then calculated normalized and relative heterozygosity across the collated segments as a function of the BGS quantile bins described in the previous section. For comparison, heterozygosity was also calculated for each TGP continental group. Across all BGS quantile bins, normalized diversity (heterozygosity/divergence) in African and European ancestry segments closely matched the values observed in their non-admixed counterparts (Fig 3A). However, normalized diversity was significantly lower in the Native American ancestry segments of admixed individuals than in the East Asian continental group (Fig 3A). This was expected given the more recent divergence of Native American populations and the strong population bottleneck they experienced migrating into the Americas [18,54,55].

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

BGS has shaped patterns of population differentiation

To test if greater genetic drift in regions of strong BGS has contributed to the observed 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) and repeated all analyses. Our choice in total number of bins resulted in a minimum of 320 SNPs per bin for estimating F_{ST} between any two populations, which should be sufficient to avoid errors when estimating F_{ST} across multiple loci [58]. After performing multiple linear regression of F_{ST} on B, recombination rate (ρ), and an interaction term 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 observed that B was a statistically significant predictor ($p < 1e\text{-}04$) for F_{ST} across all population comparisons regardless of how we partitioned sites (Table S8 in Supporting information). This result remained unchanged when performing robust regression. In contrast, recombination rate exhibited sporadic significance as an explanatory variable for F_{ST} across population comparisons and was dependent upon how sites were partitioned (i.e., whether we first partitioned by B or by recombination rate) (Table S8 in Supporting information). Furthermore, strong differences between the two binning schemes were observed for the magnitude of the recombination rate regression coefficient for certain population comparisons (e.g., African vs. East Asian and South Asian vs. East Asian), while the coefficients for B were consistent across binning schemes. The direction in which recombination rate explained F_{ST} was also inconsistent across different population comparisons, with European vs. South Asian and European vs. East Asian comparisons showing a significant positive change in F_{ST} as a function of increasing recombination rate. This result was contrary to an expectation of decreasing F_{ST} as a function of increasing recombination rate [57]. We also failed to observe consistent effects from the interaction term for B and recombination rate on F_{ST} across population comparisons or binning schemes (Table S8 in Supporting information). Performing robust regression on the model did not change these results. However, in contrast to recombination rate, when the model was performed utilizing all TGP populations (i.e., the "Global" estimate), the interaction term was significant in explaining F_{ST} across both types of binning schemes.

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

Demographic inference in putatively neutral regions of the genome

One consequence of BGS driving patterns of neutral variation within and between human populations is that demographic inference could be substantially biased [59,60]. To assess the degree of bias, we inferred a 13-parameter demographic model of African, European, and East Asian demography using only putatively neutral regions of the genome under the weakest effects of BGS ($B \geq 0.994$) from a subset of TGP individuals with high coverage whole genome sequence data (see Materials and Methods). Our demographic model followed that of Gutenkunst et al. [7], with an ancient human expansion in Africa and a single out-of-Africa bottleneck followed by European- and East Asian-specific bottlenecks, as well as exponential growth in both non-African populations and migration between all populations. To make comparisons to previous studies that have used sequence data from coding regions or genes [7,61,62], which may be under strong BGS, we also inferred demographic parameters using coding four-fold degenerate synonymous sites. Our inferred parameters for human demography were strikingly different between the two sets of sequence data (Fig S1, Table S1 in Supporting information). Notably, inferred effective population size parameters were larger for contemporary population sizes when using four-fold degenerate synonymous sites versus weak BGS neutral regions, with N_e inferred to be 22%, 23%, and 29% larger for AFR, EUR, and EASN populations, respectively. This is despite the fact that the ancestral N_e was inferred to be lower for four-fold degenerate synonymous sites (N_e = 18,449 and 17,118, for weak BGS neutral regions and four-fold degenerate sites, respectively). This result may stem from the expected decrease in N_e going into the past in regions of BGS, which can lead to inflated estimates of recent population growth [60] and has also been shown in simulation studies of synonymous sites under BGS [59].

Simulations confirm that demographic effects can impact background selection

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

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

Discussion

In our analyses of thousands of genomes from globally distributed human populations, we have confirmed that neutral variation in humans is governed largely by the processes of both linked selection and demography. While this observation is not unexpected, we have characterized for the first time the dynamic consequences of non-equilibrium demographic processes in regions of BGS. We find that demography's effect in these regions do not result in the same relative magnitude of decreased neutral diversity across populations. Instead, relative diversity in regions of BGS is highly dependent on a population's demographic history, with bottlenecks playing a particularly strong role. For example, the strongest overall decreases in relative diversity occur in the East Asian continental group, which is expected if they also suffered the strongest population bottlenecks. To remove any possible biases that would influence our results, we controlled for functional effects of mutations, variability in mutation along the genome, potential sequencing artifacts, GC-based gene conversion, and the potential mutagenic effects of recombination hotspots. None of these factors qualitatively affected our results.

We do recognize that one caveat of our controls is the fact that divergence itself is not independent of BGS [63], and this may present biases when using divergence to control for variation in mutation rate along the genome. This is because the rate of coalescence in the ancestral population of two groups will be faster in regions of strong BGS compared to regions of weak BGS due to the lower N_e of the former, leading to a decrease in overall divergence in those regions. To limit the contribution of such biases in ancestral N_e to divergence, we use rhesus macaque since it is more distantly related to humans than other primate species such as orangutan or chimpanzee (human-rhesus divergence: 29.6 MYA; [64]). However, biases in divergence due to BGS can be still be observed between species as distantly related as human and mouse [63] (human-mouse divergence: 75 MYA; [65]). Despite this, the apparent correlation between BGS and divergence should not qualitatively affect our results of relative diversity. Rather, it is likely that differences in normalized diversity between different BGS bins are greater than what we observe here. Normalizing by the lower divergence that is characteristic of strong BGS bins and the higher divergence that is characteristic of weak BGS bins should make any differences between the two smaller, not greater. In fact, for our calculations of relative diversity in which we skip the normalization step, the differences in diversity between the strongest 1% and weakest 1% BGS bins are greater and give a lower ratio of relative diversity (π/π_0 for AFR is 0.373 without the divergence step and 0.402 with the divergence step). A similar pattern is also observed for other continental groups (compare Fig 2 and Fig S5). More importantly though, we should not expect the potential biases of our divergence step to contribute to the differences in relative diversity between each of the continental groups since biases in divergence across the genome should impact all human populations equally.

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

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

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

While we focus specifically on the effects of BGS on generating local patterns of diversity across humans, there is still a possibility that the effects of positive selection, through recurrent-hitchhiking, soft-sweeps, and classic selective sweeps, are contributing to these patterns as well. Models taking into account both the effects of purifying selection and positive selection will be necessary to provide a full description of the impact of linked selection across the human genome and its behavior as a function of different population demographic histories [42]. Development of mathematical models characterizing the joint effects of both BGS and genetic hitchhiking [24] as well as their recent application to different species [30,44], gives promise towards this goal. Recently, a joint model of classic selective sweeps and BGS was applied to *Drosophila* and predicted that BGS has had a ~1.6 to 2.5-fold greater effect on neutral genetic diversity than classic selective sweeps [30]. We should expect this magnitude to be even greater for humans, since classic selective sweeps were found to be rare in recent human evolution [38] and adaptive substitutions in the human genome are much less frequent than *Drosophila* [5,70,71]. Despite this, we still attempt to alleviate any contribution from other modes of linked selection by masking out regions of hard-sweeps or soft-sweeps using a novel H12-like method [72] that uses an integrated haplotype homozygosity framework [36] to scan for population specific signatures of these events in the genome ([73]; S2 Appendix). While there may still be some residual contribution of these events to our estimates of π/π_0, if such signatures do exist, they are likely to localize where BGS is also expected to act. Thus, it is possible that our empirical results describe the consequences of non-equilibrium demography in the broader context of both positive and purifying modes of linked selection. Indeed, our simulations of BGS fail to capture the complete effects of linked selection on reducing π/π_0 in different human populations (compare Figs 2B and 6C) and the additional contribution of hitchhiking, which we did not simulate, may explain this discrepancy.

Non-equilibrium demography has also been been recently investigated with regard to its impact on patterns of deleterious variation across human populations. Initial work comparing non-synonymous and synonymous sites in European-Americans and African-Americans found that, while African-Americans have a greater number of segregating non-synoymous and synonymous sites than European-Americans, European-Americans have an overall greater amount of non-synonymous and synonymous homozygous derived alleles [50]. Other work showed similar findings [62,74], with Henn et al. [75] demonstrating that the number of deleterious derived alleles per genome increased with distance from Africa. In addition, similar results in human founder populations [76,77], *Arabadopsis* [78], and domesticated species such as dogs [12] and sunflowers [79] further demonstrate the pervasive impact that demography has on influencing the relative amount of deleterious variation across a variety of populations and species. Since BGS itself is a function of deleterious variation, it is not surprising that we also witness differences in π/π_0 across human populations that have experienced different demographic histories. These effects are likley ubiquitous across other species as well. However, there has been recent contention about whether the previously described patterns of increased deleterious variation are driven by a decrease in the efficacy of natural selection or are solely artifacts of the response of deleterious variation to demographic change [51,52,80–82]. Following a strong population bottleneck, a transient increase in the relative ratio of deleterious variation to neutral variation is expected, that is independent of the efficacy of selection [47,48]. Since our own investigation focuses solely on neutral variants, we can only invoke differences in population demography and drift on causing the differences between populations, supporting the conjecture that demographic processes may transiently dominate patterns of diversity at sites that may be under the direct influence (or in the case of BGS, indirect influence) of purifying selection.

Recently, Koch et al. [49] investigated the temporal dynamics of human demography on selected sites and observed that after a population contraction, heterozygosity at selected sites can undershoot its expected value at equilibrium as low-frequency variants are lost at a quicker rate before the recovery of intermediate frequency variants can occur. In the context of BGS, which skews the site frequency spectrum of linked neutral mutations towards rare variants [23,83], we also expect a transient decrease in diversity as low-frequency variants are lost quickly during a population contraction. This effect may be driving the results of the greater losses of relative diversity in non-African populations, but further work is needed to better understand the impact of population contractions on perturbing the site-frequency spectrum of regions under BGS. Koch et al. demonstrate that this effect is only temporary and that long-term diversity at selected sites approaches greater values once equilibrium is reached. We stress that these temporal effects also impact the patterns of neutral diversity caused by BGS and likely explain the differences of π/π_0 that we observed in humans. We predict that as evolution continues forward through time, the magnitude and rank order of relative diversity among populations may change. However, this will be highly dependent on the demographic effects that different populations experience. Although more flexible modeling frameworks that can account for BGS under changing population size have been recently developed [45,46], they assume that selection is strong enough that the impact of demography does not perturb mutation-selection balance, and they also ignore the additional impact that demography itself has on perturbing neutral variants.

A greater contemporary N_e in non-Africans could theoretically result in a greater efficacy of purifying selection and, consequently, a stronger impact of BGS. However, it is very unlikely that this is occuring or driving the observed patterns of relative diversity that we see. The greater contemporary population size of non-Africans has transpired only in the very recent past, with accelerated growth in Europeans occurring within the last few hundred generations [62,84–87]. Thus, most population specific mutations in non-Africans have arisen only during this recent population expansion and are both young and very rare [88,89]. It is doubtful that these population-specific variants have had enough time to exercise a large effect on the absolute strength of BGS itself. Rather, our simulations indicate that the response of π in regions under BGS is driven by population contractions, with reductions in π/π_0 occurring concomitantly with the out-of-Africa bottleneck and European-East Asian split bottleneck events (Fig 6) and continuing even after the European and East Asian expansion events. Interestingly though, our simulations reveal an additional factor that can influence the impact of BGS within populations – migration between populations. We observe that the exchange of deleterious variants between populations can impact linked selection and lead to decreases in π/π_0 in populations of constant size, such as the recent history inferred for Africans (Fig 6B). In particular, a decrease in π/π_0 for Africans coincides with decreases in π/π_0 for non-Africans. This trend disappears, though, once migration is removed (Fig S7). This result suggests that the migration of nearly neutral or mildly deleterious alleles in Europeans/East Asians into Africans results in those variants being more strongly selected in Africans due to their historically larger N_e, resulting in stronger BGS and lower π. However, more work is needed to definitively test this. Conversely, migration of African alleles into Europeans/East Asians has no noticeable effect on π in our simulations with BGS (compare solid lines in Fig 6B and Fig S7B). In simulations where U is lower than our base level of 0.2046, we also witness differences in π/π_0 between simulations with and without migration. Some of these involve rank order changes in π/π_0 between Africans, Europeans, and East Asians (Fig S8, U = 0.02046 with migration vs. no migration), which is likely a consequence of the increased stochasticity that migration between populations introduces, as simulations without migration restored the consistent patterns of π/π_0. These results demonstrate the complexities that migration introduces to patterns of diversity in regions of linked selection and to the dynamic changes of π/π_0 over time.

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

Since linked selection is a pervasive force in shaping patterns of diversity across the genomes in a range of biological species [1], it has been provided as an argument for why neutral diversity and estimates of N_e are relatively constrained across species in spite of the large variance in census population sizes that exist [44,97]. However, since population bottlenecks are common among species and have an inordinate influence on N_e [20], demography has also been argued as a major culprit for constrained diversity [2,97–99]. Yet, as we show in humans, it is likely that patterns of neutral diversity are in fact jointly impacted by both of these forces, magnifying one another to deplete levels of diversity beyond what is expected by either one independently. In our work, we also identify a potentially substantial role for migration from smaller populations that harbor more strongly deleterious alleles on patterns of linked neutral diversity in large populations. Together, these combined effects may help provide additional clues for the puzzling lack of disparity in genetic diversity among different species [100].

Our results give caution that not properly accounting for the impacts of demography on patterns driven by linked selection may lead to spurious rejections of certain models of linked selection in favor of others. For example, when comparing humans to other primates, a recent study by Nam et al. [101] found a greater relative decrease of neutral genetic diversity in regions close to genes (where linked selection is greatest) versus regions far away from genes in non-human primate species. The authors suggest that these patterns are best described by the action of selective sweeps since non-human primate populations usually exhibit a larger N_e and thus should experience a greater influx of adaptive alleles per unit time. However, certain orangutan, gorilla, and chimpanzee species have also experienced large population contractions in their recent history that are comparable in time and potentially more extreme in scale to what humans have also experienced [102–105]. Thus, the consequences of strong demographic change to regions of BGS can provide an alternative explanation for why these primate species exhibit a larger relative reduction in neutral diversity near genes, despite their larger historical N_e. While Nam et al. also test for BGS using simulations with demography, their model assumed a more limited population contraction (50% population reduction) over a shorter time scale than the demographic model we test.

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

Materials and methods

Data

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

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

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

Filtering and ascertainment scheme

Positions in the genome were annotated for background selection by using the background selection coefficient, *B*, which was inferred by McVicker et al. [6] and downloaded from http://www.phrap.org/othersoftware.html. *B* was inferred by applying a classical model of BGS [53], which treats its effects as a simple reduction in N_e at neutral sites as a function of their recombination distance from conserved loci, the strength of purifying selection at those conserved loci, and the deleterious mutation rate. *B* can be interpreted as the reduced fraction of neutral genetic diversity at a particular site along the genome that is caused by BGS, with a value of 0 indicating a near complete removal of neutral genetic diversity due to BGS and a *B* value of 1 indicating little to no effect of BGS on neutral genetic diversity ($B = \pi/\pi_0 = N_e/N_0$). Positions for *B* were lifted over from hg18 to hg19 using the UCSC liftOver tool. Sites that failed to uniquely map from hg18 to hg19 or failed to uniquely map in the reciprocal direction were excluded. Sites lacking a *B* value were also ignored. We focused our analyses on those regions of the genome within the top 1%, 5%, 10%, and 25% of the genome-wide distribution of *B* and within the bottom 1% of the genome-wide distribution of *B*. These quantiles correspond to the *B* values 0.095, 0.317, 0.463, 0.691, and 0.994, respectively.

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

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

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

2. phyloP: Sites with phyloP [111] scores > 1.2 or < -1.2 were removed to limit the effects of natural selection due to conservation or accelerated evolution. Scores were downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phyloP46way/.
3. phastCons: Regions in the UCSC conservation 46-way track (table: phastCons46wayPlacental) [112] were removed to limit the effects of natural selection due to conservation.
4. CpG: CpG islands in the UCSC CpG islands track were removed because of their potential role in gene regulation and/or being conserved.
5. ENCODE blacklist: Regions with high signal artifacts from next-generation sequencing experiments discovered during the ENCODE project [113] were removed.
6. Accessible genome mask: Regions not accessible to next-generation sequencing using short reads, according to the phase 3 TGP "strict" criteria, were removed (downloaded from ftp://ftp.ncbi.nlm.nih.gov/1000genomes/ftp/release/20130502/supporting/accessible_genome_masks/StrictMask/).
7. Simple repeats: Regions in the UCSC simple repeats track were removed due to potential misalignments with outgroups and/or being under natural se-

lection.

8. Gaps/centromeres/telomeres: Regions in the UCSC gap track were removed, including centromeres and telomeres.
9. Segmental duplications: Regions in the UCSC segmental dups track [114] were removed to limit potential effects of natural selection and/or misalignments with rhesus macaque.
10. Transposons: Active transposons (HERVK retrotransposons, the AluY subfamily of Alu elements, SVA elements, and L1Ta/L1pre-Ta LINEs) in the human genome were removed.
11. Recent positive selection: Regions inferred to be under hard and soft selective sweeps (using iHS and iHH12 regions from selscan v1.2.0 [73]; S2 Appendix) within each phase 3 population were removed.
12. Non-coding transcripts: Non-coding transcripts from the UCSC genes track were removed to limit potential effects of natural selection.
13. Synteny: Regions that did not share conserved synteny with rhesus macaque (rheMac2) from UCSC syntenic net filtering were removed (downloaded from http://hgdownload.soe.ucsc.edu/goldenPath/hg19/vsRheMac2/syntenicNet/).

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

14. GC-biased gene conversion (gBGC): Regions in UCSC phastBias track [115] from UCSC genome browser were removed to limit regions inferred to be under strong GC-biased gene conversion.
15. Recombination hotspots: All sites within 1.5 kb (i.e., 3 kb windows) of sites

with recombination rates ≥ 10 cM/Mb in the 1000G OMNI genetic maps for non-admixed populations (downloaded from ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/working/20130507_omni_recombination_rates/) and the HapMap II genetic map [116] were removed. 1.5 kb flanking regions surrounding the center of hotspots identified by Ref. [117] (downloaded from http://science.sciencemag.org/content/sci/suppl/2014/11/12/346.6211.1256442.DC1/1256442_DatafileS1.txt) were also removed, except for the cases in which the entire hotspot site was greater than 3 kb in length (in which case just the hotspot was removed).

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

Demographic inference

The inference tool dadi (v1.6.3) [7] was used to fit, via maximum likelihood, the 3-population 13-parameter demographic model of Gutenkunst et al. [7] to the 179 YRI, CEU, and CHS samples from the high coverage CG dataset of TGP. This sample set consisted of 53 YRI (African), 64 CEU (European), and 62 CHS (East Asian) samples. The demographic model incorporates an ancient human expansion in Africa and a single out-of-Africa bottleneck followed by European- and East Asian-specific bottlenecks, as well as exponential growth in both non-African populations and migration between populations. During the inference procedure, each population was projected down to 106 chromosomes, corresponding to the maximum number of chromosomes available in the CG YRI population. Sites were polarized with chimpanzee to identify putative ancestral/derived alleles using the chain and netted alignments of hg19 with panTro4 (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/vsPanTro4/axtNet/), and the correction for ancestral misidentification [119] option in dadi was used. The 13-filter set described previously was applied to the CG data set, and an additional filter keeping only the weakest ~1% of autosomal sites under background selection ($B \geq 0.994$) was also applied in order to mitigate potential biases in inference due to BGS [59,60] or other forms of linked selection [120]. After site filtering and correction for ancestral misidentification, a total of 110,582 segregating sites were utilized by dadi for the inference procedure. For optimization, grid points of 120, 130, and 140 were used, and 15 independent optimization runs were conducted from different initial parameter points to ensure convergence upon a global optimum. An effective sequence length (L) of 7.15 Mb was calculated from the input sequence data after accounting for the fraction of total sites removed due to filtering. In addition to the 13-filter set, this filtering included sites violating HWE, sites without B value information, sites that did not have at least 106 sampled chromosomes in each population, sites with more than two alleles, sites that did not have tri-nucleotide information for the correction for ancestral misidentification step, and sites treated as missing data. For calculating the reference effective population size, a mutation rate (μ) of 1.66 x 10^{-8} (inferred from Ref. [121]) was used. Using the optimized θ from dadi after parameter fitting, the equation $\theta = 4N_e\mu L$ was solved for N_e to generate the reference effective population size, from which all other population N_e's were calculated. This same procedure was also used to infer demographic parameters from four-fold degenerate synonymous sites across the same set of samples. After site filtering (note that *B* and the 13-filter set were not included in the filtering step for four-fold degenerate synonymous sites), 41,260 segregating sites were utilized by dadi for the inference procedure, and an effective sequence length of 2.37 Mb was used for calculating the reference effective population size.

Simulations

Forward simulations incorporating the results from the demographic inference procedure described above and a model of background selection were conducted using SFS_CODE [122]. For the model of background selection, the recombination rate, ρ, and the fraction of the genome experiencing deleterious mutation, *U*, was calculated using the 2 Mb region of chr3: 48,600,000-50,600,000, which has been subject to the strongest amount of BGS in the human genome (mean *B* = 0.002). A population-scaled recombination rate (ρ) of 6.0443 x 10^{-5} was calculated for this region using the HapMap II GRCh37 genetic map [116]. For generating *U*, the number of non-coding "functional" sites in this region was first calculated by taking the union of all phastCons sites and phyloP sites with scores > 1.2 (indicating conservation) that did not intersect with any coding exons. This amount totaled to 270,348 base pairs. Additionally, the number of coding sites was calculated by summing all coding exons within this region from GENCODE v19, which totaled to 138,923 base pairs. From these totals, a *U* of 0.2046 was generated.

The background selection model was simulated using a middle 30 kb neutral region flanked by two 1 Mb regions under purifying selection. From the calculated *U* described above, 20.46% of sites in the two 1 Mb flanking regions were simulated as being deleterious. Two distributions of fitness effects were used for the deleterious sites, with 66.06% of deleterious sites using the gamma distribution of fitness effects inferred across conserved non-coding regions by Ref. [123] (β = 0.0415, α = 0.00515625) and 33.94% of deleterious sites using the gamma distribution of fitness effects inferred across coding regions by Ref. [5] (β = 0.184, α = 0.00040244). The relative number of sites receiving each distribution of fitness effects in our simulations was determined by the relative number of non-coding "functional" sites and coding exons described above. Gamma distribution parameters are scaled to the ancestral population size of the demographic models used in Refs. [5,123]. To simulate varying levels of background selection strength, different total fractions of our calculated *U* were used (5%, 10%, 25%, 50%, and 100% of 0.2046) but with the same relative percentage of non-coding and coding sites just described. To simulate only the effects of demography without background selection (i.e., U = 0), only the 30 kb neutral region was simulated. 2,000 independent simulations were conducted for each particular *U* (12,000 total). Simulations output population genetic information every 100 generations and also at each generation experiencing a population size change (22,117 total generations were simulated), for which mean pairwise nucleotide diversity (π) was calculated across the 2,000 simulations.

Calculating recombination rate across the genome

We used recombination rate estimates from the HapMap II GRCh37 genetic map. To annotate sites in phase 3 that were not in HapMap II, recombination rates were interpolated to the midway point between the preceding and following positions in HapMap II. If the difference between successive HapMap II positions was greater than 18,848 base pairs (the first standard deviation for the distribution of distances between positions in HapMap II), then the recombination rate was only extended out 9,424 base pairs beyond the focal position. Positions beyond this distance were then ignored during analysis in which the recombination rate was used. Recombination rate quantiles were calculated using the genome-wide distribution of recombination rates (i.e., the distribution of recombination rates across all sites, including those that are not polymorphic in the data set) resulting from the procedure described above.

Population-specific calculations of diversity

Mean pairwise genetic diversity (π) was calculated as a function of the *B* quantile bins described in “Filtering and ascertainment scheme” for each of the 20 non-admixed populations in phase 3 TGP and across 4 broad populations that grouped the 20 non-admixed populations together by continent (African, European, South Asian, and East Asian, see Table S11 in Supporting information). Additionally, only regions of the genome passing the 13-filter set were used in the diversity calculations (see Table S4 in Supporting information for total number of Mb used in diversity calculations for each *B* quantile). For estimates of diversity controlling for gBGC or recombination hotspots, the additional corresponding filters described in “Filtering and ascertainment scheme” were also used. Only 100 kb regions of the genome with at least 10 kb of divergence information with Rhesus macaque were used in the diversity calculations (see "Normalization of diversity and divergence calculations with Rhesus macaque" below).

Ancestry specific calculations of diversity

To calculate genetic diversity as a function of local ancestry across the admixed phase 3 population samples, we used the ancestry deconvolution results generated by the 1000 Genomes Project Consortium (see ftp://ftp.ncbi.nlm.nih.gov/1000genomes/ftp/technical/working/20140818_ancestry_deconvolution/README_20140721_phase3_ancestry_deconvolution). Briefly, the local-ancestral inference tool, RFMix [124], was run across the ACB, ASW, CLM, MXL, PEL, and PUR phase 3 TGP samples. For the reference panel, 50 unrelated shapeit2 [125] trio-phased YRI and CEU samples each (from phase 3 TGP) and 43 shapeit2 population-phased Native American samples (from Ref. [126]) were used. We utilized local ancestry tracks that were inferred by RFMix using "trio-phased" mode.

Admixed samples were then parsed for all genomic segments homozygous for each particular ancestry (i.e., African, European, or Native American). These homozygous segments were also filtered according to the 13-filter set described in "Filtering and ascertainment scheme." Heterozygosity was calculated across admixed samples for each set of homozygous ancestries and *B* quantile bins described previously. Samples were included in the analyses only if the total length of their genome that passed all filters for the particular ancestry and *B* quantile bin was greater than 1 Mb. Additionally, only 100 kb segments that had at least 10 kb of divergence information with Rhesus macaque were used (see below). Per-site heterozygosity estimates for each ancestry and *B* quantile set were averaged across all admixed samples, regardless of their TGP population of origin. See Table S5 in Supporting information for total number of Mb used in the analyses. Additionally, heterozygosity was also calculated across the 4 continental groups using the same 13-filter set and as a function of the same *B* quantile bins.

Normalization of diversity and divergence calculations with Rhesus macaque

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

When normalizing diversity by divergence, only 100 kb bins that had at least 10 kb of divergence information were used (21,100 bins total for 13-filter set, 20,935 bins total for the 13-filter set plus the additional gBGC and hotspot filters). Bins with less than 10 kb of divergence information were ignored. To make estimates comparable, in those measurements of diversity that did not normalize by divergence, diversity was still calculated using the same set of 100 kb bins that had at least 10 kb for estimating divergence.

Calculations of population differentiation (F_{ST}) and linear regression

F_{ST} calculations were performed as a function of B and recombination rate between every pair of non-admixed phase 3 TGP populations not belonging to the same continental group (150 pairs total). We followed the recommendations in Bhatia et al. [56] to limit biases in F_{ST} due to 1) type of estimator used, 2) averaging over SNPs, and 3) SNP ascertainment. Specifically, we 1) used the Hudson-based F_{ST} estimator [127], 2) used a ratio of averages for combining F_{ST} estimated across different SNPs, and 3) ascertained SNPs based on being polymorphic in an outgroup (i.e., the KhoeSan). For ascertaining SNPs in the KhoeSan, we also performed filtering according to the filtering scheme described under “Filtering and ascertainment scheme.” For a position to be considered polymorphic in the KhoeSan, at least one alternate allele and one reference allele had to be called across the 13 genomes we utilized (see “Data”). These criteria left 3,497,105 total sites in the genome in the phase 3 dataset for F_{ST} to be estimated across.

F_{ST} was calculated across either 2% quantile bins of B or 2% quantile bins of recombination rate in order to perform simple linear regression, with either B or recombination rate acting as an explanatory variable using the linear model $F_{\mathrm{ST}} = \beta_0 + \beta_1 X + \varepsilon$ (where X represents either B or recombination rate [ρ]). This was done for all pairwise comparisons of populations between a specific pair of continental groups (25 pairs total) or across all pairwise comparisons using all continental groups (150 pairs total). Recombination rate was scaled to be between 0 and 1 (the minimum and maximum observed recombination rate was 0.0 cM/Mb and 126.88 cM/Mb, respectively) to aid in the comparison of the regression coefficient with *B*. Additionally, the mean of the bounds defining each quantile bin was used when defining the explanatory variables for the regression. Linear regression, robust linear regression, and simple correlation were performed using the lm(), rlm(), and cor() functions, respectively, in the R programming language (www. r-project.org). To generate standard errors of the mean, this same procedure was performed on F_{ST} results generated from each of 1,000 bootstrapped iterations of the data.

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

Bootstrapping

Diversity Estimates. To control for the structure of linkage disequilibrium and correlation between SNPs along the genome, we partitioned the human genome into non-overlapping 100 kb bins (these bins were identical to the 100 kb bins used for estimating divergence) and calculated mean pairwise diversity (π) or heterozygosity within each bin. We also normalized the diversity estimates by divergence within each bin. We then bootstrapped individual genomes by sampling, with replacement, the 100 kb bins until the number of sampled bins equaled the number of bins used for calculating the diversity point estimates (i.e., 21,100 bins or 20,935 bins total, depending on whether filters for gBGC and hotspots were applied). 1,000 total bootstrap iterations were completed and standard errors of the mean were calculated by taking the standard deviation from the resulting bootstrap distribution.

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

Acknowledgements

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

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

University Press; 2016;33: 3308–3313. doi:10.1093/molbev/msw195

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

on neutral variation. Genetics. 2000;155: 1415–1427. doi:drosophila-melanogaster; deleterious mutations; recombination; polymorphism; rates; population; evolution; variance; linkage; locus

25. Begun DJ, Aquadro CF. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature. Nature Publishing Group; 1992;356: 519–520. doi:10.1038/356519a0
26. Charlesworth B. Background selection and patterns of genetic diversity in Drosophila melanogaster. Genet Res. 1996;68: 131–149. doi:10.1017/S0016672300034029
27. Andolfatto P. Hitchhiking effects of recurrent beneficial amino acid substitutions in the Drosophila melanogaster genome. Genome Res. Cold Spring Harbor Laboratory Press; 2007;17: 1755–1762. doi:10.1101/gr.6691007
28. Sella G, Petrov DA, Przeworski M, Andolfatto P. Pervasive natural selection in the Drosophila genome? [Internet]. Nachman MW, editor. PLoS Genetics. Plenum; 2009. p. e1000495. doi:10.1371/journal.pgen.1000495
29. Comeron JM. Background Selection as Baseline for Nucleotide Variation across the Drosophila Genome. Begun DJ, editor. PLoS Genet. 2014;10: e1004434. doi:10.1371/journal.pgen.1004434
30. Elyashiv E, Sattath S, Hu TT, Strutsovsky A, McVicker G, Andolfatto P, et al. A Genomic Map of the Effects of Linked Selection in Drosophila. Barton NH, editor. PLoS Genet. Public Library of Science; 2016;12: e1006130. doi:10.1371/journal.pgen.1006130
31. Flowers JM, Molina J, Rubinstein S, Huang P, Schaal BA, Purugganan MD. Natural selection in gene-dense regions shapes the genomic pattern of polymorphism in wild and domesticated rice. Mol Biol Evol. Oxford University Press; 2012;29: 675–687. doi:10.1093/molbev/msr225
32. Xu X, Liu X, Ge S, Jensen JJDJJDJ, Hu F, Li X, et al. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat Biotechnol. 2012;30: 105–11. doi:10.1038/nbt.2050
33. Andersen EC, Gerke JP, Shapiro JA, Crissman JR, Ghosh R, Bloom JS, et al. Chromosome-scale selective sweeps shape Caenorhabditis elegans genomic diversity. Nat Genet. 2012;44: 285–90. doi:10.1038/ng.1050
34. Cutter AD, Payseur BA. Selection at linked sites in the partial selfer Caenorhabditis elegans. Mol Biol Evol. 2003;20: 665–673. doi:10.1093/molbev/msg072
35. Reed FA, Akey JM, Aquadro CF. Fitting background-selection predictions to levels of nucleotide variation and divergence along the human autosomes. Genome Res. 2005;15: 1211–1221. doi:10.1101/gr.3413205
36. 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
37. Cai JJ, Macpherson JM, Sella G, Petrov DA. Pervasive hitchhiking at coding and regulatory sites in humans. PLoS Genet. 2009;5: e1000336. doi:10.1371/journal.pgen.1000336
38. Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, et al. 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

54. Moreno-Estrada A, Gignoux CR, Fernández-López JC, Zakharia F, Sikora M, Contreras A V, et al. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. Science (80-). 2014;344: 1280–5. doi:10.1126/science.1251688

55. Hey J. On the number of new world founders: A population genetic portrait of the peopling of the Americas. Clark AG, editor. PLoS Biol. Springer; 2005;3: 0965–0975. doi:10.1371/journal.pbio.0030193

56. Bhatia G, Patterson N, Sankararaman S, Price AL. Estimating and interpreting FST: The impact of rare variants. Genome Res. 2013;23: 1514–1521. doi:10.1101/gr.154831.113

57. Keinan A, Reich D. Human population differentiation is strongly correlated with local recombination rate. Begun DJ, editor. PLoS Genet. Public Library of Science; 2010;6: e1000886. doi:10.1371/journal.pgen.1000886

58. Willing EM, Dreyer C, van Oosterhout C. Estimates of genetic differentiation measured by fst do not necessarily require large sample sizes when using many snp markers. PLoS One. 2012;7: 1–7. doi:10.1371/journal.pone.0042649

59. Messer PW, Petrov DA. Frequent adaptation and the McDonald-Kreitman test. Proc Natl Acad Sci U S A. 2013;110: 8615–20. doi:10.1073/pnas.1220835110

60. Ewing GB, Jensen JD. The consequences of not accounting for background selection in demographic inference. Mol Ecol. Wiley-Blackwell; 2016;25: 135–141. doi:10.1111/mec.13390

61. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, et al. Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci. 2011;108: 11983–11988. doi:10.1073/pnas.1019276108

62. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. Science (80-). 2012;337: 64–69. doi:10.1126/science.1219240

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

64. Burgess R, Yang Z. Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. Mol Biol Evol. Oxford University Press; 2008;25: 1979–1994. doi:10.1093/molbev/msn148

65. Mouse Genome Sequencing Consortium MGS, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002;420: 520–62. doi:10.1038/nature01262

66. Charlesworth B, Nordborg M, Charlesworth D. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet Res. 1997;70: 155–174. doi:10.1017/S0016672397002954

67. Hu XS, He F. Background selection and population differentiation. J Theor Biol. 2005;235: 207–219. doi:10.1016/j.jtbi.2005.01.004

68. Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Absher D, et al. The role of geography in human adaptation. Schierup MH, editor. PLoS Genet. R Foundation

for Statistical Computing; 2009;5: e1000500. doi:10.1371/journal.pgen.1000500

69. Bhatia G, Patterson N, Pasaniuc B, Zaitlen N, Genovese G, Pollack S. Genome-Wide Comparison of African-Ancestry Populations from CARe and Other Cohorts Reveals Signals of Natural Selection. 2011;89: 368–381. doi:10.1016/j.ajhg.2011.07.025.

70. Eyre-Walker A, Keightley PD. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol. 2009;26: 2097–2108. doi:10.1093/molbev/msp119

71. Galtier N. Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis. Schierup MH, editor. PLoS Genet. Oxford University Press; 2016;12: e1005774. doi:10.1371/journal.pgen.1005774

72. 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

73. 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

74. Fu W, Gittelman RM, Bamshad MJ, Akey JM. Characteristics of neutral and deleterious protein-coding variation among individuals and populations. Am J Hum Genet. Springer-Verlag, New York; 2014;95: 421–436. doi:10.1016/j.ajhg.2014.09.006

75. Henn BM, Botigué LR, Peischl S, Dupanloup I, Lipatov M, Maples BK, et al. Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. Proc Natl Acad Sci. National Academy of Sciences; 2015;113: 201510805. doi:10.1073/pnas.1510805112

76. Casals F, Hodgkinson A, Hussin J, Idaghdour Y, Bruat V, de Maillard T, et al. Whole-Exome Sequencing Reveals a Rapid Change in the Frequency of Rare Functional Variants in a Founding Population of Humans. Williams SM, editor. PLoS Genet. Public Library of Science; 2013;9: e1003815. doi:10.1371/journal.pgen.1003815

77. Lim ET, Würtz P, Havulinna AS, Palta P, Tukiainen T, Rehnström K, et al. Distribution and Medical Impact of Loss-of-Function Variants in the Finnish Founder Population. Cutler D, editor. PLoS Genet. Public Library of Science; 2014;10: e1004494. doi:10.1371/journal.pgen.1004494

78. Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, et al. Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet. 2011;43: 956–963. doi:10.1038/ng.911

79. Renaut S, Rieseberg LH. The accumulation of deleterious mutations as a consequence of domestication and improvement in sunflowers and other compositae crops. Mol Biol Evol. Oxford University Press; 2015;32: 2273–2283. doi:10.1093/molbev/msv106

80. Balick DJ, Do R, Cassa CA, Reich D, Sunyaev SR. Dominance of Deleterious Alleles Controls the Response to a Population Bottleneck. Coop G, editor. PLoS Genet. Cambridge University Press; 2015;11: e1005436. doi:10.1371/journal.pgen.1005436

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

hunter-gatherers have been the largest population throughout most of modern-human demographic history. Nat Commun. 2014;5: 5692. doi:10.1038/ncomms6692

111. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. Cold Spring Harbor Laboratory Press; 2010;20: 110–121. doi:10.1101/gr.097857.109

112. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15: 1034–1050. doi:10.1101/gr.3715005

113. Consortium ENCODEP, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489: 57–74. doi:10.1038/nature11247

114. Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. Segmental duplications: organization and impact within the current human genome project assembly. Genome Res. 2001;11: 1005–17. doi:10.1101/gr.187101

115. Capra JA, Hubisz MJ, Kostka D, Pollard KS, Siepel A. A Model-Based Analysis of GC-Biased Gene Conversion in the Human and Chimpanzee Genomes. Coop G, editor. PLoS Genet. Public Library of Science; 2013;9: e1003684. doi:10.1371/journal.pgen.1003684

116. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449: 851–61. doi:10.1038/nature06258

117. Pratto F, Brick K, Khil P, Smagulova F, Petukhova G V, Camerini-Otero RD. DNA recombination. Recombination initiation maps of individual human genomes. Science. 2014;346: 1256442. doi:10.1126/science.1256442

118. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. Oxford University Press; 2010;38: e164–e164. doi:10.1093/nar/gkq603

119. Hernandez RD, Williamson SH, Zhu L, Bustamante CD. Context-dependent mutation rates may cause spurious signatures of a fixation bias favoring higher GC-content in humans. Mol Biol Evol. 2007;24: 2196–2202. doi:10.1093/molbev/msm149

120. Schrider DR, Shanku AG, Kern AD. Effects of linked selective sweeps on demographic inference and model selection. Genetics. 2016;204: 1207–1223. doi:10.1534/genetics.116.190223

121. Palamara PF, Francioli LC, Wilton PR, Genovese G, Gusev A, Finucane HK, et al. Leveraging Distant Relatedness to Quantify Human Mutation and Gene-Conversion Rates. Am J Hum Genet. 2015;97: 775–789. doi:10.1016/j.ajhg.2015.10.006

122. Hernandez RD. A flexible forward simulator for populations subject to selection and demography. Bioinformatics. 2008;24: 2786–2787. doi:10.1093/bioinformatics/btn522

123. Torgerson DG, Boyko AR, Hernandez RD, Indap A, Hu X, White TJ, et al. Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. PLoS Genet. 2009;5. 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* ≥ 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%

BGS bin (π/π_0). Error bars represent ±1 SEM calculated from 1,000 bootstrapped datasets.

Fig S3. Diversity for TGP non-admixed populations without normalizing by divergence with Rhesus macaque. (A) Diversity (π) measured across the strongest 1% BGS bin. (B) Diversity measured across the weakest 1% BGS bin. (C) Relative diversity: the ratio of diversity for the strongest 1% BGS bin to diversity for the weakest 1% BGS bin (π/π_0). Error bars represent ±1 SEM calculated from 1,000 bootstrapped datasets.

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

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

Fig S6. F_{ST} measured across joint bins of *B* and recombination rate for different TGP continental groups. The left panels of Figures S6 A-E show F_{ST} measured as a function of 25 4% quantile recombination rate bins conditional on three 2% quantile *B* 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

human populations. Right column plots show results of simulations under a demographic model with no migration. Colored lines represent different populations though time and are identical to those in Figure 6 and Figure S7. The demographic model used is also identical to that in Figure 6 (for simulations with migration) and Figure S7 (for simulations without migration). Simulation data was sampled every 100 generations.

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

Main Figure Legends

Fig 1. Normalized diversity and relative diversity for non-admixed populations of the Thousand Genomes Project (TGP).

(A) Normalized diversity (π/divergence) measured across the strongest 1% background selection (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% BGS bin (π/π_0). BGS bins were classified by *B*. TGP population labels are indicated below each bar (see Table S11 in Supporting information for population label descriptions), with African populations colored by gold shades, European populations colored by blue shades, South Asian populations colored by violet shades, and East Asian populations colored by green shades. Error bars represent ±1 SEM calculated from 1,000 bootstrapped datasets.

Fig 2. Normalized and relative diversity for Thousand Genomes Project (TGP) continental groups.

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

Fig 3. Comparing patterns of diversity between local ancestry segments of admixed samples and continental groups.

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

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

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

Fig 5. *F*$_{\mathrm{ST}}$ between African (AFR) and South Asian (SASN) populations jointly across *B* and recombination rate.

(A) F_{ST} as a function of 25 recombination rate bins (4% quantile bins) conditional on three different 2% quantile *B* bins (note log scale of x-axis for recombination rate). (B) F_{ST} as a function of 25 *B* bins (4% quantile bins) conditional on three different 2% quantile recombination rate bins. Smaller transparent points and lines show the F_{ST} estimates and corresponding lines of best fit (using linear regression) for each of the pairwise comparisons of AFR vs. SASN Thousand Genomes Project (TGP) populations (25 comparisons total). Larger opaque points are mean F_{ST} estimates across all pairwise comparisons of AFR vs. SASN TGP populations (with bold lines showing their corresponding lines of best fit).

Fig 6. Simulations confirm that demographic events shape the impact of background selection (BGS).

(A) Inferred demographic model from Complete Genomics Thousand Genomes Project (TGP) data showing population size changes for Africans (AFR), Europeans (EUR), and East Asians (EASN) as a function of time that was used for the simulations of BGS. (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.

Figure 1

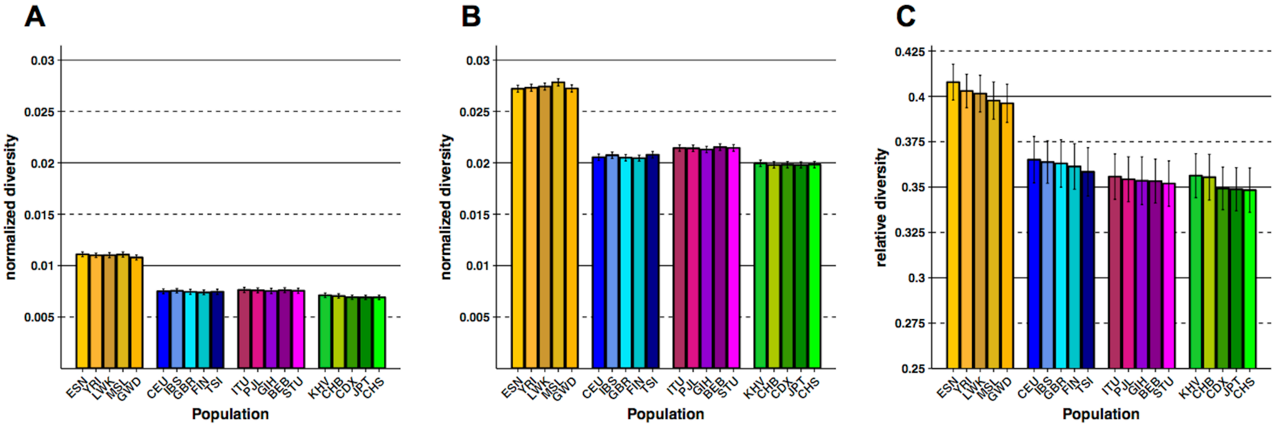

A
B
C
normalized diversity
normalized diversity
relative diversity
0.03
0.025
0.02
0.015
0.01
0.005
0.425
0.4
0.375
0.35
0.325
0.3
0.275
0.25
ESN
YRI
LWK
MSL
GWD
CEU
IBS
GBR
FIN
TSI
ITU
PJL
GIH
BEB
STU
KHV
CHB
CDX
JPT
CHS
Population
Population
Population

Figure 2

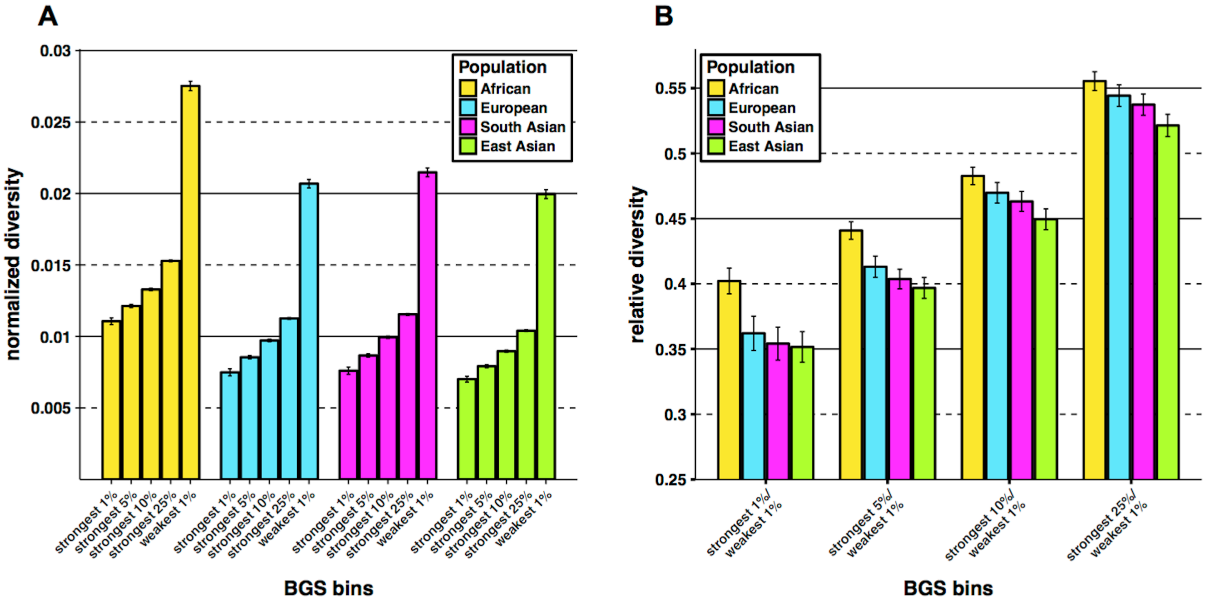

A
0.03
0.025
0.02
0.015
0.01
0.005
normalized diversity
Population
🟨 African
🟦 European
🟪 South Asian
🟩 East Asian
strongest 1%
strongest 5%
strongest 10%
strongest 25%
weakest 1%
BGS bins
B
0.55
0.5
0.45
0.4
0.35
0.3
0.25
relative diversity
Population
🟨 African
🟦 European
🟪 South Asian
🟩 East Asian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

Figure 3

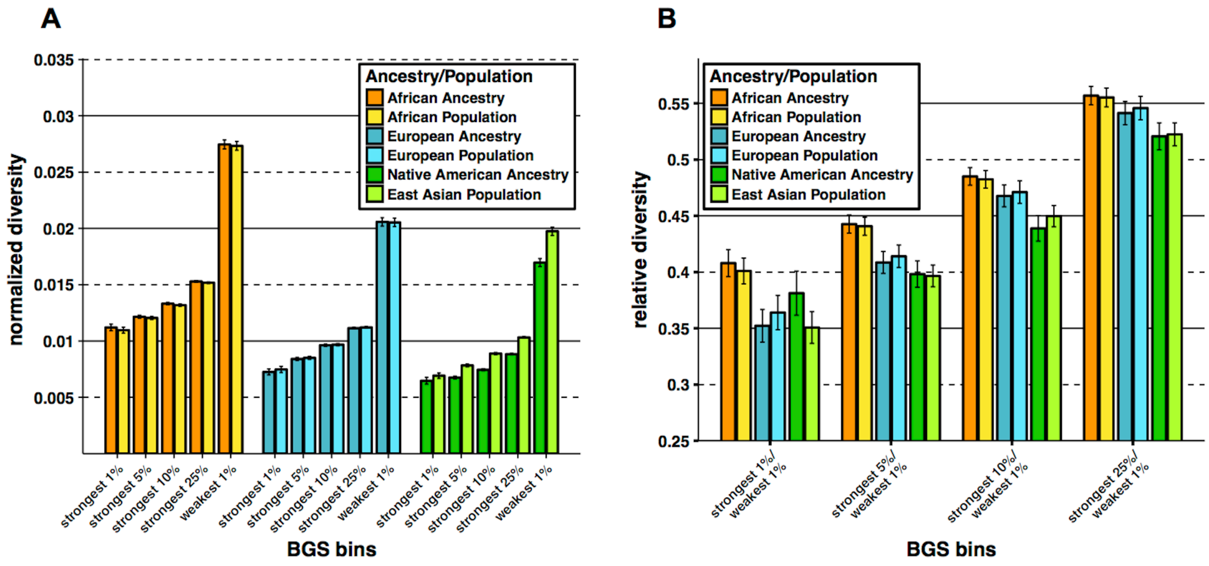

A
0.035
0.03
0.025
0.02
0.015
0.01
0.005
normalized diversity
Ancestry/Population
African Ancestry
African Population
European Ancestry
European Population
Native American Ancestry
East Asian Population
strongest 1%
strongest 5%
strongest 10%
strongest 25%
weakest 1%
BGS bins
B
0.55
0.5
0.45
0.4
0.35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

Figure 4

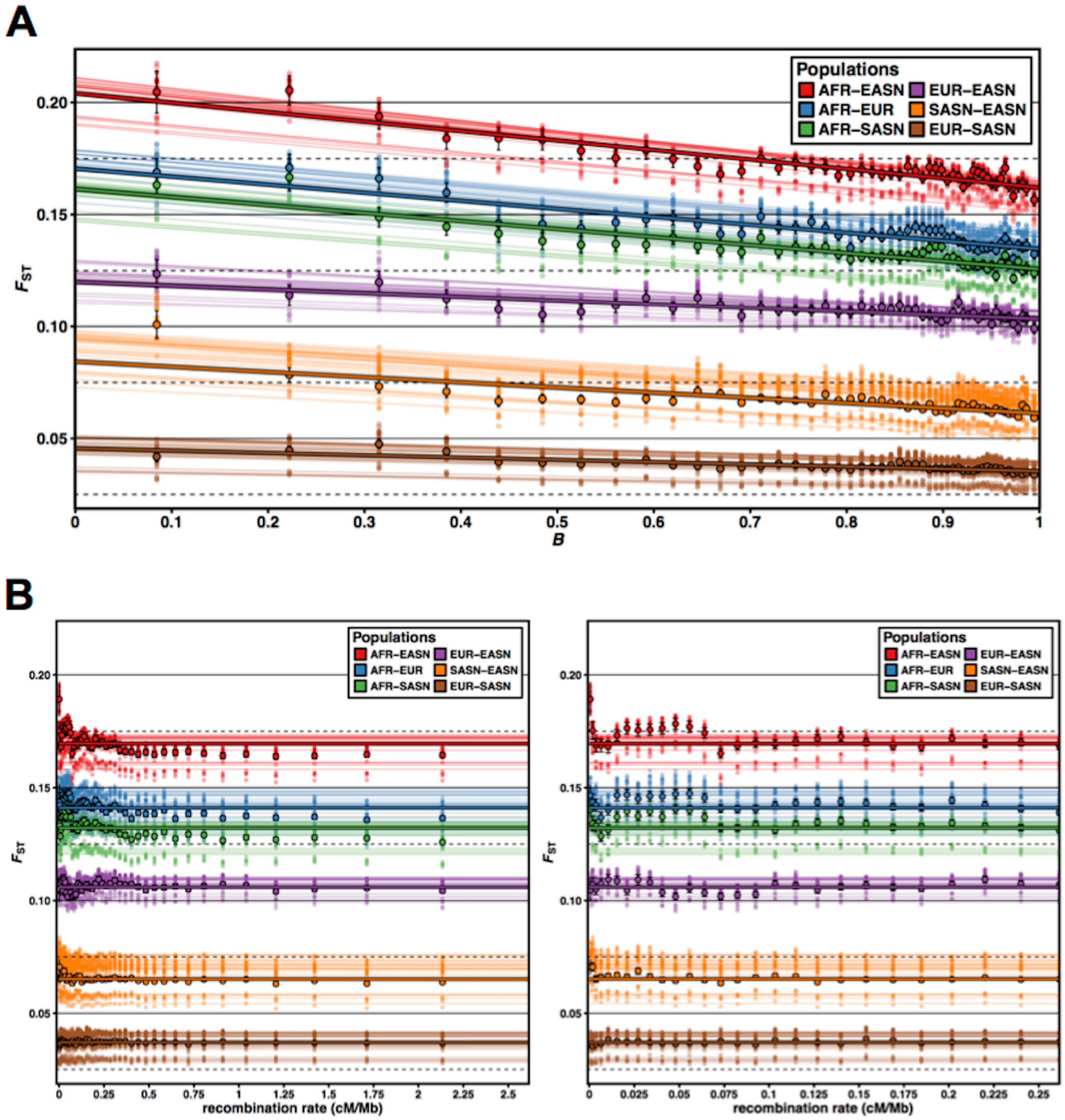

A
Populations
🟥 AFR–EASN
🟦 AFR–EUR
🟩 AFR–SASN
🟪 EUR–EASN
🟧 SASN–EASN
🟫 EUR–SASN
F_{ST}
0.20
0.15
0.10
0.05
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
B
B
Populations
🟥 AFR–EASN
🟦 AFR–EUR
🟩 AFR–SASN
🟪 EUR–EASN
🟧 SASN–EASN
🟫 EUR–SASN
F_{ST}
0.20
0.15
0.10
0.05
0
0.25
0.5
0.75
1
1.25
1.5
1.75
2
2.25
2.5
recombination rate (cM/Mb)
Populations
🟥 AFR–EASN
🟦 AFR–EUR
🟩 AFR–SASN
🟪 EUR–EASN
🟧 SASN–EASN
🟫 EUR–SASN
F_{ST}
0.20
0.15
0.10
0.05
0
0.025
0.05
0.075
0.1
0.125
0.15
0.175
0.2
0.225
0.25
recombination rate (cM/Mb)

Figure 5

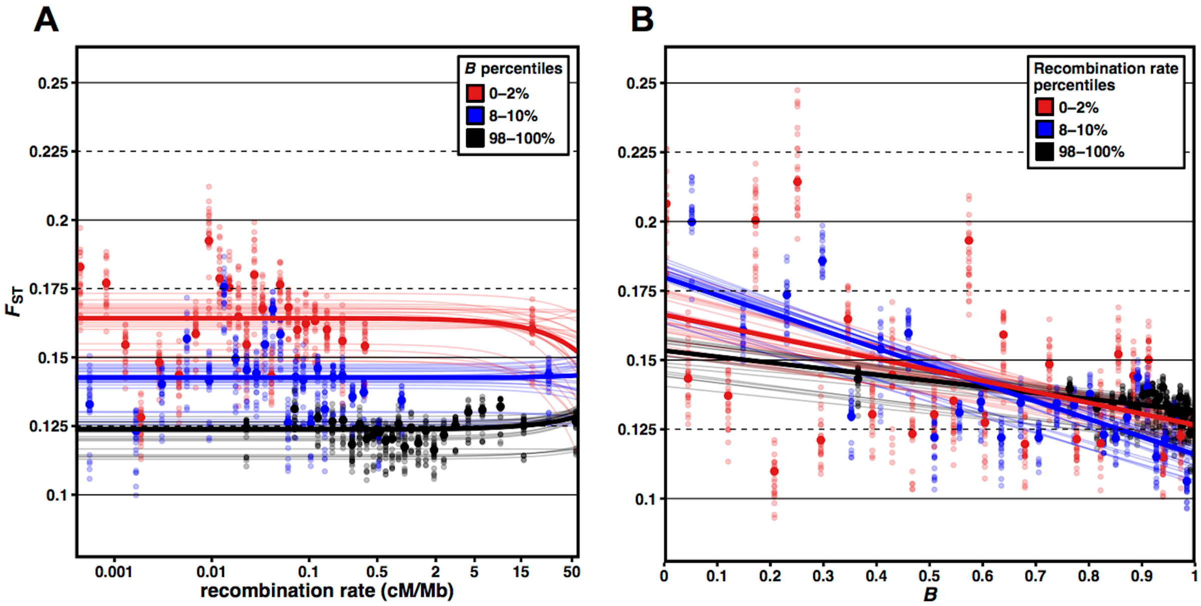
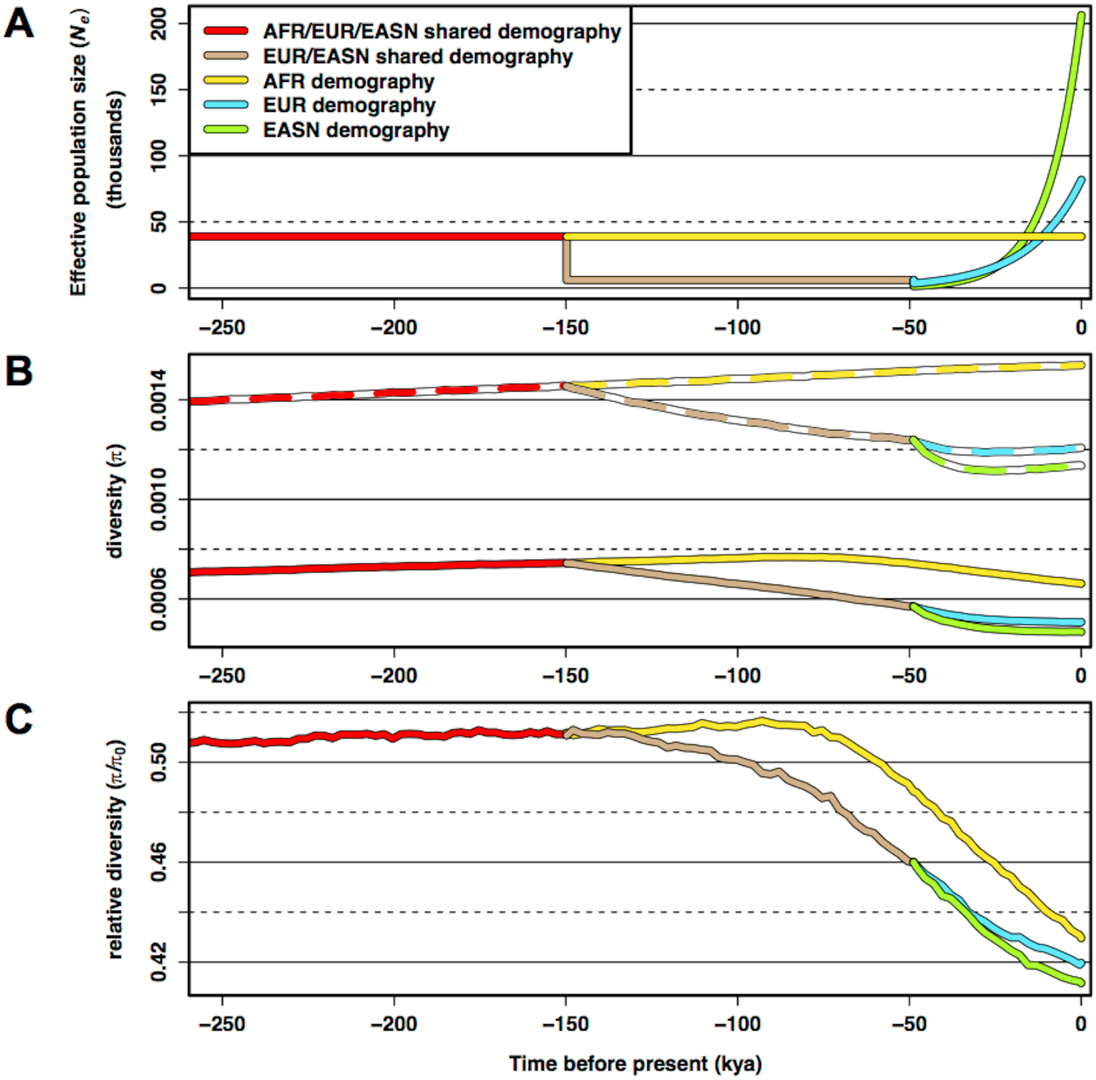

A
B percentiles
🟥 0–2%
🟦 8–10%
⬛ 98–100%
F_{ST}
0.25
0.225
0.2
0.175
0.15
0.125
0.1
0.001
0.01
0.1
0.5
2
5
15
50
recombination rate (cM/Mb)
B
Recombination rate percentiles
🟥 0–2%
🟦 8–10%
⬛ 98–100%
0.25
0.225
0.2
0.175
0.15
0.125
0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
B

Figure 6


A
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
−250
−200
−150
−100
−50
0
B
diversity (π)
0.0006
0.0010
0.0014
−250
−200
−150
−100
−50
0
C
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* ≥ 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
$m_{AFR\text{-}Bott}$ (x10^{-5})	7.83	7.02
$m_{AFR\text{-}EUR}$ (x10^{-5})	0.51	0.47
$m_{AFR\text{-}EASN}$ (x10^{-5})	0.13	0.18
$m_{EUR\text{-}EASN}$ (x10^{-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* ≥ 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

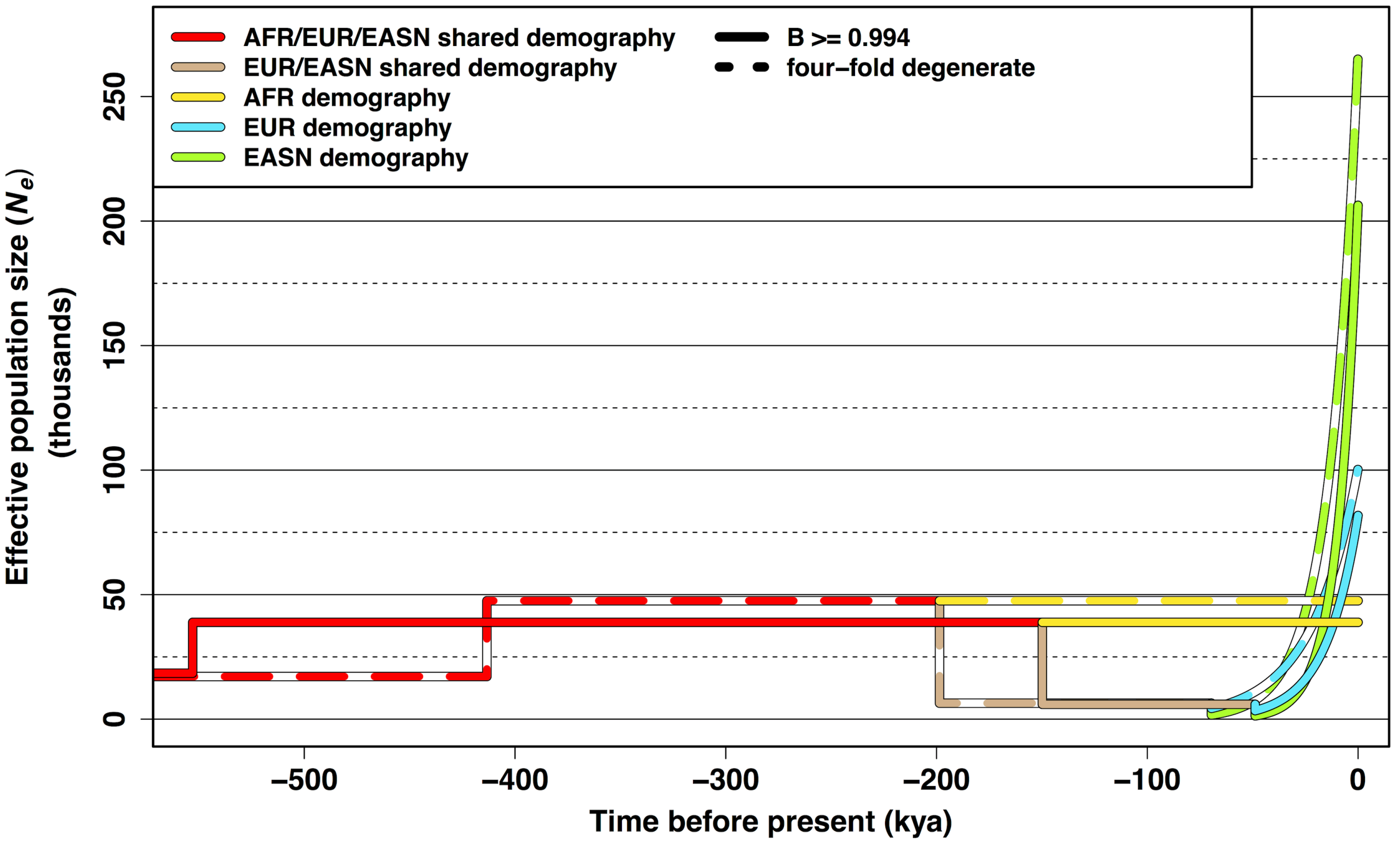

AFR/EUR/EASN shared demography
EUR/EASN shared demography
AFR demography
EUR demography
EASN demography
B >= 0.994
four−fold degenerate
Effective population size (N_e) (thousands)
0
50
100
150
200
250
−500
−400
−300
−200
−100
0
Time before present (kya)

Figure S2

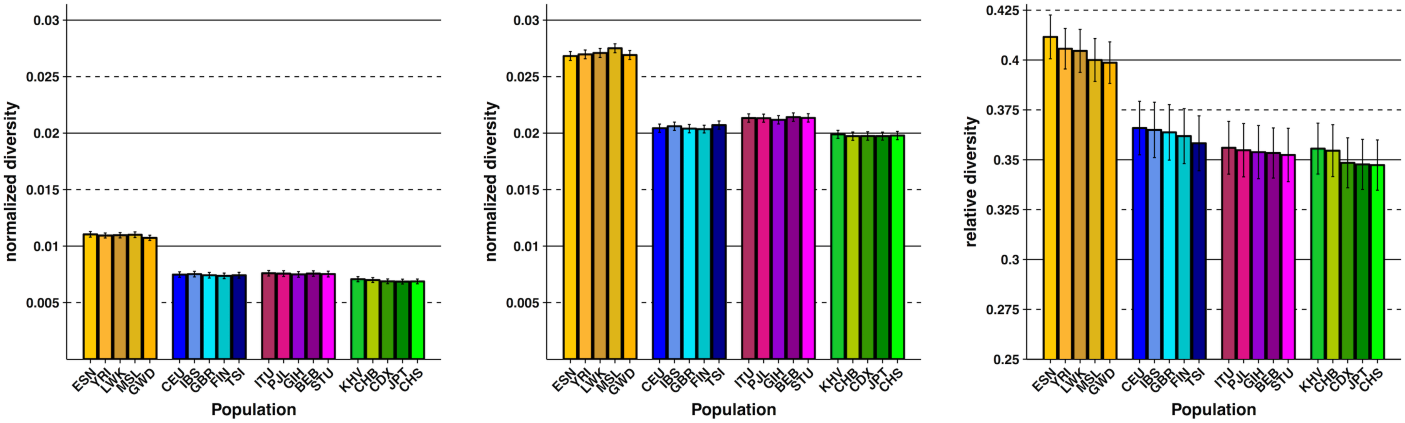

normalized diversity
0.03
0.025
0.02
0.015
0.01
0.005
ESN
YRI
LWK
MSL
GWD
CEU
IBS
GBR
FIN
TSI
ITU
PJL
GIH
BEB
STU
KHV
CHB
CDX
JPT
CHS
Population
normalized diversity
0.03
0.025
0.02
0.015
0.01
0.005
ESN
YRI
LWK
MSL
GWD
CEU
IBS
GBR
FIN
TSI
ITU
PJL
GIH
BEB
STU
KHV
CHB
CDX
JPT
CHS
Population
relative diversity
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

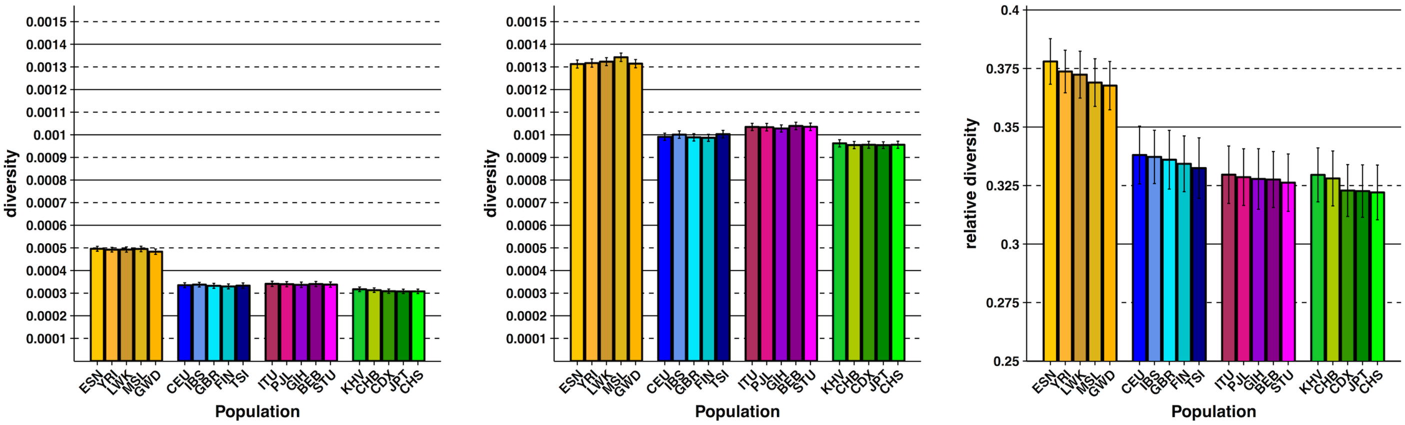

diversity
0.0015
0.0014
0.0013
0.0012
0.0011
0.001
0.0009
0.0008
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
ESN
YRI
LWK
MSL
GWD
CEU
IBS
GBR
FIN
TSI
ITU
PJL
GIH
BEB
STU
KHV
CHB
CDX
JPT
CHS
Population
relative diversity
0.4
0.375
0.35
0.325
0.3
0.275
0.25

Figure S4

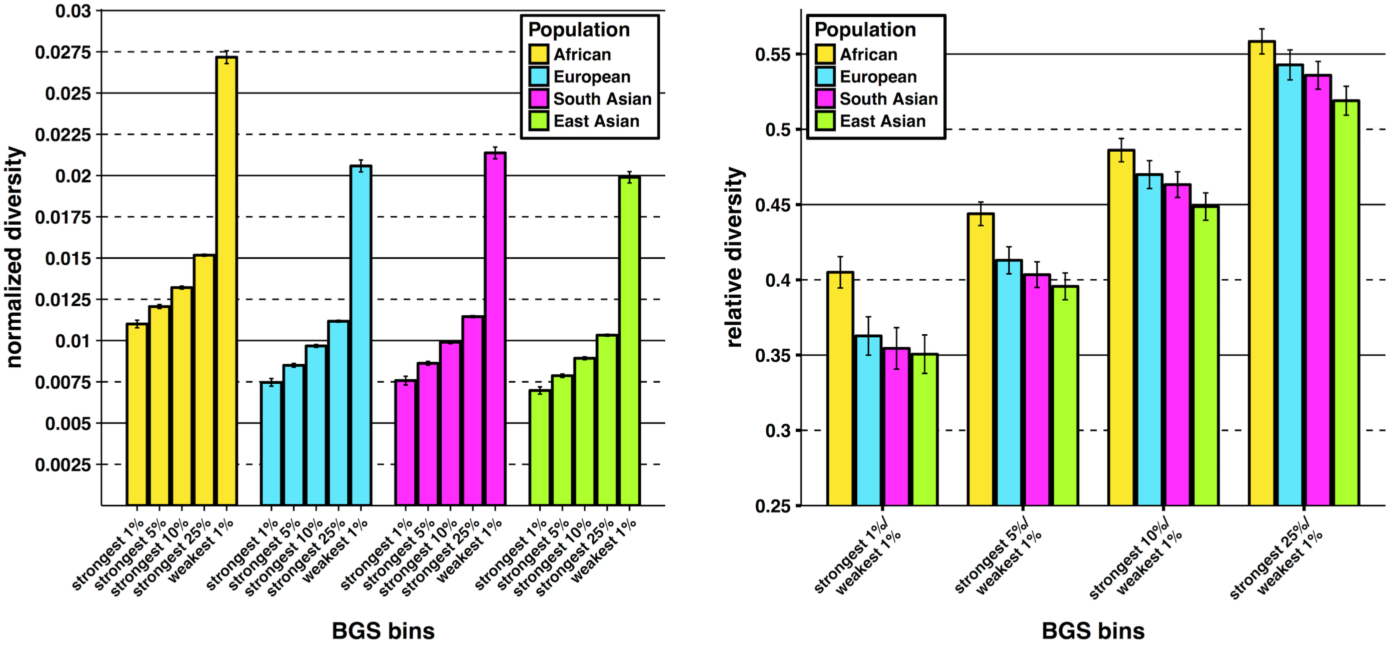

Population
African
European
South Asian
East Asian
normalized diversity
0.03
0.0275
0.025
0.0225
0.02
0.0175
0.015
0.0125
0.01
0.0075
0.005
0.0025
strongest 1%
strongest 5%
strongest 10%
strongest 25%
weakest 1%
strongest 1%
strongest 5%
strongest 10%
strongest 25%
weakest 1%
strongest 1%
strongest 5%
strongest 10%
strongest 25%
weakest 1%
strongest 1%
strongest 5%
strongest 10%
strongest 25%
weakest 1%
BGS bins
Population
African
European
South Asian
East Asian
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

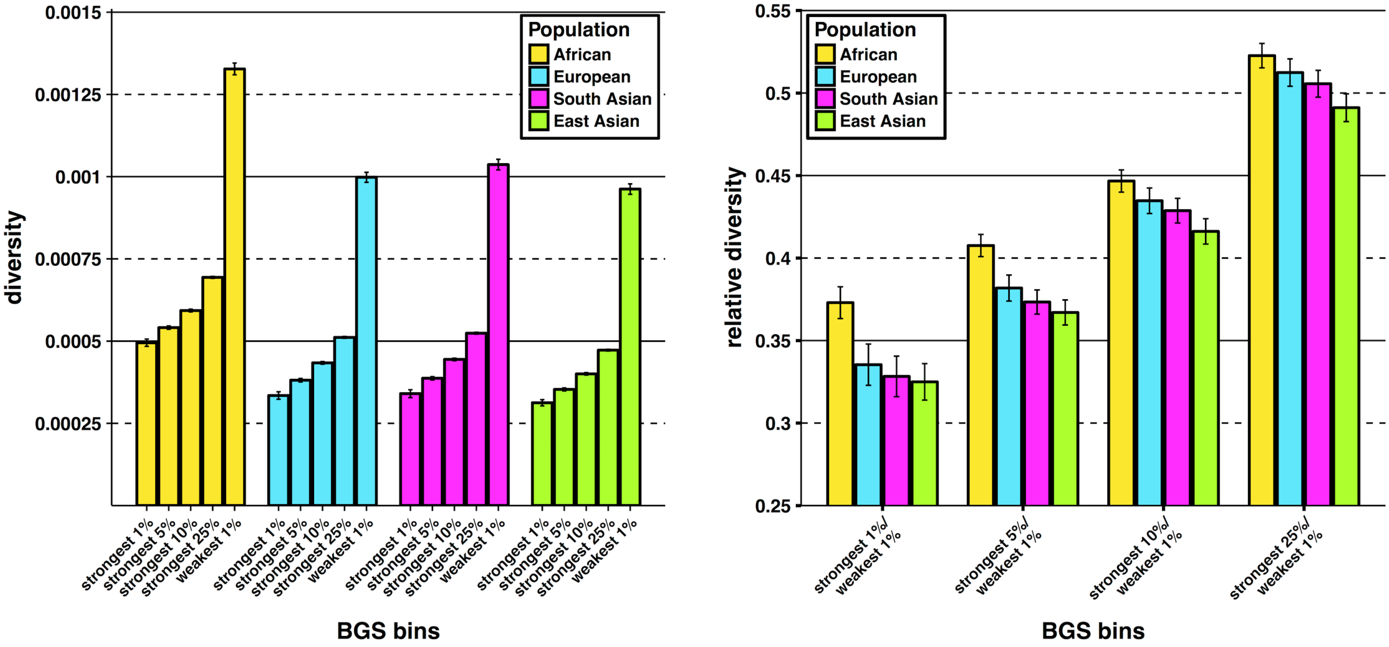

Population
African
European
South Asian
East Asian
diversity
0.0015
0.00125
0.001
0.00075
0.0005
0.00025
strongest 1%
strongest 5%
strongest 10%
strongest 25%
weakest 1%
BGS bins
relative diversity
0.55
0.5
0.45
0.4
0.35
0.3
0.25
strongest 1%/
weakest 1%
strongest 5%/
weakest 1%
strongest 10%/
weakest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

Figure S6

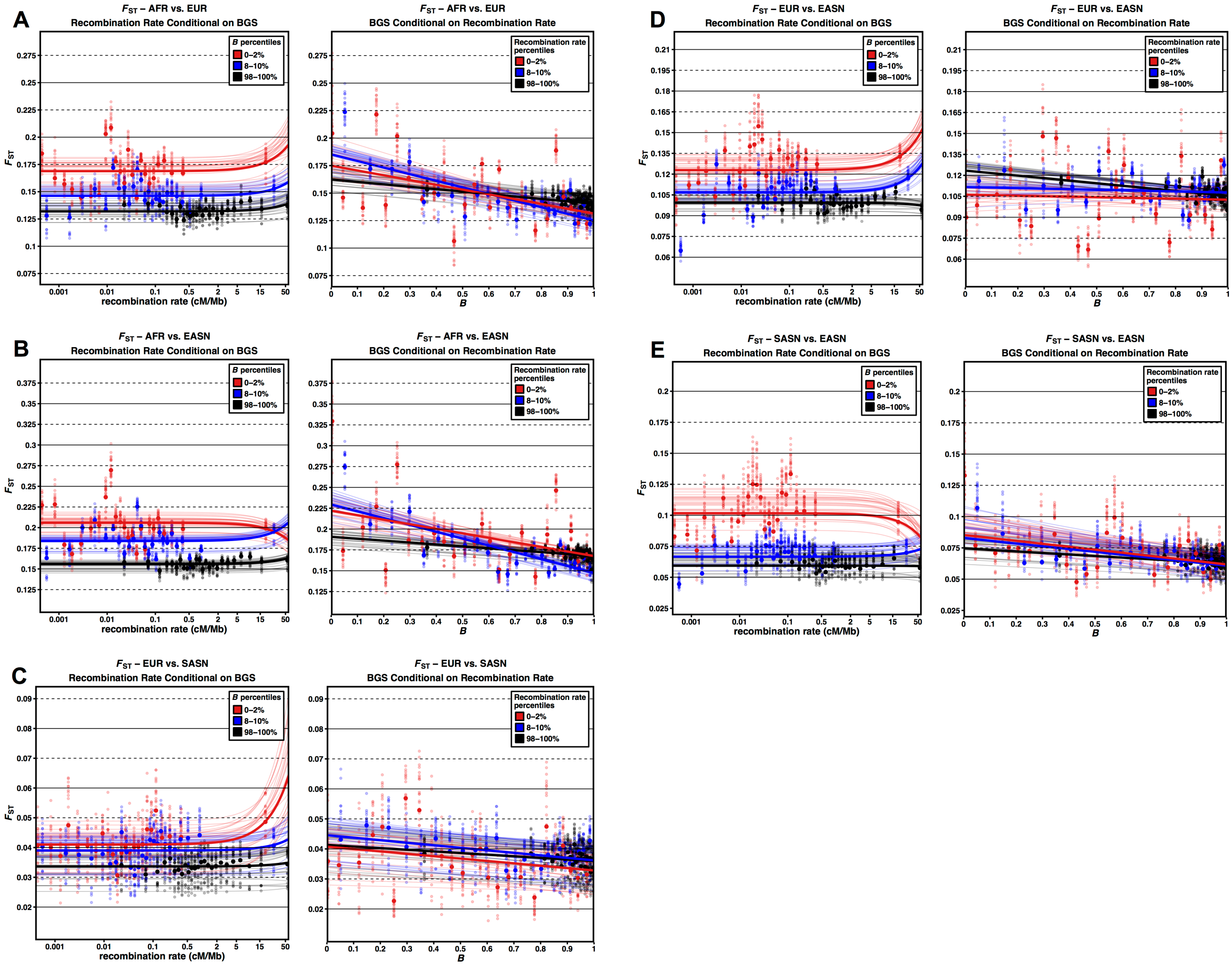

A
F_{ST} – AFR vs. EUR
Recombination Rate Conditional on BGS
BGS Conditional on Recombination Rate
B
F_{ST} – AFR vs. EASN
C
F_{ST} – EUR vs. SASN
D
F_{ST} – EUR vs. EASN
E
F_{ST} – SASN vs. EASN
B percentiles
Recombination rate percentiles
0–2%
8–10%
98–100%
F_{ST}
recombination rate (cM/Mb)
0.001
0.01
0.1
0.5
2
5
15
50
B
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Figure S7

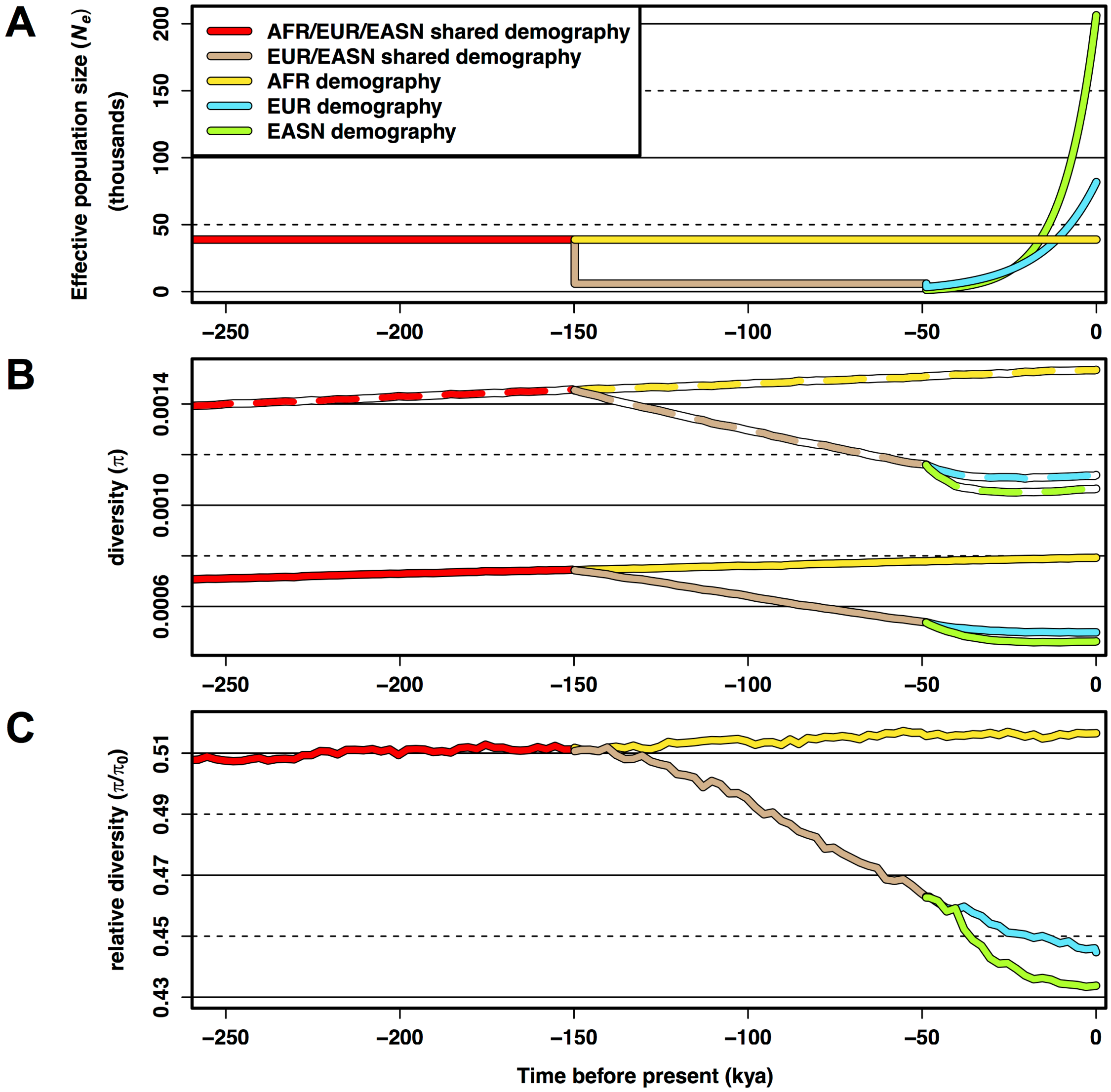
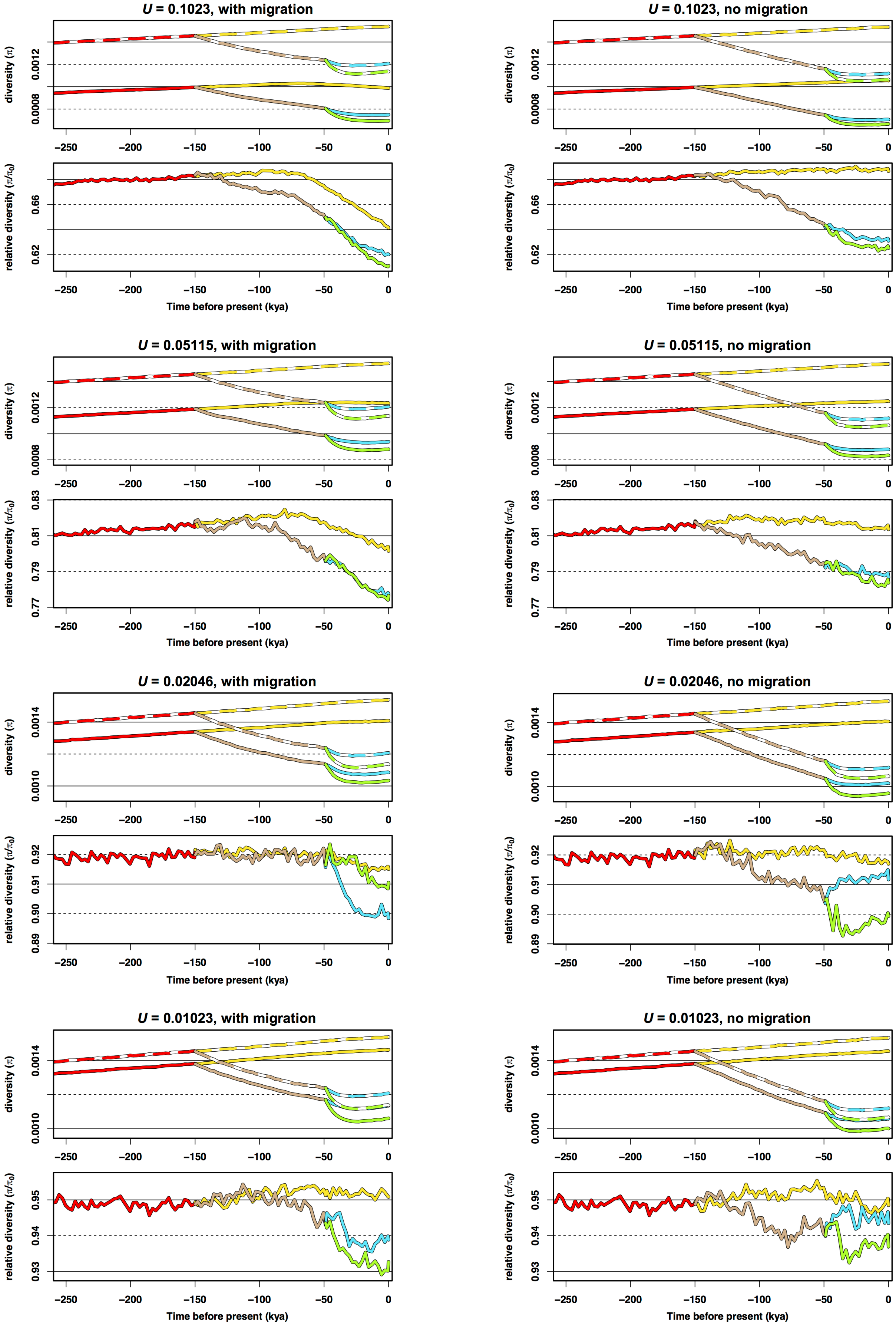

A
Effective population size (N_e) (thousands)
0
50
100
150
200
AFR/EUR/EASN shared demography
EUR/EASN shared demography
AFR demography
EUR demography
EASN demography
−250
−200
−150
−100
−50
0
B
diversity (π)
0.0006
0.0010
0.0014
−250
−200
−150
−100
−50
0
C
relative diversity (π/π_0)
0.43
0.45
0.47
0.49
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)
0.0008
0.0012
0.0010
0.0014
0.62
0.66
0.77
0.79
0.81
0.83
0.89
0.90
0.91
0.92
0.93
0.94
0.95
−250
−200
−150
−100
−50
0
Time before present (kya)

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 i^{th} 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

$$\begin{aligned} iHH12 \; &= \sum_{i=1}^{|\mathcal{D}|} \frac{1}{2}\left(EHH12(x_{i-1}) - EHH(x_i)\right) g(x_{i-1}, x_i) \\ &\quad + \sum_{i=1}^{|\mathcal{U}|} \frac{1}{2}\left(EHH12(x_{i-1}) - EHH(x_i)\right) g(x_{i-1}, x_i) \end{aligned}$$

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%.

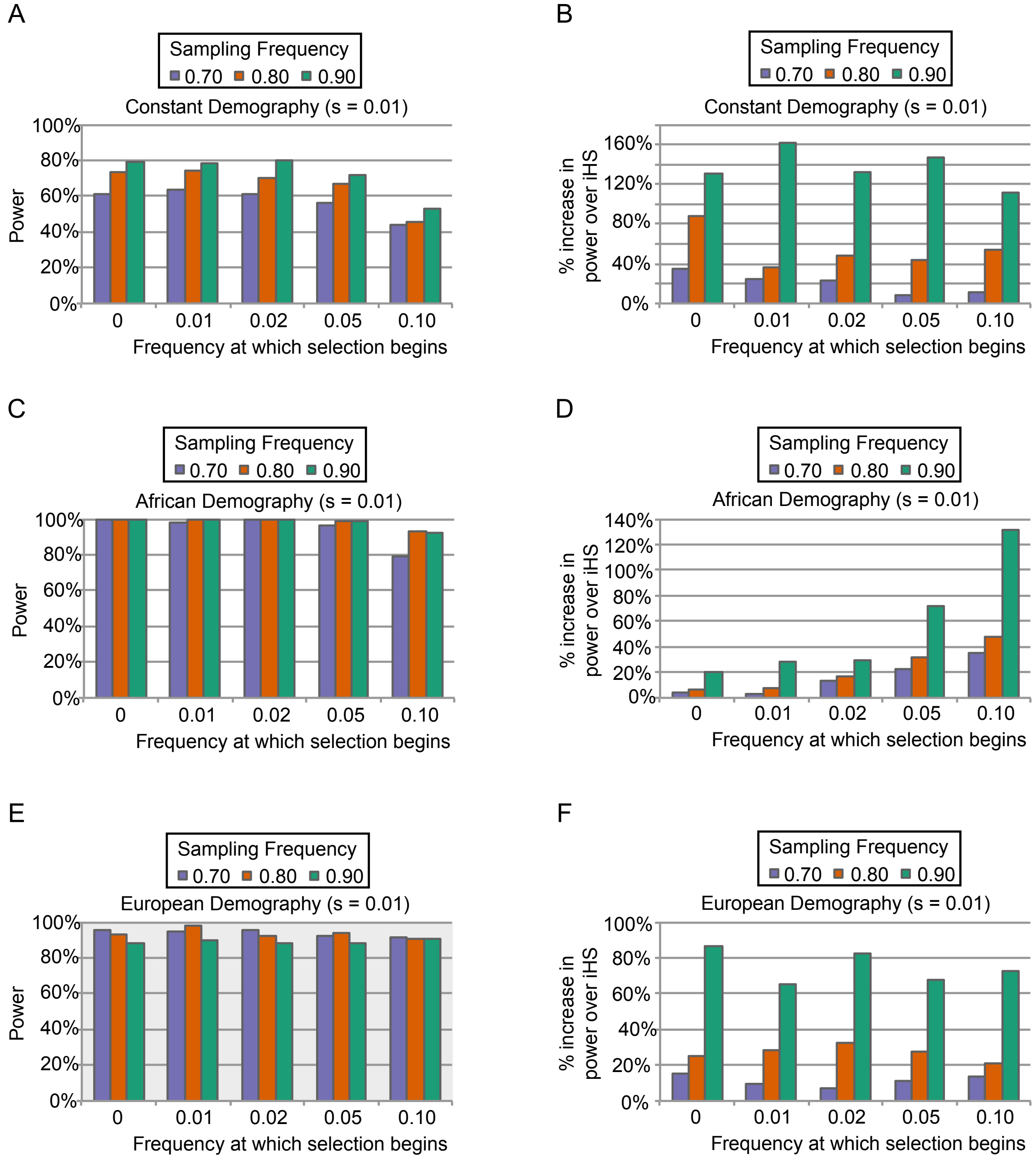
$iHH12$ has good power to detect hard and soft sweeps (Appendix Fig A1 A, C, E) and has improved power to identify both types of sweeps over iHS (Appendix Fig A1 B, D, F), particularly under realistic models of human demography.

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