RESEARCH ARTICLE

Human demographic history has amplified the effects of background selection across the genome

Raul Torres[1], **Zachary A. Szpiech**[2], **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

* ryan.hernandez@ucsf.edu

OPEN ACCESS

Citation: Torres R, Szpiech ZA, Hernandez RD (2018) Human demographic history has amplified the effects of background selection across the genome. PLoS Genet 14(6): e1007387. https://doi.org/10.1371/journal.pgen.1007387

Editor: Graham Coop, University of California Davis, UNITED STATES

Received: August 29, 2017

Accepted: April 30, 2018

Published: June 18, 2018

Data Availability Statement: All genetic data is stored at 1000genomes.org. The specific version of the data used can be accessed from the following EBI ftp site: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/.

Funding: This work was supported by National Institutes of Health (R01 HG007644 RDH) and the National Science Foundation Graduate Research Fellowship Program to RT. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Natural populations often grow, shrink, and migrate over time. Such demographic processes can affect genome-wide levels of genetic diversity. Additionally, genetic variation in functional regions of the genome can be altered by natural selection, which drives adaptive mutations to higher frequencies or purges deleterious ones. Such selective processes affect not only the sites directly under selection but also nearby neutral variation through genetic linkage via processes referred to as genetic hitchhiking in the context of positive selection and background selection (BGS) in the context of purifying selection. While there is extensive literature examining the consequences of selection at linked sites at demographic equilibrium, less is known about how non-equilibrium demographic processes influence the effects of hitchhiking and BGS. 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 selection at linked sites across the human genome. When binning the genome by inferred strength of BGS, we observe that, compared to Africans, non-African populations have experienced larger proportional decreases in neutral genetic diversity in strong BGS regions. We replicate these findings in admixed populations by showing that non-African ancestral components of the genome have also been affected more severely in these regions. We attribute these differences to the strong, sustained/recurrent 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 the inferred strength of BGS, suggesting a stronger rate of genetic drift. Forward simulations of human demographic history with a model of BGS support these observations. Our results show that non-equilibrium demography significantly alters the consequences of selection at linked sites and support the need for more work investigating the dynamic process of multiple evolutionary forces operating in concert.

Competing interests: The authors have declared that no competing interests exist.

Author summary

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

Introduction

Genetic diversity within a species is shaped by 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 shaped by population size changes, yielding patterns 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 advances 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 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 and founder events 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 long lasting signatures of decreased diversity, which may be depressed even after a population has recovered to, or surpassed, its ancestral size [20,21]. Such examples are evident in humans, where non-African populations exhibit a lower amount of genetic diversity compared to Africans [9], despite the fact that they have been inferred to have undergone a greater population expansion in recent times [22,23].

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 [24]). For example, mutations with functional effects may be removed from the population due to purifying selection or become fixed 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 [25] causes variants lying on the same haplotype as the selected allele to rise to high frequency during the selection process (note that we will use the term "genetic hitchhiking" here only in the positive selection context of selection at linked sites). Conversely, under purifying selection, background selection (BGS) [26] causes linked neutral variants to decrease in frequency or be removed from the population. Both of these processes of selection at linked sites 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 [27].

Evidence for genetic hitchhiking and BGS has been obtained from the genomes of several species, including *Drosophila melanogaster* [28–33], wild and domesticated rice [34,35], nematodes [36,37], humans [3,6,38–42], and others (see [1] for a review). While the relative contributions of genetic hitchhiking and BGS to shaping patterns of human genomic diversity have been actively debated [40,43–45], the data strongly support the large role of BGS in shaping genome-wide patterns of neutral genetic variation [41,42]. Indeed, recent arguments have been made in favor of BGS being treated as the null model when investigating the effect of selection at linked sites across recombining genomes [1,32,45–48], with one study in humans showing that BGS has reduced genetic diversity by 19–26% if other modes of selection at linked sites are assumed to be minor [6].

Although the effects of selection at linked sites across the genome have been described in a multitude of studies, it is still less obvious whether populations that have experienced different demographic histories, such as African and non-African human populations, should exhibit similar relative effects in those regions. Much of the theory developed in the context of BGS has been developed under the assumption that the population is at equilibrium, and recent work has demonstrated that this assumption likely holds under changing demography if selection is strong enough (or populations are large enough) such that mutation-selection balance is maintained [49,50]. However, strong, sustained population bottlenecks may lead to violations of that assumption, and the effect of genetic drift may dominate the influence of selection at linked sites on determining patterns of genetic variation. Finally, the effect of demography on influencing patterns of diversity in regions experiencing selection at linked sites through time has also been underappreciated (although see Ref. [51] for a recent study in maize). Since most, if not all, natural populations are in a state of changing demography, differences in neutral diversity between populations within regions experiencing selection at linked sites should not only be expected, they should also be expected to change temporally as a function of each population's specific demographic history.

While little attention has been given to the potential consequences of demography on patterns of neutral variation in regions experiencing selection at linked sites (but see [52,53] for how selection at linked sites may affect the inference of demography itself), recent studies have suggested that alleles directly under natural selection experience non-linear dynamics in the context of non-equilibrium demography. For the case of purifying selection, 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 [54,55]. This can result in relative differences in deleterious allele frequencies among populations with different demographic histories. Such effects are especially apparent in populations suffering bottlenecks [56] and have been tested and observed between different human populations with founder populations exhibiting a greater proportion of non-synonymous variants relative to synonymous variants [57–59].

We hypothesized that these non-equilibrium dynamics could also perturb nearby neutral variants due to linkage. In support of our hypothesis, a recent simulation study modeling *Drosophila* observed that population bottlenecks can result in different rates of recovery of neutral genetic diversity depending on the strength of BGS [48]. Another recent study [51] analyzed neutral diversity surrounding putatively deleterious loci in domesticated versus wild maize. They found that the extreme domestication bottleneck of maize reduced the efficiency of purifying selection, which has resulted in higher diversity in regions experiencing BGS relative to neutral regions in the domesticated population compared to the wild population (which has likely experienced a much more stable demographic history). Together, these studies provide further evidence that non-equilibrium demography should have a strong effect on patterns of diversity in the presence of selection at linked sites.

To investigate the effect of non-equilibrium dynamics in regions experiencing selection at linked sites, we measure patterns of average pairwise neutral genetic diversity (π) as a function of the strength of BGS, *B* (background selection coefficient; inferred by Ref. [6]), within a global set of human populations from phase 3 of the Thousand Genomes Project (TGP) [9]. We focus on the ratio of neutral diversity in regions of strong BGS (low *B*) to regions of weak BGS (high *B*; the closest proxy available for neutral variation in humans), which we term "relative diversity." Due to the inference procedure used to infer specific *B* values in Ref. [6], there are many caveats that may plague their direct interpretation (e.g., positive selection is not modeled, the distribution of fitness effects are inconsistent with other studies, and the deleterious mutation rate exceeds the per base pair mutation rate of other studies). However, we argue that the inferred *B* values nevertheless provide a decent proxy for ranking sites from most closely linked to deleterious loci (low *B*) to most unlinked from deleterious loci (high *B*) in humans since the key parameters used to infer *B*, namely recombination rate and local density of selected sites, are fundamental for defining regions of the genome most susceptible to selection at linked sites.

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 selection at linked sites can explain the differences of relative diversity across human populations using simulations incorporating a parametric demographic model of human history [7] with and without a model of BGS. We also investigate how genetic differentiation between TGP populations is shaped by selection at linked sites by measuring F_{ST} as a function of *B*. Finally, we demonstrate that back migration from Europeans and Asians into Africa re-introduces sufficient deleterious variation to affect patterns of BGS, leading to decreased relative diversity in Africans. Our results demonstrate the strong effect that changing demography has on perturbing levels of diversity in regions experiencing selection at linked sites and have implications for population genetic studies seeking to characterize selection at linked sites across any species or population that is not at demographic equilibrium.

Results

Differential effects of selection at linked sites across human populations

We measured mean pairwise genetic diversity (π) in the autosomes (we ignore the sex chromosomes and the mitochondrial genome for all analyses) 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; population labels and groupings reported in Table L in S1 Text). A set of stringent filters, including the masking of sites inferred to be under selective sweeps, were first applied to all 20 populations to identify a high-quality set of putatively neutral sites in the genome (see Materials and Methods). Sites were then divided into quantile bins based on estimates of *B* [6]. For our initial set of analyses, we focused on the bins corresponding to the 1% of sites inferred to be under the strongest amount of BGS (i.e., sites having the lowest inferred *B* values) and the 1% of sites inferred to be under the weakest amount BGS (i.e., sites having the highest inferred *B* values). Mean diversity was normalized by divergence from rhesus macaque within these bins for each population and is shown in Fig 1A and 1B. As expected, normalized diversity was highest in African populations and lowest in East Asian populations across both 1% *B* quantile bins.

To estimate the effect that selection at linked sites has had on neutral diversity, we calculated a statistic called "relative diversity" for each population. We define relative diversity as the ratio of normalized diversity in the lowest 1% *B* bin to normalized diversity in the highest 1% *B* bin, which should capture the relative consequences of selection at linked sites within the genome. While this statistic is analogous to "π/π_0" in the BGS literature [26,60], we caution that this interpretation is not completely accurate in the context of observed data since even regions estimated to have the highest *B* values in the human genome may still experience a minimal effect of selection at linked sites. We will use "π/π_{min}" in the context of *observed* relative diversity to make clear that we are attempting to minimize selection at linked sites. Fig 1C shows that observed relative diversity was lower in non-African populations (0.348–0.365 for non-Africans, 0.396–0.408 for Africans), demonstrating that these populations have experienced a greater reduction in diversity in regions with strong selection at linked sites and also suggesting that demography may have contributed to these patterns.

To characterize these effects across a broader distribution of sites experiencing selection at linked sites, we grouped populations together according to their continental group (i.e., African, European, South Asian, and East Asian, see Table L in S1 Text for a detailed description) and estimated relative diversity at neutral sites for each of the continental groups in bins

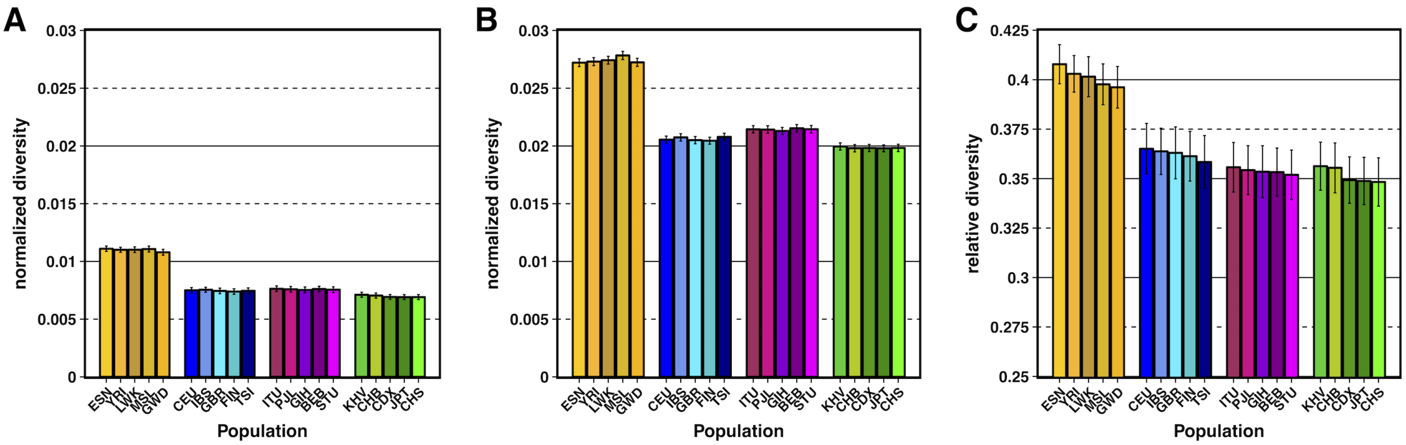

Fig 1. Normalized diversity and relative diversity for non-admixed populations of the Thousand Genomes Project (TGP). (A) Normalized diversity (π/divergence) measured across the lowest 1% *B* quantile bin (strong BGS). (B) Normalized diversity measured across the highest 1% *B* quantile bin (weak BGS). (C) Relative diversity: the ratio of normalized diversity in the lowest 1% *B* bin to normalized diversity in the highest 1% *B* bin (π/π_{min}). TGP population labels are indicated below each bar (see Table L in S1 Text 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. See S1 Table for underlying data.

https://doi.org/10.1371/journal.pgen.1007387.g001

corresponding to the lowest 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 effect on the reduction of diversity caused by selection at linked sites. 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 regions experiencing selection at linked sites also held for broader 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 further suggested an effect of demography on the diversity-reducing effect of selection at linked sites, with the strongest effects 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 (S2 and S4 Figs in S1 Text) or if we did not normalize diversity by divergence (S3 and S5 Figs in S1 Text). Patterns of relative diversity in regions of local ancestry (i.e., African, European, or Native American) across admixed TGP populations also largely recapitulated the patterns observed in their continental group counterparts across *B* quantile bins, with the largest reductions in relative diversity occurring for the Native American and European ancestral segments (S11 Fig, S1 Text).

To test if demography has influenced selection at linked sites more recently in time, we also calculated the number of singletons observed per site (normalizing by divergence and using the same set of neutral filters as was used for the calculations of π) across the lowest and highest 1% *B* quantile bins (S13 Fig in S1 Text). While it has been shown theoretically and observed empirically that selection at linked sites skews the site-frequency spectrum towards a higher proportion of singleton variants among segregating sites, the absolute number of singletons among all sites should be lower in regions of strong selection at linked sites when compared to

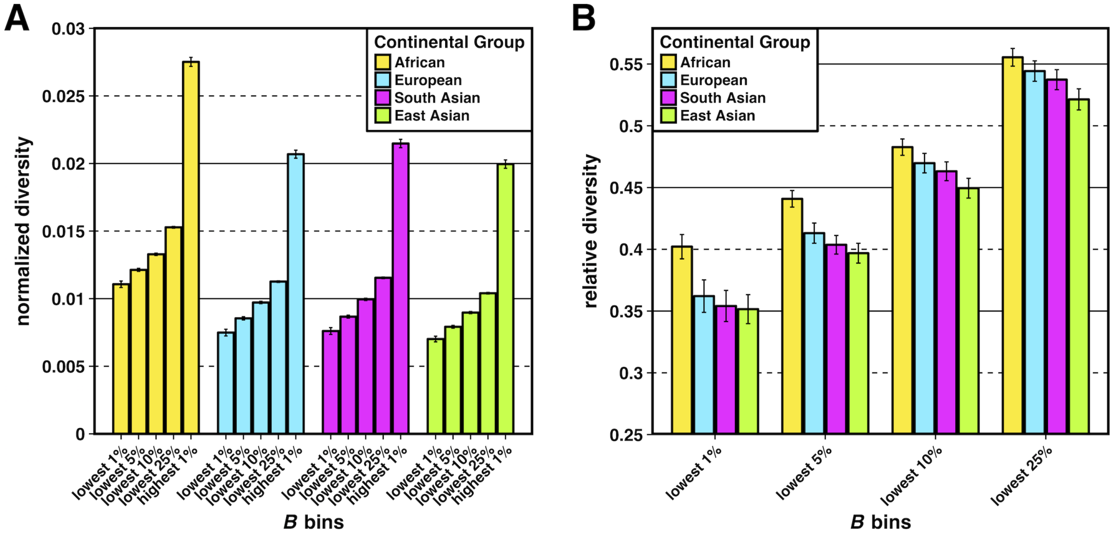

Fig 2. Normalized and relative diversity for Thousand Genomes Project (TGP) continental groups. (A) Normalized diversity (π/divergence) measured across the lowest 1%, 5%, 10% and 25% *B* quantile bins (strong BGS) and the highest 1% *B* quantile bin (weak BGS). (B) Relative diversity: the ratio of normalized diversity in the lowest *B* quantile bins (strong BGS) in (A) to normalized diversity in the highest 1% *B* quantile bin (weak BGS). Error bars represent ±1 SEM calculated from 1,000 bootstrapped datasets. See S1 Table for underlying data.

https://doi.org/10.1371/journal.pgen.1007387.g002

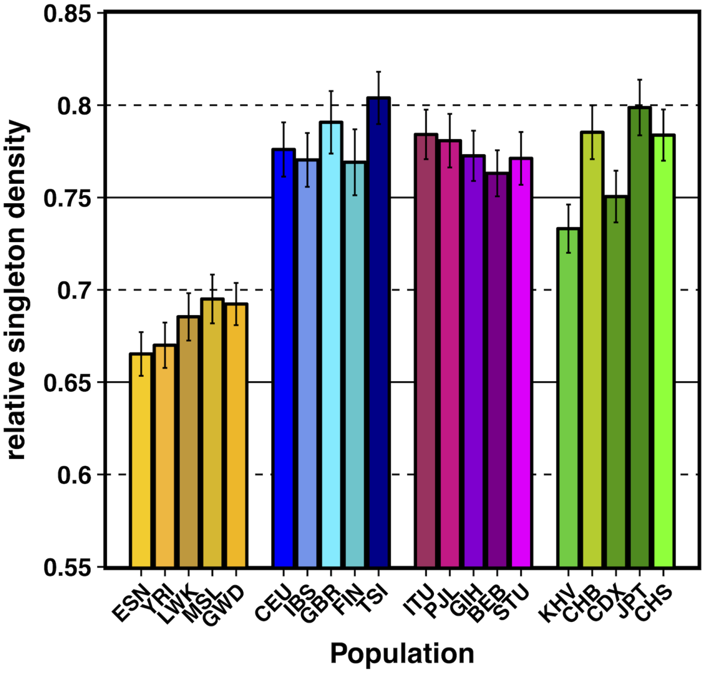

Fig 3. Relative singleton density for non-admixed populations of the Thousand Genomes Project (TGP). Relative singleton density measured by taking the ratio of singleton density in the lowest 1% *B* quantile bin to singleton density in the highest 1% *B* quantile bin ($\psi/\psi_{\min}$). Singleton density was normalized by divergence with Rhesus macaque. TGP population labels are indicated below each bar (see Table L in S1 Text 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. See S3 Table for underlying data.

https://doi.org/10.1371/journal.pgen.1007387.g003

neutral regions. In addition, since singletons are, on average, the youngest variants within the genome, they should better capture signals about very recent population history. Thus, we took the ratio of singletons observed per-site across these extreme *B* quantile bins to create a statistic called relative singleton density, which we term "$\psi/\psi_{\min}$." We accounted for differences in population sample size by first projecting down all populations to 2N = 170 (Materials and Methods). Qualitatively, our measurements of $\psi/\psi_{\min}$ showed patterns in the opposite direction to our estimates of $\pi/\pi_{\min}$, with Africans exhibiting a lower ratio of $\psi/\psi_{\min}$ when compared to non-Africans (0.665–0.695 for Africans, 0.733–0.804 for non-Africans; Fig 3). These patterns suggest that the effect of demography on regions experiencing selection at linked sites is transient, with patterns of relative diversity between populations dependent on the time frame in which they are captured (see Discussion).

Selection at linked sites has shaped patterns of population differentiation

Our results described above offered evidence that demography can affect patterns of neutral diversity in regions of selection at linked sites. Such patterns may be caused by accelerated drift in these regions, which can be amplified by demographic changes, thus leading to accelerated population differentiation. An increase in population differentiation is obvious in the context of hitchhiking (where linked neutral loci sweep to high frequency) but is also expected with BGS [61,62]. Here we quantified the magnitude of the effect of BGS on population differentiation in humans and found that population differentiation at neutral loci is indeed highly correlated with *B* (the inferred strength of BGS; Fig 4 and Table 1). Specifically, we divided the genome into 2% quantile bins based on the genome-wide distribution of *B* and measured F_{ST}

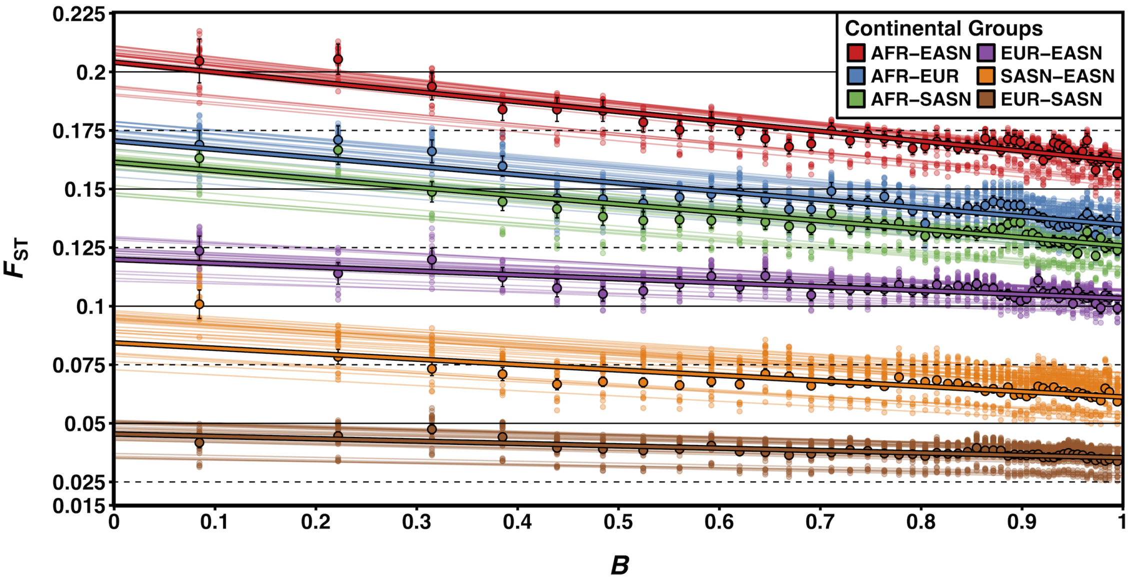

Fig 4. F_{ST} is correlated with B. F_{ST} between TGP populations measured across 2% quantile bins of B. Smaller transparent points and lines show the estimates and corresponding lines of best fit (using linear regression) for F_{ST} between every pairwise population comparison within a particular pair of continental groups (25 pairwise comparisons each). Larger opaque points and lines are mean F_{ST} estimates and lines of best fit across all population comparisons within a particular pair of continental groups. Error bars represent ±1 SEM calculated from 1,000 bootstrapped datasets.

https://doi.org/10.1371/journal.pgen.1007387.g004

in each bin for all pairs of populations from different continental groups [63]. We then 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 across all 150 population comparisons (i.e., the "Global" estimate in Table 1), B explained 26.9% of the change in F_{ST} across the most extreme B values. This result was robust to outliers [64] (Table F in S1 Text) and dominated the effects of local recombination rate (see S1 Text).

Demographic inference in putatively neutral regions of the genome

One consequence of BGS and hitchhiking in driving patterns of neutral variation within and between human populations is that demographic inference could be substantially biased

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.

https://doi.org/10.1371/journal.pgen.1007387.t001

[52,53,65]. To assess the degree of bias in the context of human data, we fit 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 selection at linked sites ($B \geq 0.994$) 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 among all populations. To make comparisons to previous studies that have used sequence data from coding regions or genes [7,22,23], which may be under strong BGS or hitchhiking effects, 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 (S1 Fig, Table A in S1 Text). Notably, inferred effective population size parameters were larger for contemporary population sizes when using four-fold degenerate synonymous sites versus ascertained neutral regions with $B \geq 0.994$, 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 neutral regions with $B \geq 0.994$ and four-fold degenerate sites, respectively). This result may stem from the expected decrease in N_e going into the past in regions of strong BGS, which can lead to inflated estimates of recent population growth [53] and has been found in simulation studies of synonymous sites under BGS [65]. Put more simply, the skew of the site-frequency spectrum towards rare variants in regions experiencing selection at linked sites [66–68] mimics a population expansion, thus leading to erroneous inference.

Simulations confirm that demographic effects can affect patterns of diversity under background selection

Using the demographic parameters inferred from neutral regions where $B \geq 0.994$, we simulated patterns of neutral diversity with and without the effects of BGS (see Materials and Methods). To measure the relative effect of BGS for each population, we 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). These results are comparable to, but not quite as extreme as, the effects we observed in the regions of the genome with the strongest effects of BGS for these population groups (Fig 1C) and may therefore reflect the weaker signatures of BGS shown in Fig 2B. 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, and the ratio π/π_0 (Fig 5). We observed that there is a distinct drop in π and π_0 at each population bottleneck experienced by non-Africans, with East Asians (who had a more severe bottleneck) experiencing a larger drop than Europeans. Fig 5C shows that the population bottlenecks experienced by non-African populations also reduces π/π_0. Surprisingly, 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 (S7 Fig in S1 Text). This finding highlights an evolutionary role that deleterious alleles can play when they are transferred across populations through migration (see Discussion).

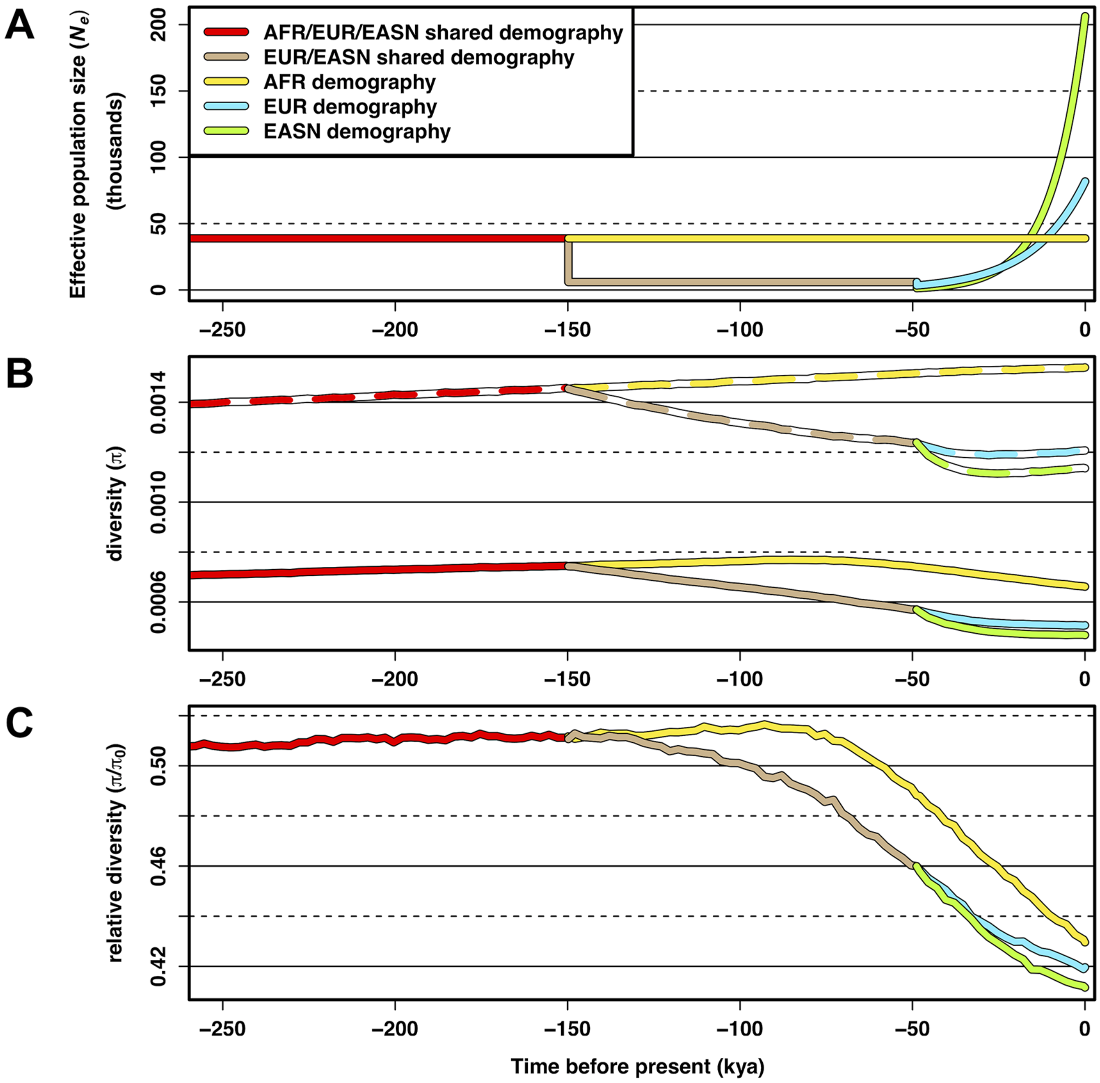

Fig 5. Simulations confirm that demographic events shape the effect of background selection (BGS). (A) Inferred demographic model from Complete Genomics TGP data showing population size changes for Africans (AFR), Europeans (EUR), and East Asians (EASN) as a function of time that was used for the simulations 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 (see S5 Table for exact values of mean π).

https://doi.org/10.1371/journal.pgen.1007387.g005

We also observed the effects of demography and BGS on singleton density by calculating ψ/ψ_0 (i.e., the ratio of singletons observed among all sites in simulations with BGS relative to simulations without BGS) and again qualitatively observed patterns similar to, but not as extreme as, our empirical estimates of ψ/ψ_{min} (S12 Fig A in S1 Text). Calculating ψ and ψ_0 through time showed that the population bottlenecks experienced by non-Africans led to strong decreases in both ψ and ψ_0, with recent expansion in these populations then leading to large, rapid recoveries. Strong decreases in ψ/ψ_0 after each population bottleneck were also observed, including a slight decrease in ψ/ψ_0 in Africans that disappeared in the simulations without migration (S12 Fig B in S1 Text). While ψ/ψ_0 for the European/East Asian ancestral population in the simulations with migration remained below that of Africans during the course of the Out-of-Africa bottleneck, we observed a rapid recovery in ψ/ψ_0 for this population in the simulations without migration (compare bottoms panels, S12 Fig A and B in S1 Text). This suggests that for populations experiencing a sustained population bottleneck, the response of singletons to the weakened intensity of BGS is quite rapid, especially when compared to patterns of π/π_0 (compare S7 Fig C to S12 Fig B bottom panel in S1 Text). However, population migration mitigates this pattern. Regardless of whether migration between populations was simulated, BGS had little effect on singleton density recovery in Europeans and Asians once population expansion occurred.

Our simulations were based on the functional density found in a 2 Mb region of the human genome with the lowest *B* values and, thus, where BGS was inferred to be strongest (chr3: 48,600,000–50,600,000). There, 20.46% of sites were either coding or conserved non-coding (see Materials and Methods) which is why the fraction of the genome experiencing deleterious mutation in our simulations of strong BGS was 0.2046. Our simulations were intended to represent the strongest effect of BGS inferred for humans. However, we did not model the specific genomic locations of coding and conserved non-coding sites in our simulations (since the structure would be specific to each region of the genome), so while the patterns we simulated are qualitatively similar to the patterns we observed in real data, there were slight quantitative differences. Since the strength of BGS is dependent upon the density of sites experiencing deleterious mutation within a given region (or more formally, *U*, which is the product of the per-site deleterious mutation rate and the number of sites experiencing deleterious mutation [69]), we simulated weaker effects of BGS by reducing the fraction of sites experiencing purifying selection while keeping the distribution of selective effects constant (see Materials and Methods). When the fraction of sites experiencing selection was decreased 2–4 fold in our simulations, we continued to observe a stepwise decrease in π/π_0 while maintaining the specific rank order of African, followed by European, and then East Asian populations (S8 Fig in S1 Text). As expected, π/π_0 increased for all populations as the fraction of sites that were simulated as deleterious decreased ($\pi/\pi_0 = 0.641$ vs. 0.802, 0.62 vs. 0.777, and 0.611 vs. 0.777 for AFR, EUR, and EASN when the fraction of sites experiencing selection was reduced to 0.1023 and 0.05115, respectively). These simulations resulted in π/π_0 values much larger than the observed values of π/π_{min} (Figs 1C and 2B).

Discussion

In our analyses of thousands of genomes from globally distributed human populations, we have confirmed that the processes of demography and selection at linked sites influence neutral variation across the genome. While this observation is not unexpected, we have characterized the dynamic consequence of non-equilibrium demographic processes in regions experiencing selection at linked sites in humans. We find that demography (particularly population bottlenecks) can amplify the consequences of selection at linked sites. 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-biased gene conversion, and the potential mutagenic effects of recombination hotspots. None of these factors qualitatively affected our results. However, because divergence itself is not independent of BGS [70], biases may arise 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, thereby leading to a decrease in overall divergence in those regions. While we attempt to limit the contribution of such biases by using a more diverged primate species (rhesus macaque), our calculations of π/π_{min} show that our results are actually conservative when normalizing by divergence (π/π_{min} for AFR is 0.373 without the divergence step and 0.402 with the divergence step). Moreover, the population comparisons we make should be robust to such biases since all human populations are equally diverged from rhesus macaque and estimates of *B* are constant across populations.

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* unlikely to be precisely inferred. In fact, the *B* values provided by McVicker et al. range from 0 to 1, suggesting that some regions of the genome should be essentially devoid of diversity (but we do not observe this to be the case). Since our own analyses show that relative diversity has a lower bound at only ~0.35 in humans, the exact value of *B* itself should not be taken at face value. Rather, our primary motivation for using *B* was to ascertain regions that should be on the extreme ends of the genome-wide distribution of regions experiencing selection at linked sites, for which *B* should provide a good assessment. A study by Comeron [32] 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 regions experiencing the weakest and strongest effects of selection at linked sites within the human genome, with model misspecification unlikely to change our empirical results.

While the effects of selection at linked sites captured in our analyses could in principle include the consequences of positive selection (such as soft-sweeps and classic selective sweeps), we applied stringent filters to remove any such regions before our analyses (Materials and Methods, S1 Appendix). Nonetheless, we cannot rule out all contributions from hitchhiking to our results. In fact, our simulations of BGS fail to capture the complete effects of selection at linked sites on reducing π/π_0 in different human populations (compare Figs 1C and 5C), and the additional contribution of hitchhiking to humans may explain this discrepancy (though proper modeling of linkage among deleterious loci could also improve our quantitative results). Further investigation will be needed to in order to more fully characterize the effect demography has on influencing the various modes of selection at linked sites, including BGS, selective sweeps, and interference selection [67].

Non-equilibrium demography has also been of recent interest in regards to its effect on patterns of deleterious variation across human populations (often referred to as genetic load), with initial work showing that non-African populations have a greater proportion of segregating non-synonymous deleterious variants compared to synonymous variants [57]. Similar results in human founder populations [58,71], *Arabadopsis* [72], and domesticated species such as dogs [12] and sunflowers [73] further demonstrate the pervasive effect that demography has on influencing the relative amount of deleterious variation across a variety of populations and species. Since BGS is a function of deleterious variation, it is not surprising that we also witness differences in π/π_{min} across human populations that have experienced different demographic histories. These effects are probably ubiquitous across other species as well. However, there has been recent contention about whether the previously described patterns of increased deleterious variants are driven by a decrease in the efficacy of natural selection (thus resulting in increased genetic load) or are solely artifacts of the response of deleterious variation to demographic change [59,74–77]. Recently, Koch et al. [56] investigated the temporal dynamics of demography on selected sites within humans 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 both BGS and hitchhiking, which skew the site frequency spectrum of linked neutral mutations towards rare variants [26,69,78,79], we also expect a transient decrease in diversity as low-frequency variants are lost quickly during a population contraction. Indeed, as evident from our simulations of BGS and demography, immediately after a population bottleneck, rapid losses in singleton density can occur, leading to transient decreases in ψ/ψ_0. However, the recovery in singleton density is also quite rapid, while the recovery in π and π/π_0 is quite slow. This is due to the fact that higher frequency variants, which contribute a greater amount to π, take a longer amount of time to recover after a population contraction compared to lower-frequency variants such as singletons. Furthermore, Koch et al. also demonstrated that the effect of demography on diversity is only temporary and that long-term diversity at selected sites approaches greater values once equilibrium is reached.

The temporal effects of non-equilbrium demographics on patterns of π/π_{min} and ψ/ψ_{min} may also explain the conflicting results obtained in a similar study of selection at linked sites in teosinte and its domesticated counterpart, maize [51]. In that study, the authors observed that π/π_{min} was higher in maize, which underwent a population bottleneck during domestication (no bottleneck event was inferred for the teosinte population) but that ψ/ψ_{min} was lower. This result is contrary to what we observed qualitatively between non-African and African human populations. However, the demographic models that have been inferred for maize and humans are quite different. Maize is inferred to have had a recent, major domestication bottleneck that was essentially instantaneous and followed by rapid exponential growth [51]. In contrast, demographic models for non-African humans suggest a much more distant bottleneck that was sustained over a longer period of time, and only recently have non-African populations experienced rampant growth (coinciding with the advent of agriculture). Thus, depending on how far in the past a particular demographic event occurred and how strong the population size change was, different qualitative observations of π/π_{min} and ψ/ψ_{min} will result. Importantly, our simulations show changing values of these statistics through time (Fig 5, S12 Fig in S1 Text), which can lead to different qualitive results that are dependent on the time frame in which populations are observed.

Broadly, our results show that contemporary patterns of neutral diversity cannot easily be attributable to contemporary forces of selection but instead may be exhibiting signatures that are still dominated by older demographic events. Interestingly though, our simulations reveal an additional factor that can influence the effect of BGS within populations–migration between populations. We observe that the exchange of deleterious variants from populations that have experienced extensive bottlenecks to populations with a more stable demography can magnify the strength of selection at linked sites. In particular, our simulations show that both π/π_0 and ψ/ψ_0 decrease in Africans despite the fact that they are inferred to have been constant in size in their recent evolutionary history (Fig 5B). These patterns disappear when migration is removed (S7 Fig, S12 Fig B in S1 Text); however, more work is needed to definitively test this.

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. These studies have shown that population contractions have affected 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 humans [80] and later modeled and shown in other species [81], with the explanation 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 [82–84]. These studies, plus others [85], have also shown that there is strong evidence for a more dominant effect of selection at linked sites on the sex chromosomes relative to the autosomes in humans.

Since selection at linked sites 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 [47,86]. 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 [2,86–88]. Yet, as we show in humans, it is likely that patterns of neutral diversity are in fact jointly affected by both of these forces, magnifying one another to deplete levels of diversity beyond what is expected by either one independently. This may play an even larger role in higher N_e species such as *Drosophila*, where the overall distribution of *B* was inferred to be even smaller (i.e., exhibiting stronger BGS) than in humans [32]. 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 [89].

Finally, our results also have implications for 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 selection at linked sites 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 affects power for disease association tests [90]. Our results indicate that this 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 [91]. It should also further motivate investigators to simultaneously account for demography and selection at linked sites when performing tests to uncover disease variants within the genome [90,92,93].

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.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/. vcftools (v0.1.12a) [94] 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.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/. After filtering out related individuals via pedigree analyses, we analyzed 53 YRI, 64 CEU, and 62 CHS samples (Table B in S1 Text). 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 C in S1 Text) from 3 studies [95–97] 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 [60], which treats its effects as a simple reduction in N_e at neutral sites as a function of their recombination distance from conserved and exonic loci, the strength of purifying selection at those 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 lowest 1%, 5%, 10%, and 25% of the genome-wide distribution of *B* and within the highest1% 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. These filters were applied to all samples in phase 3 TGP (all filters are in build hg19) for all sets of analyses (see Table D in S1 Text 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.1000genomes.ebi.ac.uk/vol1/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 [98] 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) [99] 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 [100] 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.1000genomes.ebi.ac.uk/vol1/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 selection.
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 [101] 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 [102]; S1 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 [103] 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 $\geq$ 10 cM/Mb in the 1000G OMNI genetic maps for non-admixed populations (downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20130507_omni_recombination_rates/) and the HapMap II genetic map [104] were removed. 1.5 kb flanking regions surrounding the center of hotspots identified by Ref. [105] (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 [106] with Gencode V19 annotations. ANNOVAR and Gencode V19 annotations were also used to gather an autosome-wide set of four-fold degenerate sites (i.e., all possible sites, regardless of being polymorphic), 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 [107] 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 autosomal sites in the top 1% of B ($B \geq 0.994$) was also applied in order to mitigate potential biases in inference due to BGS [53,65] or other forms of selection at linked sites [52]. 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. [108]) 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 [109]. For the model of background selection, the recombination rate, ρ, and the fraction of the genome experiencing deleterious mutation were 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} (raw recombination rate of 8.19 x 10^{-10}) was calculated for this region using the HapMap II GRCh37 genetic map [104]. For ascertaining the fraction of sites experiencing deleterious mutation, 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, the total fraction of deleterious sites, 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 fraction of deleterious sites described above, 20.46% of sites in the two 1 Mb flanking regions were simulated as being deleterious. The mutation rate in our simulations for the deleterious sites and for neutral sites were both set to 1.66 x 10^{-8} [108]. Two distributions of fitness effects were used for the deleterious sites, with 66.06% of deleterious sites using the gamma distribution (parameters: mean = α/β, variance = α/β^2) of fitness effects inferred across conserved non-coding regions by Ref. [110] ($\beta = 0.0415$, $\alpha = 0.00515625$) and 33.94% of deleterious sites using the gamma distribution of fitness effects inferred across coding regions by Ref. [5] ($\beta = 0.184$, $\alpha = 0.00040244$). Gamma distribution parameters were scaled to the ancestral population size of the demographic models used in Refs. [5,110]. Their unscaled values are ($\beta = 0.0415$, $\alpha = 80.11$) and ($\beta = 0.184$, $\alpha = 6.25$) for conserved non-coding regions and coding regions, respectively. The relative number of non-coding "functional" sites and coding exons described above determined the relative number of sites receiving each distribution of fitness effects in our simulations. An example of the SFS_CODE command for our simulations is in S1 Text. To simulate varying levels of background selection strength, different total fractions of our original calculated deleterious fraction of 0.2046 were used (i.e., 5%, 10%, 25%, 50%, and 100% of 0.2046). However, the same relative percentage of non-coding and coding sites and mutation rate were used. These different simulated fractions of deleterious sites resulted in a reduced total deleterious mutation rate, *U*, which is the product of the per-site deleterious mutation rate and the total number of sites experiencing deleterious mutation [69]. Thus, weaker effects of BGS were simulated. To simulate only the effects of demography without background selection, only the 30 kb neutral region was simulated. 2,000 independent simulations were conducted for each particular set of the deleterious site fraction (2,000 x 6 = 12,000 total). Simulations output population genetic information for 100 samples every 100 generations and also at each generation experiencing a population size change (22,117 total generations were simulated), from which mean pairwise nucleotide diversity (π) and singleton density (ψ) was calculated across the 2,000 simulations.

Population-specific calculations of diversity and singleton density

Mean pairwise genetic diversity (π) and singleton density (ψ) 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, for π, across 4 broad populations that grouped the 20 non-admixed populations together by continent (African, European, South Asian, and East Asian, see Table L in S1 Text). Additionally, only regions of the genome passing the 13-filter set were used in the calculations of π and ψ (see Table D in S1 Text for total number of Mb used in diversity calculations for each *B* quantile). When calculating ψ for each non-admixed phase 3 TGP population, the site-frequency spectrum was first projected down to 2N = 170 samples (the number of chromosomes in MSL, the smallest phase 3 population sample) using a hypergeometric distribution [7] from each population's full (unfolded) site-frequency spectrum. This allowed for unbiased comparisons of singleton density between all populations. Additionally, when identifying singletons for calculating ψ, only sites annotated with high confidence calls for polarizing ancestral and derived states were used when creating the unfolded site-frequency spectrum. These high confidence sites were ascertained from the GRCh37 ancestral sequence (downloaded from ftp://ftp.ensembl.org/pub/release-71/fasta/ancestral_alleles/homo_sapiens_ancestor_GRCh37_e71.tar.bz2). 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 π and ψ calculations (see "Normalization of diversity and divergence calculations with Rhesus macaque" below).

Normalization of diversity/singleton density 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 and singleton density 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 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. [63] 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 [111], 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 2% quantile bins of B (based on the genome-wide distribution of B) 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). Simple linear regression was performed with the model $F_{ST} = \beta_0 + \beta_1 B + \varepsilon$. The mean of the bounds defining each quantile bin was used when defining the explanatory variables for the regression. Linear regression, robust linear regression [64], 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.

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 B 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.

Supporting information

S1 File. SFS_CODE implementation used for simulations of human demography under a model of BGS with two negative gamma distributions of fitness effects.
(GZ)

S1 Appendix. Soft sweep detection and implementation in selscan v1.2.0.
(PDF)

S1 Text. Admixed population analyses, linear regression of F_{ST} on recombination-rate, SFS_CODE simulation commands, supplemental tables, and supplemental figures.
(PDF)

S1 Table. Diversity (π), normalized diversity (π/divergence), and relative diversity for phase 3 TGP populations and continental groups. Population and continental group labels are given in the first column and their corresponding information (described in the second column) is given on each row. Each population/continental group has information corresponding to its observed per-site diversity for the lowest 1%, 5%, 10% and 25% *B* quantile bins and the highest 1% *B* quantile bin (i.e., rows with descriptor 'pi' [or 'pi_D' if normalized by divergence] in the second column) and the ratio of the lowest 1%, 5%, 10% and 25% *B* quantile bins to the highest 1% *B* quantile bin (i.e., rows with descriptor 'pi_pimin' [or 'pi_pimin_D' if normalized by divergence] in the second column). Rows that have 'SEM' in the second column contain information that corresponds to calculated standard errors of the mean.
(TXT)

S2 Table. Diversity (π), normalized diversity (π/divergence), and relative diversity while controlling for GC-biased gene conversion and recombination hotspots for phase 3 TGP populations and continental groups. Population and continental group labels are given in the first column and corresponding information is given on each row. Before calculating diversity, regions of GC-biased gene conversion and recombination hotspots were filtered out. The table structure is identical to that of S1 Table.
(TXT)

S3 Table. Singleton density (ψ), normalized singleton density (ψ/divergence), and relative singleton density for phase 3 TGP populations. Population labels are given in the first column and corresponding information is given on each row. The table structure is identical to that of S1 Table.
(TXT)

S4 Table. Diversity (heterozygosity), normalized diversity (heterozygosity/divergence), and relative diversity for phase 3 TGP continental groups and local ancestry segments. Continental group and ancestry labels are given in the first column and corresponding information is given on each row. The table structure is identical to that of S1 Table.
(TXT)

S5 Table. Simulation result calculations of diversity (π) from a model of human demography with BGS using various deleterious site fractions with and without migration. Values represent mean diversity (π) calculated from 2000 total simulations for 100 samples every 100 generations (or immediately after a demographic event).
(TXT)

S6 Table. Simulation result calculations of singleton density (ψ) from a model of human demography with BGS using various deleterious site fractions with and without migration. Values represent mean singleton density (ψ) calculated from 2000 total simulations for 100 samples every 100 generations (or immediately after a demographic event).
(TXT)

Acknowledgments

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

Author Contributions

Conceptualization: Raul Torres, Zachary A. Szpiech, Ryan D. Hernandez.

Formal analysis: Raul Torres, Zachary A. Szpiech.

Funding acquisition: Raul Torres.

Investigation: Raul Torres.

Methodology: Raul Torres, Zachary A. Szpiech, Ryan D. Hernandez.

Project administration: Ryan D. Hernandez.

Resources: Ryan D. Hernandez.

Software: Zachary A. Szpiech, Ryan D. Hernandez.

Supervision: Ryan D. Hernandez.

Writing – original draft: Raul Torres.

Writing – review & editing: Raul Torres, Ryan D. Hernandez.

References

1. Cutter AD, Payseur BA. Genomic signatures of selection at linked sites: unifying the disparity among species. Nat Rev Genet. 2013; 14: 262–274. https://doi.org/10.1038/nrg3425 PMID: 23478346

2. Ellegren H, Galtier N. Determinants of genetic diversity. Nat Rev Genet. 2016; 17: 422–433. https://doi.org/10.1038/nrg.2016.58 PMID: 27265362

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. 2002; 419: 832–837. https://doi.org/10.1038/nature01140 PMID: 12397357

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. 2005; 102: 7882–7887. https://doi.org/10.1073/pnas.0502300102 PMID: 15905331

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. https://doi.org/10.1371/journal.pgen.1000083 PMID: 18516229

6. McVicker G, Gordon D, Davis C, Green P. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. 2009; 5: e1000471. https://doi.org/10.1371/journal.pgen.1000471 PMID: 19424416

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. https://doi.org/10.1371/journal.pgen.1000695 PMID: 19851460

8. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011; 475: 493–496. https://doi.org/10.1038/nature10231 PMID: 21753753

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. https://doi.org/10.1038/nature15393 PMID: 26432245

10. Lack JB, Lange JD, Tang AD, Corbett-Detig RB, Pool JE. A thousand fly genomes: An expanded *Drosophila* genome nexus. Mol Biol Evol. 2016; 33: 3308–3313. https://doi.org/10.1093/molbev/msw195 PMID: 27687565

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. Science. 2009; 324: 528–532. https://doi.org/10.1126/science.1167936 PMID: 19390050

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. https://doi.org/10.1073/pnas.1512501113 PMID: 26699508

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. 2007; 3: 1745–1756. https://doi.org/10.1371/journal.pgen.0030163 PMID: 17907810

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. https://doi.org/10.1038/365548a0 PMID: 8413609

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. 2005; 15: 790–799. https://doi.org/10.1101/gr.3541005 PMID: 15930491

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. 2005; 22: 2119–2130. https://doi.org/10.1093/molbev/msi207 PMID: 15987874

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: 240–243. https://doi.org/10.1126/science.1140462 PMID: 17431170

18. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, 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. 2005; 102: 15942–15947. https://doi.org/10.1073/pnas.0507611102 PMID: 16243969

19. Henn BM, Cavalli-Sforza LL, Feldman MW. The great human expansion. Proc Natl Acad Sci. 2012; 109: 17758–17764. https://doi.org/10.1073/pnas.1212380109 PMID: 23077256

20. Charlesworth B. Effective population size and patterns of molecular evolution and variation. Nat Rev Genet. 2009; 10: 195–205. https://doi.org/10.1038/nrg2526 PMID: 19204717

21. Nei M, Maruyama T, Chakraborty R. The bottleneck effect and genetic variability in populations. Evolution. 1975; 29: 1–10. https://doi.org/10.1111/j.1558-5646.1975.tb00807.x PMID: 28563291

22. 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. https://doi.org/10.1073/pnas.1019276108 PMID: 21730125

23. 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. 2012; 337: 64–69. https://doi.org/10.1126/science.1219240 PMID: 22604720

24. Charlesworth D. Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet. 2006; 2: 379–384. https://doi.org/10.1371/journal.pgen.0020064 PMID: 16683038

25. Maynard Smith J, Haigh J. The hitch-hiking effect of a favourable gene. Genet Res. 1974; 23: 23–35. https://doi.org/10.1017/S0016672308009579 PMID: 4407212

26. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993; 134: 1289–1303. PMID: 8375663

27. Kim Y, Stephan W. Joint effects of genetic hitchhiking and background selection on neutral variation. Genetics. 2000; 155: 1415–1427. PMID: 10880499

28. Begun DJ, Aquadro CF. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature. 1992; 356: 519–520. https://doi.org/10.1038/356519a0 PMID: 1560824

29. Charlesworth B. Background selection and patterns of genetic diversity in *Drosophila melanogaster*. Genet Res. 1996; 68: 131–149. https://doi.org/10.1017/S0016672300034029 PMID: 8940902

30. Andolfatto P. Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome. Genome Res. 2007; 17: 1755–1762. https://doi.org/10.1101/gr.6691007 PMID: 17989248

31. Sella G, Petrov DA, Przeworski M, Andolfatto P. Pervasive natural selection in the *Drosophila* genome? PLoS Genet. 2009; 5: e1000495. https://doi.org/10.1371/journal.pgen.1000495 PMID: 19503600

32. Comeron JM. Background selection as baseline for nucleotide variation across the *Drosophila* genome. PLoS Genet. 2014; 10: e1004434. https://doi.org/10.1371/journal.pgen.1004434 PMID: 24968283

33. 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. PLoS Genet. 2016; 12: e1006130. https://doi.org/10.1371/journal.pgen.1006130 PMID: 27536991

34. 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. 2012; 29: 675–687. https://doi.org/10.1093/molbev/msr225 PMID: 21917724

35. Xu X, Liu X, Ge S, Jensen JD, 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: 105–111. https://doi.org/10.1038/nbt.2050 PMID: 22158310

36. 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: 285–290. https://doi.org/10.1038/ng.1050 PMID: 22286215

37. Cutter AD, Payseur BA. Selection at linked sites in the partial selfer *Caenorhabditis elegans*. Mol Biol Evol. 2003; 20: 665–673. https://doi.org/10.1093/molbev/msg072 PMID: 12679551

38. 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. https://doi.org/10.1101/gr.3413205 PMID: 16140989

39. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006; 4: 0446–0458. https://doi.org/10.1371/journal.pbio.0040072 PMID: 16494531

40. Cai JJ, Macpherson JM, Sella G, Petrov DA. Pervasive hitchhiking at coding and regulatory sites in humans. PLoS Genet. 2009; 5: e1000336. https://doi.org/10.1371/journal.pgen.1000336 PMID: 19148272

41. Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, et al. Classic selective sweeps were rare in recent human evolution. Science. 2011; 331: 920–924. https://doi.org/10.1126/science.1198878 PMID: 21330547

42. 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. https://doi.org/10.1371/journal.pgen.1002326 PMID: 22022285

43. Alves I, Šrámková Hanulová A, Foll M, Excoffier L. Genomic data reveal a complex making of humans. PLoS Genet. 2012; 8: e1002837. https://doi.org/10.1371/journal.pgen.1002837 PMID: 22829785

44. 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. https://doi.org/10.1534/genetics.112.144071 PMID: 22960214

45. Enard D, Messer PW, Petrov DA. Genome-wide signals of positive selection in human evolution. Genome Res. 2014; 24: 885–895. https://doi.org/10.1101/gr.164822.113 PMID: 24619126

46. Bank C, Ewing GB, Ferrer-Admettla A, Foll M, Jensen JD. Thinking too positive? Revisiting current methods of population genetic selection inference. Trends Genet. 2014; 30: 540–546. https://doi.org/10.1016/j.tig.2014.09.010 PMID: 25438719

47. Corbett-Detig RB, Hartl DL, Sackton TB. Natural selection constrains neutral diversity across a wide range of species. PLoS Biol. 2015; 13: e1002112. https://doi.org/10.1371/journal.pbio.1002112 PMID: 25859758

48. Comeron JM. Background selection as null hypothesis in population genomics: insights and challenges from *Drosophila* studies. Philos Trans R Soc B. 2017; 372: 20160471. https://doi.org/10.1098/rstb.2016.0471 PMID: 29109230

49. Zeng K. A coalescent model of background selection with recombination, demography and variation in selection coefficients. Heredity. 2013; 110: 363–371. https://doi.org/10.1038/hdy.2012.102 PMID: 23188176

50. Nicolaisen LE, Desai MM. Distortions in genealogies due to purifying selection and recombination. Genetics. 2013; 195: 221–230. https://doi.org/10.1534/genetics.113.152983 PMID: 23821597

51. Beissinger TM, Wang L, Crosby K, Durvasula A, Hufford MB, Ross-Ibarra J. Recent demography drives changes in linked selection across the maize genome. Nat Plants. 2016; 2: 16084. https://doi.org/10.1038/nplants.2016.84 PMID: 27294617

52. Schrider DR, Shanku AG, Kern AD. Effects of linked selective sweeps on demographic inference and model selection. Genetics. 2016; 204: 1207–1223. https://doi.org/10.1534/genetics.116.190223 PMID: 27605051

53. Ewing GB, Jensen JD. The consequences of not accounting for background selection in demographic inference. Mol Ecol. 2016; 25: 135–141. https://doi.org/10.1111/mec.13390 PMID: 26394805

54. Pennings PS, Kryazhimskiy S, Wakeley J. Loss and recovery of genetic diversity in adapting populations of HIV. PLoS Genet. 2014; 10: e1004000. https://doi.org/10.1371/journal.pgen.1004000 PMID: 24465214

55. Brandvain Y, Wright SI. The limits of natural selection in a nonequilibrium world. Trends Genet. 2016; 32: 201–210. https://doi.org/10.1016/j.tig.2016.01.004 PMID: 26874998

56. Koch E, Novembre J. A temporal perspective on the interplay of demography and selection on deleterious variation in humans. G3. 2017; 7: 1027–1037. https://doi.org/10.1534/g3.117.039651 PMID: 28159863

57. 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. https://doi.org/10.1038/nature06611 PMID: 18288194

58. 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. PLoS Genet. 2013; 9: e1003815. https://doi.org/10.1371/journal.pgen.1003815 PMID: 24086152

59. Simons YB, Sella G. The impact of recent population history on the deleterious mutation load in humans and close evolutionary relatives. Curr Opin Genet Dev. 2016; 41: 150–158. https://doi.org/10.1016/j.gde.2016.09.006 PMID: 27744216

60. Nordborg M, Charlesworth B, Charlesworth D. The effect of recombination on background selection. Genet Res. 1996; 67: 159–174. https://doi.org/10.1017/S0016672300033619 PMID: 8801188

61. 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. https://doi.org/10.1017/S0016672397002954 PMID: 9449192

62. Hu XS, He F. Background selection and population differentiation. J Theor Biol. 2005; 235: 207–219. https://doi.org/10.1016/j.jtbi.2005.01.004 PMID: 15862590

63. Bhatia G, Patterson N, Sankararaman S, Price AL. Estimating and interpreting F_{ST}: The impact of rare variants. Genome Res. 2013; 23: 1514–1521. https://doi.org/10.1101/gr.154831.113 PMID: 23861382

64. Yu C, Yao W. Robust linear regression: A review and comparison. Commun Stat—Simul Comput. 2017; 46: 6261–6282. https://doi.org/10.1080/03610918.2016.1202271

65. Messer PW, Petrov DA. Frequent adaptation and the McDonald-Kreitman test. Proc Natl Acad Sci. 2013; 110: 8615–8620. https://doi.org/10.1073/pnas.1220835110 PMID: 23650353

66. Neher RA, Hallatschek O. Genealogies of rapidly adapting populations. Proc Natl Acad Sci. 2013; 110: 437–442. https://doi.org/10.1073/pnas.1213113110 PMID: 23269838

67. Good BH, Walczak AM, Neher RA, Desai MM. Genetic diversity in the interference selection limit. PLoS Genet. 2014; 10: e1004222. https://doi.org/10.1371/journal.pgen.1004222 PMID: 24675740

68. Cvijović I, Good BH, Desai MM. The effect of strong purifying selection on genetic diversity. bioRxiv. 2017; https://doi.org/10.1101/211557

69. Charlesworth B. The effects of deleterious mutations on evolution at linked sites. Genetics. 2012; 190: 5–22. https://doi.org/10.1534/genetics.111.134288 PMID: 22219506

70. Phung TN, Huber CD, Lohmueller KE. Determining the effect of natural selection on linked neutral divergence across species. PLoS Genet. 2016; 12: e1006199. https://doi.org/10.1371/journal.pgen.1006199 PMID: 27508305

71. 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. PLoS Genet. 2014; 10: e1004494. https://doi.org/10.1371/journal.pgen.1004494 PMID: 25078778

72. 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. https://doi.org/10.1038/ng.911 PMID: 21874002

73. 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. 2015; 32: 2273–2283. https://doi.org/10.1093/molbev/msv106 PMID: 25939650

74. Balick DJ, Do R, Cassa CA, Reich D, Sunyaev SR. Dominance of deleterious alleles controls the response to a population bottleneck. PLoS Genet. 2015; 11: e1005436. https://doi.org/10.1371/journal.pgen.1005436 PMID: 26317225

75. 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. https://doi.org/10.1038/ng.3186 PMID: 25581429

76. Simons YB, Turchin MC, Pritchard JK, Sella G. The deleterious mutation load is insensitive to recent population history. Nat Genet. 2014; 46: 220–224. https://doi.org/10.1038/ng.2896 PMID: 24509481

77. Gravel S. When is selection effective? Genetics. 2016; 203: 451–462. https://doi.org/10.1534/genetics.115.184630 PMID: 27010021

78. Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics. 1995; 140: 783–796. PMID: 7498754

79. Stephan W. Genetic hitchhiking versus background selection: the controversy and its implications. Philos Trans R Soc B. 2010; 365: 1245–1253. https://doi.org/10.1098/rstb.2009.0278 PMID: 20308100

80. 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. https://doi.org/10.1093/oxfordjournals.molbev.a026175 PMID: 10406117

81. Pool JE, Nielsen R. Population size changes reshape genomic patterns of diversity. Evolution. 2007; 61: 3001–3006. https://doi.org/10.1111/j.1558-5646.2007.00238.x PMID: 17971168

82. 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. 2011; 43: 741–743. https://doi.org/10.1038/ng.877 PMID: 21775991

83. 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. https://doi.org/10.1016/j.ajhg.2014.04.011 PMID: 24836452

84. Wilson Sayres MA, Lohmueller KE, Nielsen R. Natural selection reduced diversity on human Y chromosomes. PLoS Genet. 2014; 10: e1004064. https://doi.org/10.1371/journal.pgen.1004064 PMID: 24415951

85. 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: 830–831. https://doi.org/10.1038/ng.651 PMID: 20802480

86. 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. 2012; 10: e1001388. https://doi.org/10.1371/journal.pbio.1001388 PMID: 22984349

87. Vucetich JA, Waite TA, Nunney L. Fluctuating population size and the ratio of effective to census population size. Evolution. 1997; 51: 2017–2021. https://doi.org/10.1111/j.1558-5646.1997.tb05123.x PMID: 28565105

88. Coop G. Does linked selection explain the narrow range of genetic diversity across species? bioRxiv. 2016; https://doi.org/10.1101/042598

89. Lewontin RC. The genetic basis of evolutionary change. New York and London: Columbia University Press; 1974.

90. 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. https://doi.org/10.1002/gepi.21866 PMID: 25417809

91. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, et al. Human demographic history impacts genetic risk prediction across diverse populations. Am J Hum Genet. 2017; 100: 635–649. https://doi.org/10.1016/j.ajhg.2017.03.004 PMID: 28366442

92. Maher MC, Uricchio LH, Torgerson DG, Hernandez RD. Population genetics of rare variants and complex diseases. Hum Hered. 2012; 74: 118–128. https://doi.org/10.1159/000346826 PMID: 23594490

93. 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. 2016; 26: 863–873. https://doi.org/10.1101/gr.202440.115 PMID: 27197206

94. 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. https://doi.org/10.1093/bioinformatics/btr330 PMID: 21653522

95. 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. 2016; 113: E440–449. https://doi.org/10.1073/pnas.1510805112 PMID: 26712023

96. Kidd JM, Sharpton TJ, Bobo D, Norman PJ, Martin AR, Carpenter ML, et al. Exome capture from saliva produces high quality genomic and metagenomic data. BMC Genomics. 2014; 15: 262. https://doi.org/10.1186/1471-2164-15-262 PMID: 24708091

97. 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. https://doi.org/10.1038/ncomms6692 PMID: 25471224

98. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010; 20: 110–121. https://doi.org/10.1101/gr.097857.109 PMID: 19858363

99. 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. https://doi.org/10.1101/gr.3715005 PMID: 16024819

100. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489: 57–74. https://doi.org/10.1038/nature11247 PMID: 22955616

101. 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: 1005–1017. https://doi.org/10.1101/gr.187101 PMID: 11381028

102. Szpiech ZA, Hernandez RD. selscan: An efficient multithreaded program to perform EHH-based scans for positive selection. Mol Biol Evol. 2014; 31: 2824–2827. https://doi.org/10.1093/molbev/msu211 PMID: 25015648

103. 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. PLoS Genet. 2013; 9: e1003684. https://doi.org/10.1371/journal.pgen.1003684 PMID: 23966869

104. 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: 851–861. https://doi.org/10.1038/nature06258 PMID: 17943122

105. Pratto F, Brick K, Khil P, Smagulova F, Petukhova G V, Camerini-Otero RD. Recombination initiation maps of individual human genomes. Science. 2014; 346: 1256442. https://doi.org/10.1126/science.1256442 PMID: 25395542

106. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010; 38: e164. https://doi.org/10.1093/nar/gkq603 PMID: 20601685

107. 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. https://doi.org/10.1093/molbev/msm149 PMID: 17656634

108. 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. https://doi.org/10.1016/j.ajhg.2015.10.006 PMID: 26581902

109. Hernandez RD. A flexible forward simulator for populations subject to selection and demography. Bioinformatics. 2008; 24: 2786–2787. https://doi.org/10.1093/bioinformatics/btn522 PMID: 18842601

110. 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: e1000592. https://doi.org/10.1371/journal.pgen.1000592 PMID: 19662163

111. Hudson RR, Slatkin M, Maddison WP. Estimation of levels of gene flow from DNA sequence data. Genetics. 1992; 132: 583–589. PMID: 1427045