Title
The Effects of Background Selection and Demography on Patterns of Neutral Variation within the Genome

Author
Torres, Raul

Publication Date
2018

Peer reviewed|Thesis/dissertation

The Effects of Background Selection and Demography on Patterns of Neutral Variation within the Genome

by
Raul Torres

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Biomedical Sciences

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Seielstad, Mark
54A1F0A578614C2...

Seielstad, Mark
Chair

DocuSigned by:
Ryan Hernandez
[illegible]

Ryan Hernandez

DocuSigned by:
Jeffrey Wall
C944684198754B9...

Jeffrey Wall

Committee Members

To my mother, Carol,
for her unconditional dedication, love, and support.

ACKNOWLEDGEMENTS

They say it takes a village to raise a child. Well, I can certainly say it takes one to get a PhD candidate through graduate school. I have many people to thank for supporting me along this hard, trying, and gratifying journey. First and foremost, I would like to thank Ryan Hernandez, my PhD advisor who took a chance on me (myself having very little background in computational biology) and provided the tutelage necessary in order for me to accomplish much of my work as a graduate student. He was officially my advisor, but he treated me as a colleague. I am incredibly grateful to have shared a professional relationship with him; being able to pass him my ideas without apprehension, being able to work independently with his trust in me as a scholar, and ultimately, completing a successful and fulfilling thesis. It was a privilege to work under his guidance.

I would also like to thank the many members of the Hernandez Lab throughout the years. Academic labs are dynamic, with people and personalities coming and going each year, making for an always unique environment. I am happy to have participated in much of that dynamic and am grateful for the people who made the lab a welcome place for the many years I was a member. I especially want to thank Nicolas Strauli, who has been with me as both a class mate and a lab mate since the beginning of my time at UCSF. We have shared the same desk space for 6 years, but his presence never grew old. He has been a wonderful colleague through that time but was an even better friend (in spite of being an Oakland A's fan), and having him close by throughout the years enriched my daily experience in the lab and made even the most difficult days worthwhile.

I want to thank my fellow SACNistas at UCSF, especially Melissa Spear, Joselyn DelCid, and Maria Mouchess, for providing a community for people to come together and celebrate the wonderful diverse backgrounds that we all bring to the scientific endeavor. Our mission as scientists is to uncover objective truths about the natural world and to benefit humanity with that

new knowledge. But it is our rich, varied experiences as people that fosters the multitude of ideas and paths that help us accomplish that mission. SACNAS at UCSF has been a wonderful organization to be a part of, and I am thankful for the friends and community that organization has provided.

I also want to give the most gracious thanks to my family, who have always provided encouragement and support throughout my academic journey. This includes my father, Raul, my brother Ryan, my stepfather Joe, my mother, Carol, several younger siblings, and countless other members of my extended family. I am incredibly blessed to be part of such a large and supportive Latino family. But the largest thank you goes to my mother who made an especially strong impact on my life and, though her dedication as a mother, helped foster me to become the academically minded individual I am today. Her selflessness and love to her family is remarkable and I could not dream of a better parent. The work of this thesis is dedicated to her. I am also grateful to have shared both my childhood and adulthood with one of my best friends – my brother Ryan. As I've gotten older, I've realized that lifelong friendships are very rare. Being my brother, the lifelong friendship I share with Ryan is a wonderful gift.

I want to thank the countless friendships I have formed throughout the years, especially those that I made while living in the Bay Area. This includes Meaza Solomon, Tracey Chiu, Loong Kwok, and Sam Kuntz. But in particular, I want thank my partner Emily Hague. She has been my most cherished friend for 6 years and an equally cherished partner for over 3 years. It's been a wonderful journey watching our friendship grow into a loving relationship and being able to share so many experiences together – going on adventures in other countries, hiking outdoors, running half-marathons (soon marathons!), sharing meals, seeing shows, walking to Lone Palm for a Manhattan and an Old Fashioned, being parents to a cute cat. All of these wonderful experiences are made better because I can share them with her – I can't wait for a lifetime more.

ACKNOWLEDGEMENT OF PREVIOUSLY PUBLISHED MATERIALS AND RESEARCH CONTRIBUTIONS

Chapter 2 of this dissertation was previously published. The text in Chapter 2 is a reprint of the material as it appears in Torres et al. 2018. Zachary Szpiech contributed data and software/analysis tools to this work. Ryan Hernandez supervised the research that forms the basis of this chapter.

The text of Chapter 3 will be submitted for review to a peer review journal. Markus Stetter* contributed data and software/analysis tools to this work. Ryan Hernandez and Jeffrey Ross-Ibarra* supervised the research that forms the basis of this chapter.

*affiliation: University of California, Davis

The text of Chapter 4 will be submitted for review to a peer review journal. Ryan Hernandez supervised the research that forms the basis of this chapter.

REFERENCES

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

ABSTRACT

The Effects of Background Selection and Demography on Patterns of Neutral Variation within the Genome

Raul Torres

Patterns of genetic diversity across the genome are affected by multiple forces of evolution, including natural selection and population demography. These effects are manifested both locally across the genome and genome-wide. While natural selection operates directly on only a small percentage of mutations within genome, it can have wide effects across neutral regions due to genetic linkage. In the context of purifying selection at linked sites, this process is called 'background selection' (BGS) and it leads to decreases in genetic diversity and skews the distribution of allele frequencies in the genome. While much theoretical and empirical investigation has gone into how BGS and demography operate independently to affect the genome, little investigation has been conducted on how these forces pattern the genome in concert. Utilizing thousands of human genomes and population genetic simulations, I have determined that the effects of BGS in humans can be magnified by population demography. I also analyzed population genetic simulations of different demographic models with BGS and found that the effects of demography and BGS are transient through time, with dips and rises in genetic diversity dependent on how far removed they are from a demographic event. Finally, in order to gain an understanding of how BGS and recent human population growth have affected patterns of the allele frequency spectrum, I analyzed genomes of varying sample size as a function of the strength of BGS. Doing so, I found that the effect of BGS on skewing the allele frequency spectrum towards rare variants in humans is dependent on sample size and leads to larger biases in demographic inference when sample size is small.

TABLE OF CONTENTS

Appendices

LIST OF FIGURES

LIST OF TABLES

Page

Chapter 1:

Introduction

Population genetics is a rich field that seeks to understand how patterns of genetic variation in populations are influenced through time by evolution. Through both the development of mathematical models and empirical observation, research in population genetics has greatly enriched our understanding of the nature of evolutionary change at the genetic level. It has been incredibly fulfilling to take part in this pursuit, especially at a time in which the floodgates of genomic sequencing have been opened. In fact, most of the work in this thesis could not have been pursued when I graduated from college a short decade ago. It has also been humbling to pursue empirical research inspired by the work of some great theoreticians who developed much of the theory and mathematical models that my research is based upon. I never cease to be amazed by the fact that the modern synthesis – and the foundation of much of population genetics theory developed by the "great trio" of Fisher, Wright, and Haldane – occurred decades before the structure of DNA was even discovered.

In this thesis, I will focus on two main forces of evolution that pattern the genome: demography and selection at linked sites. These two forces have different consequences for how variation within the genome and between populations varies. I will begin with the observation that the average amount of genetic variation differs significantly between populations. This is clearly evident for humans where a distinct cline of decreasing genetic diversity as a function of distance from Africa has been observed [1]. This result is attributed to the successive population bottlenecks (i.e., the "serial founder effect") that occurred as modern humans migrated from their origins in Africa to locations throughout the globe. This demographic history resulted in several successive losses in genetic diversity, with non-Africans seeing especially low levels of genetic diversity compared to Africans. But humans are far from the only species suffering population bottlenecks, and differences in diversity between populations of other species have also been observed. Two prominent examples include *Drosophila melanogaster*, which have an interesting demographic history similar to humans [2], and crops such as maize, which underwent a

domestication bottleneck during the course of their demographic history [3]. In fact, the latter population motivated work that forms the basis of Chapter 3 (discussed below).

While the forces of demography, such as population bottlenecks, contribute to genome-wide changes in genetic diversity, locally across the genome other modes of evolution can also result in changes. Early on, it was recognized that genetic diversity at neutral sites can be altered by nearby selected sites along the genome [4]. This is because natural selection operates rather coarsely across the genome. Although selection itself only targets mutations with phenotypic effects, the effects of selection in the genome can still be felt nearby through genetic linkage. This leads to decreases of neutral genetic diversity in regions where selection operates. This effect, referred to as "selection at linked sites", is also strongest in regions with high linkage (i.e., low recombination). In the context of selection acting on deleterious mutations, this effect is referred to as background selection (BGS) [5]. Although selection at linked sites also operates through positive selection, I will focus on the effects of BGS throughout the entirety of this thesis.

The forces of demography and selection at linked sites have both been well-researched. However, their joint impacts on patterns of diversity in the genome are less well known. In my first year as a member of the Hernandez Lab, Ryan suggested that pursuing research on the possible joint effects of demography and selection at linked sites (specifically in the form of BGS) might yield fruitful and novel insights. Phase 1 of the 1000 Genomes Project had just been completed shortly before I came to UCSF and Ryan had shown with this dataset that BGS was a predominate driver of patterns of diversity within the human genome [6]. I took a deeper dive into the effects of BGS across a more diverse array of populations by utilizing phase 3 of the 1000 Genomes Project [7] and found that patterns of diversity under BGS are also magnified by demographic change, especially in the form of population bottlenecks. This work and its results form Chapter 2 of this thesis.

As was mentioned earlier, population bottlenecks (and dynamic demography in general) are not limited to just humans. Rather, demographic change is pervasive across natural populations. Of course, natural selection is pervasive as well, so there is strong reason to believe that our work uncovering the joint effects of selection at linked sites and demography should be translatable to other natural populations experiencing both BGS and bottlenecks. Biology is never this straightforward, though, and important unforeseen nuances can enrich scientific investigation and make it a never-ending endeavor. Thus, it was perhaps not unexpected that we encountered a previously published study in maize (corn) [3] that found directly opposite results of our work on BGS and demography in humans. These conflicting results led us into a collaboration effort with the Jeffrey-Ross Ibarra Lab at UC Davis, which authored the maize study of Ref. [3]. We embarked on an extensive simulation study looking at the effects of several different demographic models, including both population bottlenecks and population expansions, in order to observe how demography impacts patterns of diversity under selection at linked sites at a more fine level of detail. From our simulations, we observed that the time span of demographic events are important for generating specific patterns of diversity under BGS across both common and rare variants in the genome. Because of this time-dependent effect, the population bottlenecks of maize and humans can yield different results if the time passed since those bottleneck events occurred is also different. This work and its results form Chapter 3.

While selection at linked sites perturbs genetic diversity locally across the genome, it also has specific effects on the frequency spectrum of mutations (referred to as the site-frequency spectrum or SFS) across regions where it operates. Because newly arising variants take time to be eliminated by BGS, younger variants will predominate the spectrum of variation in the genome compared to older variation beyond what is expected in a neutrally evolving region. Since young neutral variants are also predominantly rare, this leads to a skew in the SFS, with proportionally more rare variants than expected under a model where BGS is absent. However, the recent

explosive growth of humans, which is another interesting aspect of our demographic history, has injected a bevy of new and rare variation into the human genome [8,9]. Larger samples sizes are also needed to detect much of this rare variation. As sample sizes increase, then, the discrepancy in the skew to the SFS between regions of strong and weak BGS may become less apparent since more rare variants that arose recently in time will begin to dominate the SFS. We tested for the impact of the recent population expansion in humans on patterning the SFS by analyzing 2,416 genomes from individuals of European ancestry. We found that across regions of strong and weak BGS, the proportion of rare variants in the SFS becomes more similar with larger sample sizes. In contrast, with smaller sample sizes, larger differences among rare variants of the SFS were observed. We also found that this impacts demographic inference procedures that utilize the SFS by introducing strong biases if the sample size used for inference is small. This work and its results form Chapter 4.

The body of work in this thesis presents results that show that demography introduces unexpected, and sometimes unintuitive, patterns of diversity in regions of selection at linked sites. Much of these results have been gleaned because we now have thousands of genomes at our disposal to form and test numerous hypotheses. As even more genomes from different populations and species come under study, more unexpected results and novel insights about the interaction of demography and selection at linked sites will surely arise. This will lead to the continuing enrichment of our knowledge on how evolution operates on the genome.

REFERENCES

1. 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. doi:10.1073/pnas.0507611102

2. David JR, Capy P. Genetic variation of Drosophila melanogaster natural populations. Trends Genet. 1988;4: 106–111.

3. 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. doi:10.1038/nplants.2016.84

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

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

6. 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. doi:10.1126/science.1198878

7. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015;526: 68–74. doi:10.1038/nature15393

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

9. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2013;493: 216–220. doi:10.1038/nature11690

Chapter 2:

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

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 A.1 in Appendix A). 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 Figure 2.1. As expected, normalized diversity was highest in African populations and lowest in East Asian populations across both 1% *B* quantile bins.

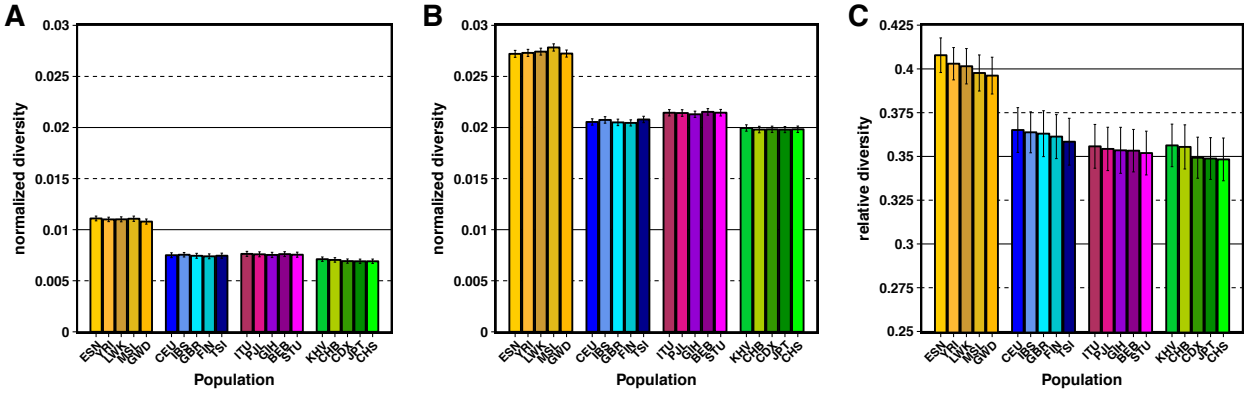


Figure 2.1. Normalized diversity and relative diversity for non-admixed populations of the Thousand Genomes Project (TGP).
(A) Normalized diversity (π/divergence) measured across the lowest 1% *B* quantile bin (strong BGS). (B) Normalized diversity measured across the highest 1% *B* quantile bin (weak 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 A.1 in Appendix A 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.

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. Figure 2.1 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 A.1 in Appendix A for a detailed description) and estimated relative diversity at neutral sites for each of the continental groups in bins 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 (Figure 2.2), 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 (Figure 2.2). These effects remained even after we controlled for the effects of GC-biased gene conversion and recombination hotspots (Figure A.1 and Figure A.2 in Appendix A) or if we did not normalize diversity by divergence (Figure A.3 and Figure A.4 in Appendix A). 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 (Figure A.5, see Appendix A).

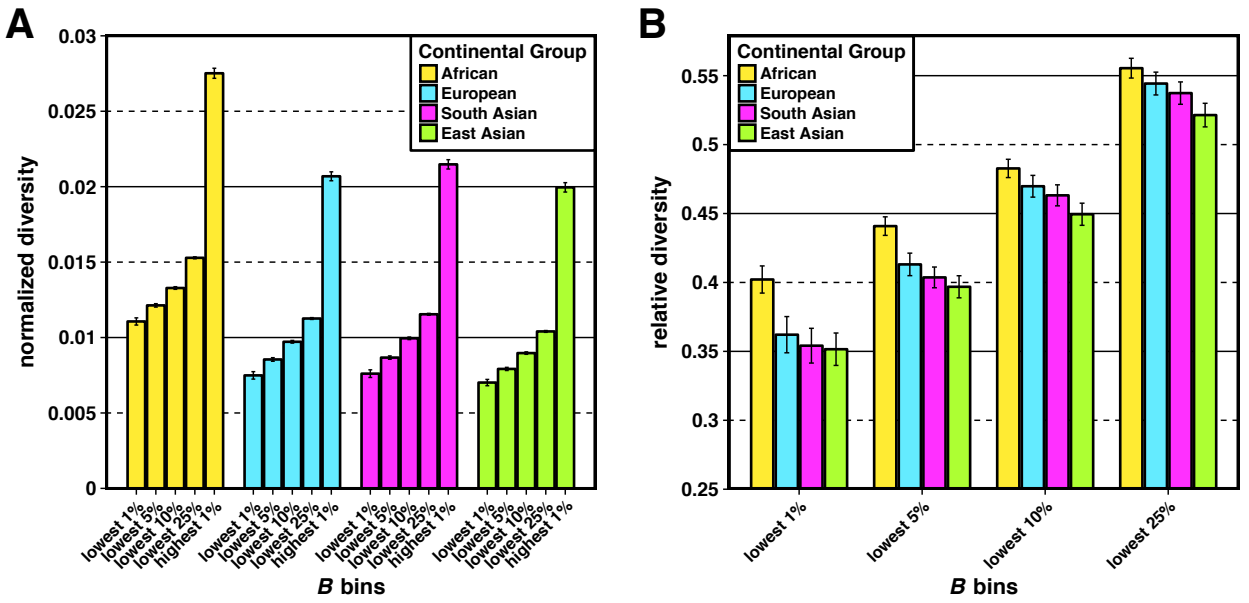


Figure 2.2. Normalized and relative diversity for Thousand Genomes Project (TGP) continental groups.
(A) Normalized diversity (π/divergence) measured across the lowest 1%, 5%, 10% and 25% *B* quantile bins (strong BGS) and the highest 1% *B* quantile bin (weak BGS). (B) Relative diversity: the ratio 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.

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 (Figure A.6 in Appendix A). 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 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 "ψ/ψ_{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 ψ/ψ_{min} showed patterns in the opposite direction to our estimates of π/π_{min}, with Africans exhibiting a lower ratio of ψ/ψ_{min} when compared to non-Africans (0.665-0.695 for Africans, 0.733-0.804 for non-Africans; Figure 2.3). 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).

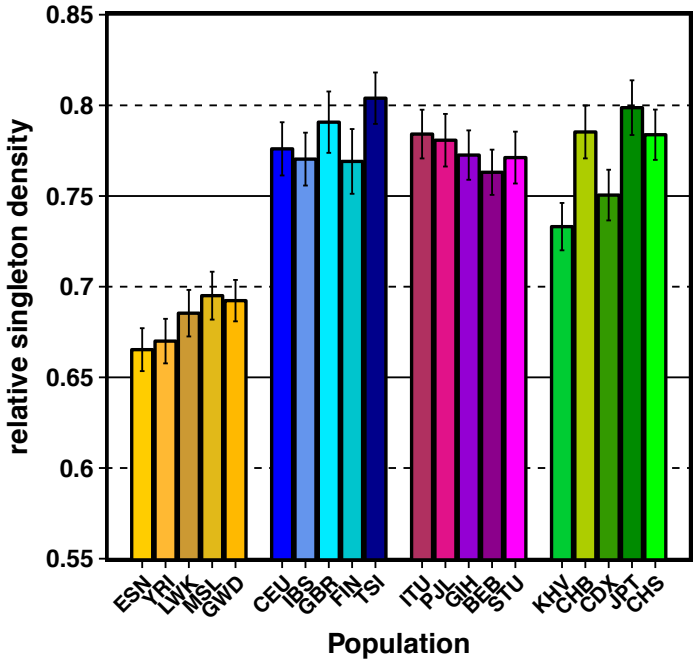


Figure 2.3. Relative singleton density for non-admixed populations of the Thousand Genomes Project (TGP).
Relative singleton density measured by taking the ratio of singleton density in the lowest 1% *B* quantile bin to singleton density in the highest 1% *B* quantile bin (ψ/ψ_{min}). Singleton density was normalized by divergence with Rhesus macaque. TGP population labels are indicated below each bar (see Table A.1 in Appendix A 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.

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; Figure 2.4 and Table 2.1). Specifically, we divided the genome into 2% quantile bins based on the genome-wide distribution of B and measured F_{ST} 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 2.1), B explained 26.9% of the change in F_{ST} across the most extreme B values. This result was robust to outliers [64] (Table A.2 in Appendix A) and dominated the effects of local recombination rate (see Appendix A).

	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

Table 2.1. Regression coefficient estimates for linear regression of F_{ST} on 2% quantile bins of B.

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.

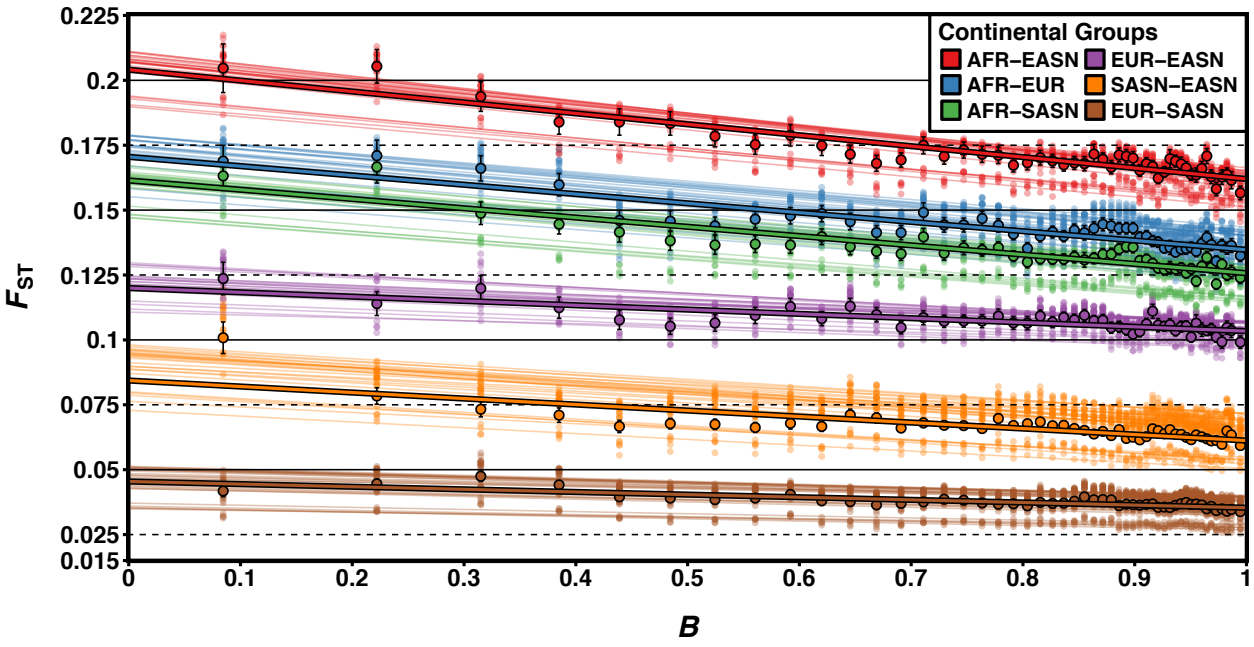


Figure 2.4. F_{ST} is correlated with B.
F_{ST} between TGP populations measured across 2% quantile bins of B. Smaller transparent points and lines show the estimates and corresponding lines of best fit (using linear regression) for F_{ST} between every pairwise population comparison within a particular 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.

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 [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 (Figure A.10, Table A.8 in Appendix A). 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 (Figure 2.1) and may therefore reflect the weaker signatures of BGS shown in Figure 2.2. 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 (Figure 2.5). 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. The bottom panel of Figure 2.5 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 (Figure A.11 in Appendix A). This finding highlights an evolutionary role that deleterious alleles can play when they are transferred across populations through migration (see Discussion).

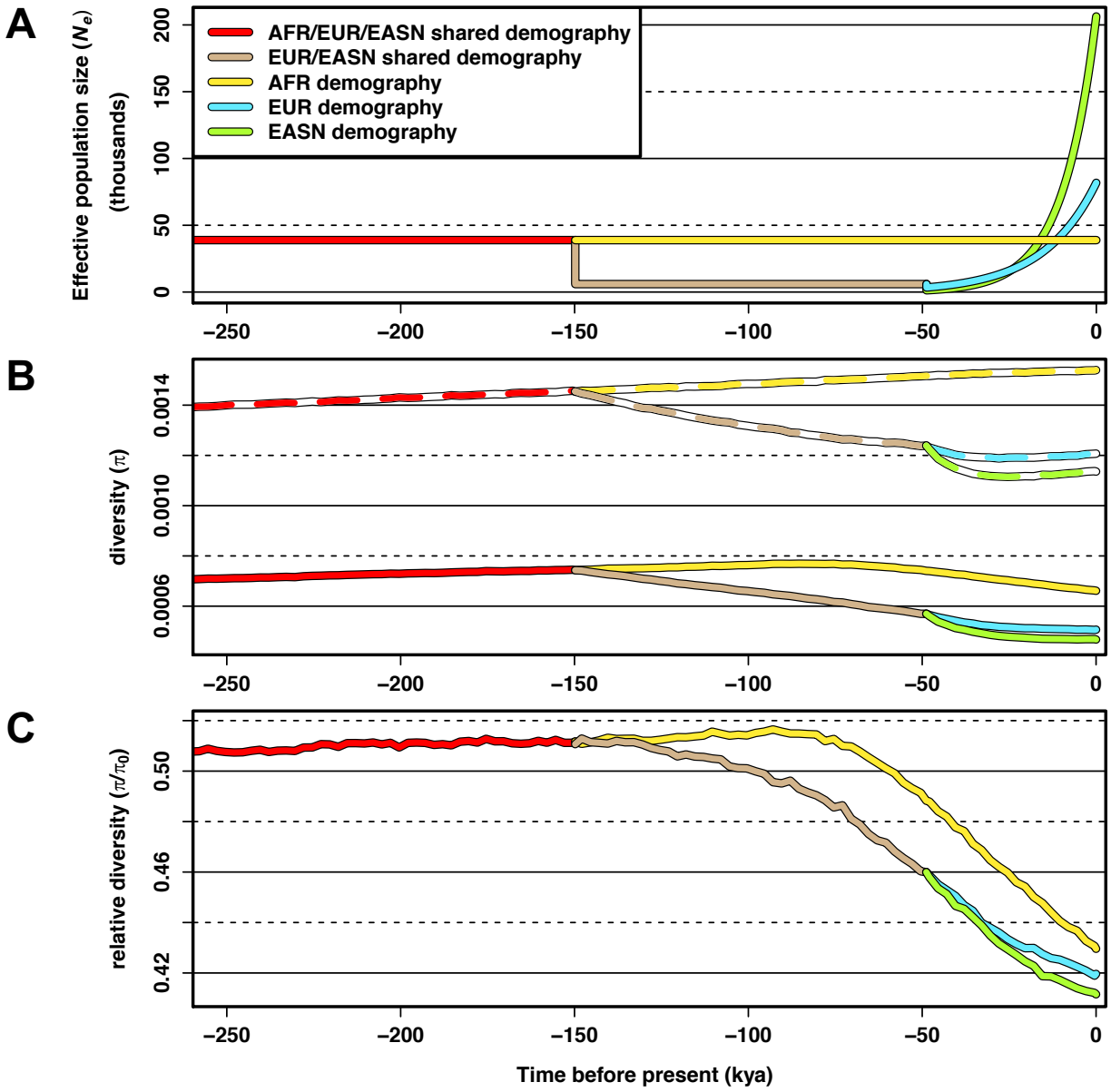


Figure 2.5. Simulations confirm that demographic events shape the effect of background selection (BGS).
(A) Inferred demographic model from Complete Genomics TGP data showing population size changes for Africans (AFR), Europeans (EUR), and East Asians (EASN) as a function of time that was used for the 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.

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} (Figure A.12 in Appendix A). 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. 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 of Figure A.12 A and Figure A.12 B in Appendix A). 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 Figure A.11 C to Figure A.12 B bottom panel in Appendix A). 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 (Figure A.13 in Appendix A). As expected, π/π_0 increased for all populations as the fraction of sites that were simulated as deleterious decreased (π/π_0 = 0.641 vs. 0.802, 0.62 vs. 0.777, and 0.611 vs. 0.777 for AFR, EUR, and EASN when 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} (Figures 2.1 and 2.2).

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). 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 Figure 2.1 C and Figure 2.5 C), 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 (Figure 2.5, Figure A.12 in Appendix A), 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 (Figure 2.5). These patterns disappear when migration is removed (Figure A.11 and Figure A.12 B in Appendix A); 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. 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 A.9 in Appendix A) 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 A.10 in Appendix A 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]) 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 theta (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 1.66×10^{-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] (α = 0.0415, β = 0.00515625) and 33.94% of deleterious sites using the gamma distribution of fitness effects inferred across coding regions by Ref. [5] (α = 0.184, β = 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 (α = 0.0415, β = 80.11) and (α = 0.184, β = 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 Appendix A. 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 A.1 in Appendix A). Additionally, only regions of the genome passing the 13-filter set were used in the calculations of π and ψ (see Table A.10 in Appendix A 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.

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. doi:10.1038/nrg3425

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

3. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002;419: 832–837. doi:10.1038/nature01027.1.

4. Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD. Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc Natl Acad Sci. 2005;102: 7882–7887. doi:10.1073/pnas.0502300102

5. Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, et al. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. 2008;4: e1000083. doi:10.1371/journal.pgen.1000083

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

7. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 2009;5: e1000695. doi:10.1371/journal.pgen.1000695

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

9. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015;526: 68–74. doi:10.1038/nature15393

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

11. Gibbs RA, Taylor JF, Van Tassell CP, Barendse W, Eversole KA, Gill CA, et al. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science. 2009;324: 528–532. doi:10.1126/science.1167936

12. Marsden CD, Ortega-Del Vecchyo D, O'Brien DP, Taylor JF, Ramirez O, Vilà C, et al. Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. Proc Natl Acad Sci. 2016;113: 152–157. doi:10.1073/pnas.1512501113

13. Caicedo AL, Williamson SH, Hernandez RD, Boyko A, Fledel-Alon A, York TL, et al. Genome-wide patterns of nucleotide polymorphism in domesticated rice. PLoS Genet. 2007;3: 1745–1756. doi:10.1371/journal.pgen.0030163

14. Begun DJ, Aquadro CF. African and North American populations of *Drosophila melanogaster* are very different at the DNA level. Nature. 1993;365: 548–550. doi:10.1038/365548a0

15. Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P. Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. Genome Res. 2005;15: 790–799. doi:10.1101/gr.3541005

16. Ometto L, Glinka S, De Lorenzo D, Stephan W. Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. Mol Biol Evol. 2005;22: 2119–2130. doi:10.1093/molbev/msi207

17. Hernandez RD, Hubisz MJ, Wheeler DA, Smith DG, Ferguson B, Rogers J, et al. Demographic histories and patterns of linkage disequilibrium in Chinese and Indian rhesus macaques. Science. 2007;316: 240–243. doi:10.1126/science.1140462

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. doi:10.1073/pnas.0507611102

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

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

21. Nei M, Maruyama T, Chakraborty R. The bottleneck effect and genetic variability in populations. Evolution. 1975;29: 1–10. doi:10.2307/2407137

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. doi:10.1073/pnas.1019276108

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. doi:10.1126/science.1219240

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

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

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

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

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

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

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

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

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

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. doi:10.1371/journal.pgen.1006130

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. doi:10.1093/molbev/msr225

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. doi:10.1038/nbt.2050

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. doi:10.1038/ng.1050

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

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. doi:10.1101/gr.3413205

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. doi:10.1371/journal.pbio.0040072

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

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. doi:10.1126/science.1198878

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. doi:10.1371/journal.pgen.1002326

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

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. doi:10.1534/genetics.112.144071

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

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. doi:10.1016/j.tig.2014.09.010

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

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. doi:10.1098/rstb.2016.0471

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

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

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. doi:10.1038/nplants.2016.84

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

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

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

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

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. doi:10.1534/g3.117.039651

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. doi:10.1038/nature06611

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. doi:10.1371/journal.pgen.1003815

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. doi:10.1016/j.gde.2016.09.006

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

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. doi:10.1017/S0016672397002954

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

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. doi:10.1101/gr.154831.113

64. Yu C, Yao W. Robust linear regression: A review and comparison. Commun Stat - Simul Comput. 2017;46: 6261–6282. doi: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. doi:10.1073/pnas.1220835110

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

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

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

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

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

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. doi:10.1371/journal.pgen.1004494

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. doi:10.1038/ng.911

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. doi:10.1093/molbev/msv106

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. doi:10.1371/journal.pgen.1005436

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. doi:10.1038/ng.3186

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. doi:10.1038/ng.2896

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

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.

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

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. doi:10.1093/oxfordjournals.molbev.a026175

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

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. doi:10.1038/ng.877

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. doi:10.1016/j.ajhg.2014.04.011

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

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. doi:10.1038/ng.651

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. doi:10.1371/journal.pbio.1001388

87. Vucetich JA, Waite TA, Nunney L. Fluctuating population size and the ratio of effective to census population size. Evolution. 1997;51: 2017–2021. doi:10.2307/2411022

88. Coop G. Does linked selection explain the narrow range of genetic diversity across species? bioRxiv. 2016; doi: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. doi:10.1002/gepi.21866

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. doi:10.1016/j.ajhg.2017.03.004

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

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. doi:10.1101/gr.202440.115

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. doi:10.1093/bioinformatics/btr330

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. doi:10.1073/pnas.1510805112

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. doi:10.1186/1471-2164-15-262

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. doi:10.1038/ncomms6692

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

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. doi:10.1101/gr.3715005

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. doi:10.1038/nature11247

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. doi:10.1101/gr.187101

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. doi:10.1093/molbev/msu211

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. doi:10.1371/journal.pgen.1003684

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. doi:10.1038/nature06258

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. doi:10.1126/science.1256442

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

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. doi:10.1093/molbev/msm149

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. doi:10.1016/j.ajhg.2015.10.006

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

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. doi:10.1371/journal.pgen.1000592

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

Chapter 3:

Complex dynamics and patterns of diversity under demography and background selection: a simulation study

INTRODUCTION

The effects of natural selection and demography on neutral genetic diversity within populations have long been of interest in evolutionary and population genetics. Recent efforts in sequencing tens of thousands of genomes across a multitude of species have yielded new and valuable insights into how these two forces of evolution have shaped extant patterns of genomic variation. Yet, while the theoretical underpinnings of the effects of natural selection and demography on genetic diversity have been investigated for decades [1–9], methodical investigation into how they jointly act to create patterns of diversity in different populations remains lacking.

It has long been observed, and described theoretically, that patterns of neutral genetic variation can vary regionally across the genome as a function of recombination rate [1,10]. This is because natural selection operating on selected sites not only decreases genetic variation at the focal site but can also lead to decreases in nearby neutral genetic diversity due to genetic linkage [11]. These effects, known as genetic hitchhiking [1] (in which neutral variants rise to high frequency with adaptive variants) and background selection [6] (BGS; in which neutral variants are removed along with deleterious variants) can be widespread across the genome. Evidence for selection at linked sites has been found across an array of species, including *Drosophila melanogaster* [10,12–16], wild and domesticated rice [17,18], *Capsella* [19], maize [20], and humans [21–27].

Demographic change can also impact patterns of diversity across the genome. For example, neutral theory predicts that the amount of genetic diversity is proportional to a population's effective population size (N_e), such that changes in N_e should result in concomitant changes to diversity [28]. However, evidence suggests that such diversity also varies much less in magnitude across species when compared to their census population sizes [29,30]. One of the most common forms of a population size change is a population bottleneck, whereby

populations suffer a large decrease followed by an expansion. Bottlenecks can occur via domestication events [31–34], seasonal or cyclical fluctuations in population size [35–38], and founder events [39–41]. Notably, while the rate of loss of diversity in response to a population contraction is quite fast, the recovery of diversity from a following population increase can be quite slow [42]. As a result, contemporarily large populations may still yield patterns of low average genetic diversity if their population size was much smaller in the recent past. In humans, this is clearly evident in European and Asian populations due to the out-of-Africa bottleneck [43].

Because selection at linked sites and demography are both pervasive forces across a multitude of species, the characterization of how these two forces interact with one another is necessary in order to develop a full picture on the determinants of neutral genetic diversity. The efficiency of natural selection scales proportionally with N_e and the impact of selection at linked sites on neutral diversity is likely to be greater in larger populations and lower in smaller populations [5,11,44], although the rate of change for lowered diversity may diminish as populations reach larger and larger sizes [45,46]. Further, demographic changes can also increase (in the case of bottlenecks) or decrease (in the case of expansions) the rate of drift. It is therefore plausible that the rate at which diversity at a neutral locus is perturbed by selection at linked sites could be highly dependent on both the current as well as long-term N_e of the population. This competition between selection at linked sites (also referred to as “genetic draft” in the literature [47], which increases with N) and genetic drift (which decreases with N) may be a key contributor to observations of limited diversity among species despite much larger observed differences in census size [44–46]. However, selection at linked sites alone may not be sufficient to explain the observed discrepancy between observed diversity and census populations sizes [48] and the action of both demography and selection at linked sites in concert may provide a better model.

Many models of selection at linked sites were also formulated with the assumption that the population (or selection itself) is large enough such that mutation-selection balance is maintained [6,49,50]. However, demography may break such assumptions and forces other than selection may drive patterns of variation in regions experiencing selection at linked sites. For example, during the course of a population bottleneck, genetic drift may transiently dominate the effects of selection at many sites, such that traditional models of selection will poorly predict patterns of genetic diversity. As was shown in Chapter 2, genetic drift is heightened in regions of selection at linked sites during a bottleneck [51], resulting in even greater losses than expected by the action of demography or selection at linked sites alone. A recent review by Comeron et al. 2017 [52] included a cursory investigation into the impact of demography on diversity in regions under BGS and suggested a dependency on demographic history. Recent empirical work in maize and humans (Chapter 2) both demonstrated a strong interaction between demography and selection at linked sites [20,51]. However, these two studies found opposite qualitative results on the impact that population bottlenecks have for patterns of diversity in regions affected by selection at linked sites, thus demonstrating that a more thorough analysis of these joint effects is warranted. Importantly, the development of tools for the accurate inference of the evolutionary forces patterning variation along the genome and the ascertainment schemes utilized to collect genomic data for such inference will require a deeper understanding of the interaction between demography and selection at linked sites.

In order to more fully explore the joint consequences of demography and selection at linked sites, we conducted extensive simulations of different demographic models jointly with the effects of BGS. We find that the time span removed from demographic events is critical for populations experiencing non-equilibrium demography and can yield contrasting patterns of diversity that reconcile the contradicting results mentioned earlier [20,51]. Additionally,

sensitivity of genetic diversity to demography is dependent on the frequency of the alleles being measured, with rare variants experiencing more dynamic changes through time.

Our results demonstrate that traditional models of selection at linked sites may be poorly suited for predicting patterns of diversity for populations experiencing recent demographic change and that the predicted forces of BGS become apparent only after populations begin to approach equilibrium. Importantly, even simple intuition about the effect of selection at linked sites may lead to erroneous conclusions if populations are assumed to be at equilibrium. These results should motivate further research into this area and support the use of models that incorporate the joint effects of both demography and selection at linked sites.

RESULTS/DISCUSSION

Patterns of diversity are dynamic under BGS after a population size change

We first present the joint effects of demography and BGS under two epoch models where there is an instantaneous decrease in size (models 2-3; Figure B.1 in Appendix B). While our simulation model incorporated a 200 kb neutral region, we first focused on patterns of diversity generated within the 10 kb window nearest to the 2 Mb locus experiencing purifying selection, as this is where BGS is strongest. Doing so allowed us to observe any change in the dynamics of π and ψ as they approached new population equilibria resulting from a change in size. For models 2 and 3, we observed an expected strong decrease in ψ and π following their population contractions in both models of BGS and neutrality, demonstrating the loss of diversity that accompanies a population reduction (Figure B.3 in Appendix B). Additionally, the values of ψ and π at the initial and final generations were observed to be lowest in models of BGS. In order to observe whether greater rates of change to diversity occurred specifically in regions of BGS (as was suggested in Ref. [51] (Chapter 2)), we normalize π and ψ generated with BGS by their equivalent statistics generated under the same demographic model in the absence of any

selection to generate two statistics: π/π_0 and ψ/ψ_0. Measuring π/π_0 and ψ/ψ_0 showed that these two statistics were dynamic through time in response to demography, with changes occurring to both their magnitude and direction (Figure 3.1). These patterns indicate that demography affects diversity differently under BGS versus neutrality. Moreover, changes to ψ/ψ_0 occurred more rapidly through time when compared to π/π_0. For example, in model 2 we observed a dip and rise in the ψ/ψ_0 statistic (i.e., relative to model 1) within the first ~0.1 N_{anc} generations (N_{anc} refers to the N_e of the population in the ancestral generation of the model). Yet, for the same model, π/π_0 remained depressed for over 0.5 N_{anc} generations (Figure 3.1). Similar patterns were observed for model 3, which experienced a greater reduction in size, although this pattern is less clear due to the greater sampling variance exhibited between successive time points for the ψ/ψ_0 statistic. This increased variance may stem from the fact that rare variants are especially sensitive to loss during a population reduction [9] and a fewer number of them will remain in the population following the size change. However, as expected for a population suffering a smaller reduction, lower sampling variance for ψ/ψ_0 was observed across each successive time point in model 2.

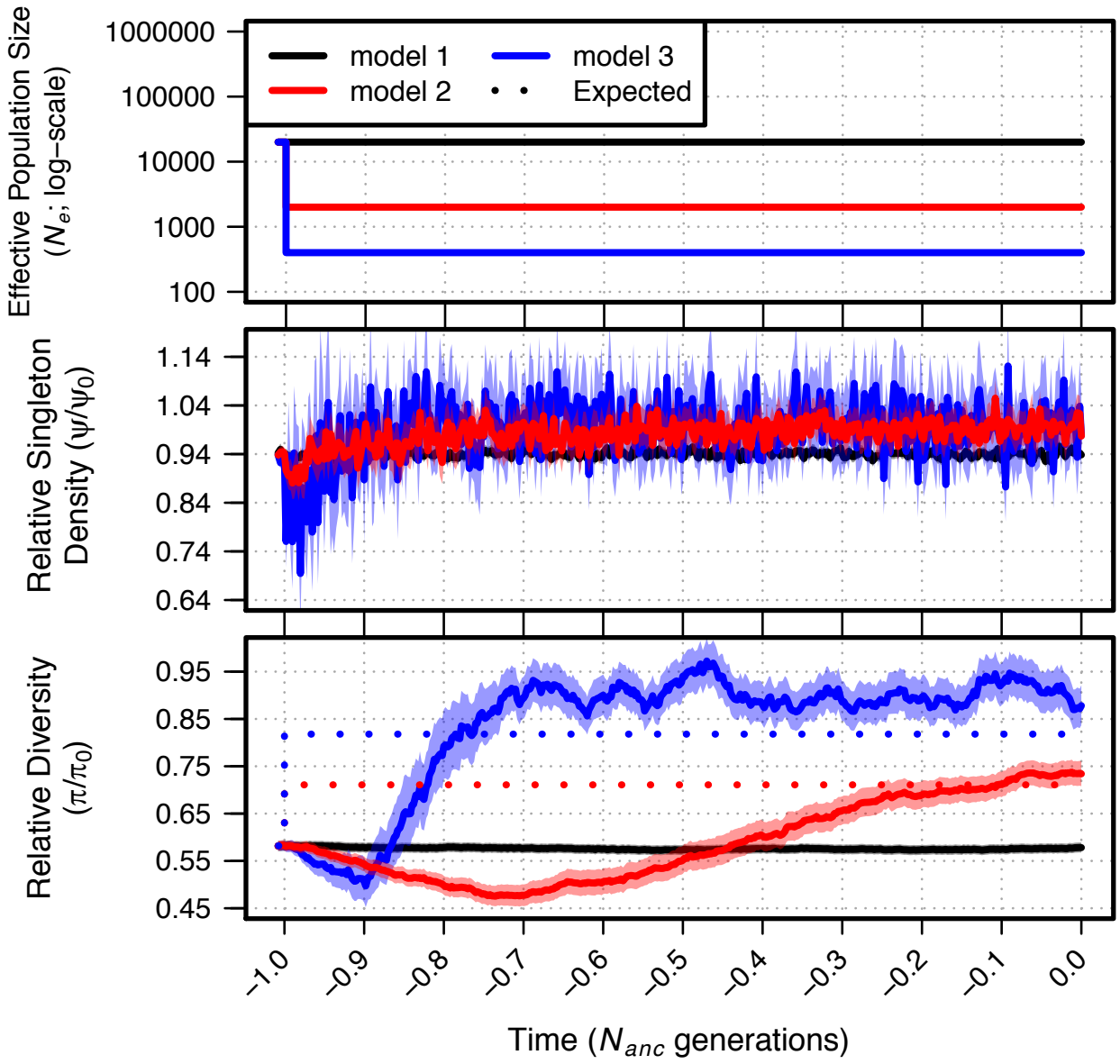


Figure 3.1 Relative singleton density (ψ/ψ_0) and relative diversity (π/π_0) across time for demographic models 1-3.
Top panel shows each demographic model (time proceeds forward from left to right; time is scaled by the N_e of the population at the initial generation (N_{anc})). Black lines show ψ/ψ_0 and π/π_0 from simulations of a constant sized population (model 1). Dotted lines in the bottom panel show the expectation of π/π_0 from Eq. (14) of Nordborg et al. 1996 for models 2 and 3 given the specific selection parameters and N_e at each time point. See Table B.1 in Appendix B for demographic model parameters. Envelopes are 95% CIs calculated from 10,000 bootstraps of the original simulation data.

Changes in population size should lead to changes in the rate of genetic drift (and possibly the efficacy of natural selection [28]) and, thus, changes in the magnitude of BGS across time. If this were the case, we would expect smaller populations to experience weaker BGS (with higher π/π_0 and ψ/ψ_0) compared to larger populations. In order to test for this type of outcome, we calculated the predicted π/π_0 given the specific N_e at each time point using the BGS model of Nordborg et al. 1996 [7] (we will refer to this as the Nordborg model; see Materials and Methods). When comparing the expectation of π/π_0 from the Nordborg model with our simulation data, we observed qualitatively opposite patterns (Figure 3.1; bottom panel). In both models 2 and 3, the Nordborg model predicted a higher value for π/π_0 immediately following the population contraction, which was in contrast to the observed transient drop in π/π_0 for both models. While the effects of BGS should be attenuated in populations with lower N_e (because the efficacy of purifying selection is weakened), the drop in π/π_0 instead demonstrated that the populations were dominated by the effects of allelic loss, which is expected for populations suffering a strong bottleneck. Additionally, more rapid effects were observed for model 3. This was despite the fact that the *expectation* of π/π_0 for model 3 was actually higher than for model 2. These observations made it clear that effects of BGS on π/π_0 immediately following a reduction in N_e were not driven by a change in the efficacy of natural selection from population decline, but rather by the increased sensitivity of allelic loss within these regions – with greater rates of loss accompanying greater reductions in population size. The diversity-reducing effects of BGS have been modeled as a reduction in N_e [53], which has been observed to increase sensitivity to drift in such regions for populations experiencing recent bottlenecks [51] (though we caution that the effects of BGS on the SFS cannot be simplified to this extent [54]). These patterns were made evident when observing the more rapid relative decrease in diversity in our selection models with BGS versus neutrality. When compared to their initial equilibrium starting points, π under BGS suffered faster rates of loss compared to the neutral

case for both models (Figure B.4 in Appendix B), with the fastest rates of loss accompanying larger reductions (i.e., model 3). Importantly, these results demonstrated that classical models that predict the impact of BGS on π/π_0, such as the Nordborg model, implicitly assume a population at equilibrium (and more explicitly, at mutation-selection balance [55]) and are inappropriate for predicting true patterns of genetic diversity for populations suffering recent size changes.

Even though ψ/ψ_0 and π/π_0 experienced drops immediately in response to the population reductions of models 2 and 3, these patterns of reduced ψ/ψ_0 and π/π_0 reversed themselves through time and, in the case of π/π_0, approached the expectation predicted by the Nordborg model. These dynamics occurred more quickly for ψ/ψ_0, which was expected since approaches to equilibrium are more rapid for rare variants relative to common variants. Additionally, the approaches to equilibrium occurred fastest for model 3, which also suffered the larger population size reduction (Figure 3.1). This faster approach was also not unexpected since, in general, the time to equilibrium is scaled by N_e [28,55] and changes resulting in smaller N_e (e.g., model 3) should also result in shorter times towards new population equilibria. Thus, despite the demographic change resulting in immediate decreases to ψ/ψ_0 and π/π_0, patterns of relative diversity in populations suffering a contraction eventually approached their expected higher equilibrium values under weakened BGS, with rates dependent on the frequency of the alleles being observed and the magnitude of the population reduction. This was evident from the fact that the final π/π_0 values for models 2 and 3 were within close approximation of the Nordborg model. However, we note that this expectation underestimated π/π_0 for model 3 because the threshold of $s < 0.15/2N_e$ was likely not conservative enough to ignore deleterious mutations that behave neutrally under the low N_e size of 400 for that model.

We next tested the effects of BGS under a demographic model with an instantaneous expansion (model 4; Figure B.1 in Appendix B). In models of both BGS and neutrality, we

observed that ψ reached higher values more rapidly than π (Figure B.3 in Appendix B), as expected [9]. However, when we observed the relative increase of ψ and π under BGS versus neutrality by measuring ψ/ψ_0 and π/π_0, we saw that the increase for ψ and π occurred at a greater rate under BGS (Figure 3.2). Thus, the patterns of ψ/ψ_0 and π/π_0 that manifested occurred in opposite directions from what we observed in demographic models with a population contraction – namely a transient increase in ψ/ψ_0 and a sustained increase in π/π_0. The latter pattern occurred despite the expectation of a decrease in π/π_0 from the Nordborg model, which would have been generated by more efficient purifying selection accompanying a larger N_e.

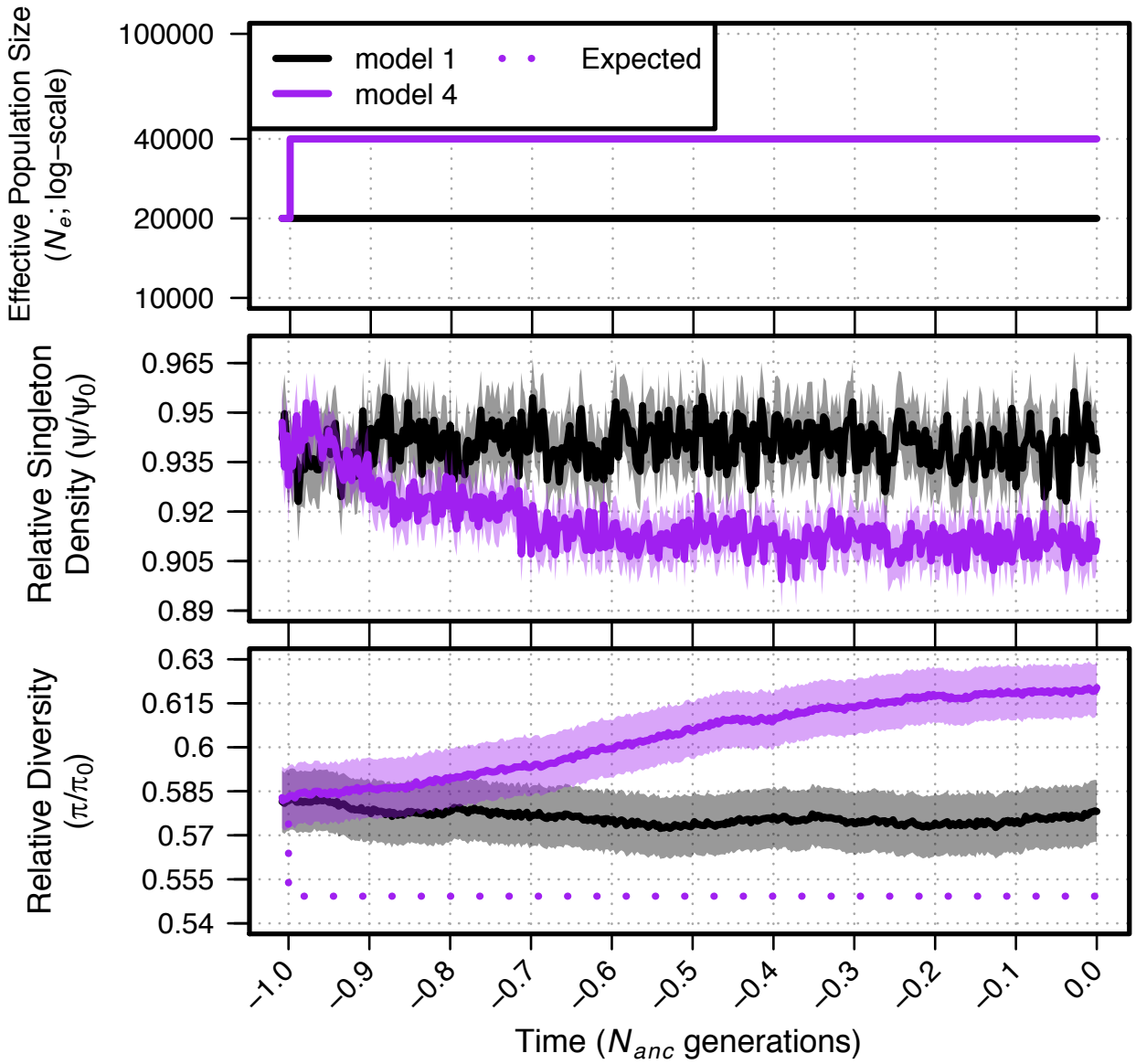


Figure 3.2. Relative singleton density (ψ/ψ_0) and relative diversity (π/π_0) across time for demographic models 1 and 4.
Top panel shows each demographic model (time proceeds forward from left to right; time is scaled by the N_e of the population at the initial generation (N_{anc})). Black lines show ψ/ψ_0 and π/π_0 from simulations of a constant sized population (model 1). Dotted lines in the bottom panel show the expectation of π/π_0 from Eq. (14) of Nordborg et al. 1996 for model 4 given the specific selection parameters and N_e at each time point. See Table B.1 in Appendix B for demographic model parameters. Envelopes are 95% CIs calculated from 10,000 bootstraps of the original simulation data.

The more sensitive and rapid response to population increase under BGS recapitulated the faster approaches to equilibrium that were exhibited in the contraction models. Intuitively, these faster approaches to new equilibrium levels under BGS in response to size changes make sense if we consider the fact that the distance between equilibrium diversity levels are closer to one another under BGS when compared to neutrality. This was evident when we compared ψ at the initial and final generations for models 2-4, which appeared close to their new equilibria at the end of their respective demographic histories (Figure B.3 and Table B.2 in Appendix B). This was also evident when observing the same differences for π for models 2 and 3 (Figure B.3 and Table B.3 in Appendix B). There, it was even more clear that the distance between π for the initial and ending generations was lower under BGS relative to neutrality. This provided a potential explanation for why we observed the specific dynamics of π/π_0 immediately following a size change. This argument likely does not hold for model 4 because, due to its higher N_e, it is unlikely that equilibrium has yet been reached. However, the observation of an increase in π/π_0 for model 4 did provide supporting evidence that a faster approach to equilibrium under BGS still existed under an expansion model. This was evident when observing the relative change in π under BGS vs. π under neutrality (Figure B.4 in Appendix B). There, we observed that π increased at a faster rate under BGS, thus explaining why π/π_0 increased following the population expansion. Presumably, this also indicates that π under BGS will reach a new equilibrium first, at which point π/π_0 will begin a downward trajectory in response to π continuing to increase under neutrality but π under BGS remaining at a constant equilibrium. This is a likely outcome if the qualitative changes in ψ/ψ_0 foreshadow the future dynamics of π/π_0. The end result would also include a decrease in π/π_0 relative to its initial starting point, with π/π_0 eventually reaching a value close to its expectation (Figure 3.2; dotted lines). Although we foresee no reason why this prediction should not hold true, more extensive simulations will be necessary to confirm this.

Population contractions dominate patterns of pairwise diversity (π) under BGS during a population bottleneck-expansion

We built upon the simple two epoch demographic scenarios to test more complex demographic scenarios and their effects on patterns of diversity under BGS. Specifically, we incorporated a bottleneck-expansion model where we simulated a population undergoing a contraction similar in size to models 2 and 3, but with a subsequent expansion to 400,000 individuals by the final generation. Although we vary the time length of the contraction and expansion events (see Table B.1 in Appendix B), qualitatively, these bottleneck-expansion models match the demography of previous empirical studies investigating the impact of BGS within dynamic populations [20,51]. They also helped to glean information about which particular demographic events – contractions or expansions – dominate the overall patterns that we witnessed for the two epoch models described previously.

For the demographic models in which the bottleneck event began -1.0 N_{anc} generations in the past and was followed by immediate expansion (models 5-6), we observed transient decreases in ψ/ψ_0 and π/π_0, recapitulating the dynamics observed for models 2 and 3 (Figure 3.3). Similar to models 2 and 3, we also observed approaches to higher values of π/π_0 later in their demographic histories, consistent with the effects of weakened BGS as a result of the initial population decline. This was in contrast to the lower π/π_0 expected for an expanding, larger population (Figure 3.3; dotted lines). The approach to higher π/π_0 values later in the demographic model also occurred more rapidly for model 6 than model 5. Thus, as was shown when comparing model 3 to model 2, the time to equilibrium after a size change is highly dependent on N_e and occurs more quickly for populations suffering larger reductions. The fact that these patterns of increasing π/π_0 exist despite the population expansion events of models 5 and 6 also demonstrates the dominant effects that a population decline has on patterns of diversity under selection at linked sites.

While the population expansion eventually did have an effect on increasing π under both BGS and neutrality after the initial reduction in size, this increase occurred at a higher relative rate under BGS and was further accelerated by larger reductions in population size. This is evident in Figure B.5 (in Appendix B) where we compared π relative to its initial value through time. There, we observed a sharper increase in π under BGS following its minimum point for model 6 when compared to model 5. However, for both models, the faster rate of recovery of π under BGS in response to the expansion also ensured that π/π_0 continued to remain higher in later generations. So while bottlenecks led to a sharper rate of decrease in π under BGS when compared to neutrality, they also aided in a faster rate of recovery during the expansion, thereby leading to an increase in π/π_0 in the face of growth and mimicking patterns evident for models with no expansion. The fact that the approach to higher π/π_0 occurred *despite* the increasing population size of both demographic models clearly demonstrated that patterns of π/π_0 were primarily being driven in response to earlier demographic events (i.e., the initial population contraction). Yet the stronger action of BGS in response to population expansion should eventually arise for both models and is suggested by the decreasing π/π_0 predicted by the Nordborg model (Figure 3.3; dotted lines). Such patterns may take much longer to manifest than the time span we have simulated here if approaches to equilibrium take on the order of $4N_e$ generations [28,55].

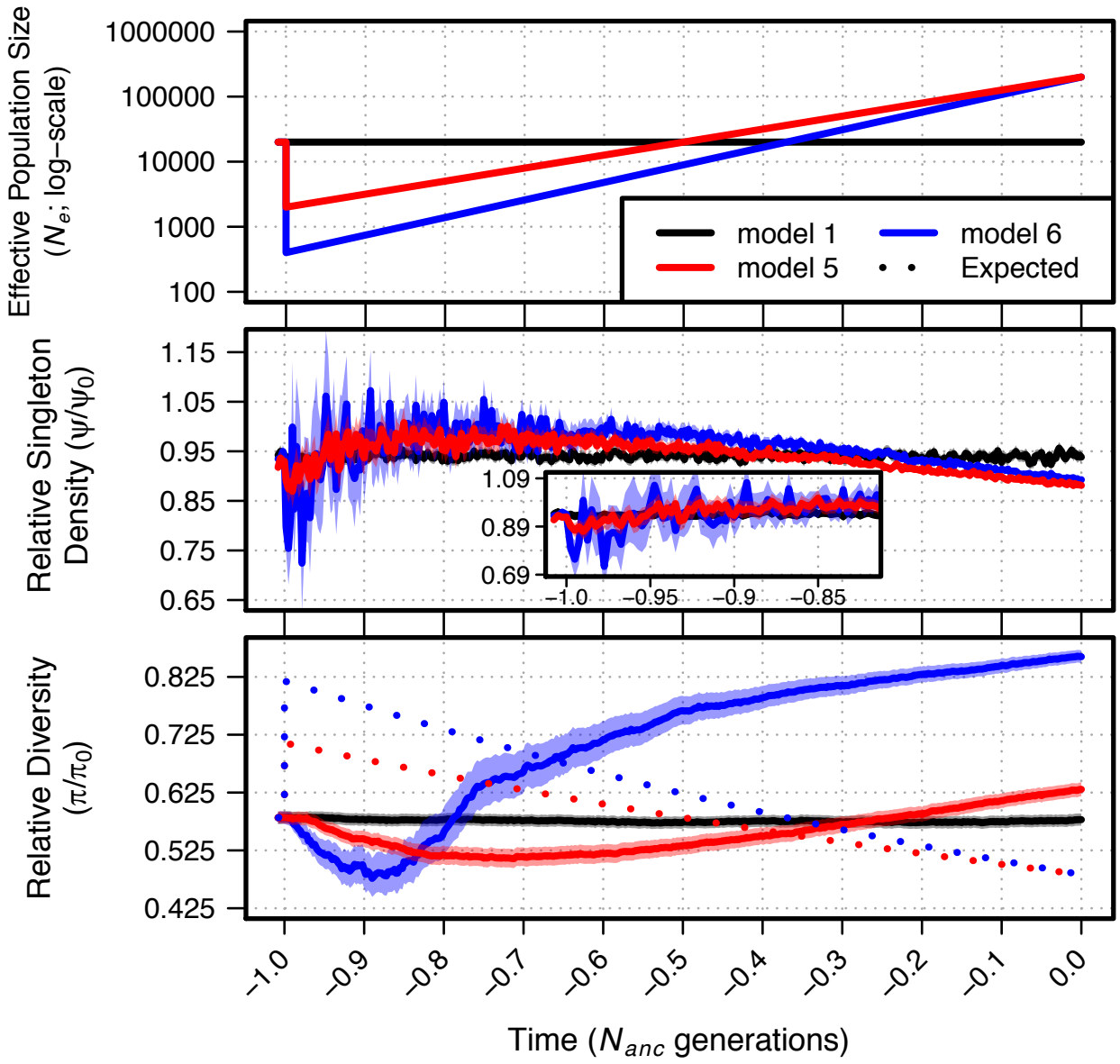


Figure 3.3. Relative singleton density (ψ/ψ_0) and relative diversity (π/π_0) across time for demographic models 1 and 5-6.
Top panel shows each demographic model (time proceeds forward from left to right; time is scaled by the N_e of the population at the initial generation (N_{anc})). Black lines show ψ/ψ_0 and π/π_0 from simulations of a constant sized population (model 1). Dotted lines in the bottom panel show the expectation of π/π_0 from Eq. (14) of Nordborg et al. 1996 for models 5 and 6 given the specific selection parameters and N_e at each time point. Inset shows a smaller range along the x and y axes for greater detail. See Table B.1 in Appendix B for demographic model parameters. Envelopes are 95% CIs calculated from 10,000 bootstraps of the original simulation data.

Observations of the dominant impact of genetic drift and weakened BGS following a population reduction were also apparent when measuring patterns of π/π_0 in models that had a more sustained contraction (i.e., models 7 and 8). There, a decline in N_e was sustained for an additional 0.5 N_{anc} generations before the expansion event began (Figure B.2 in Appendix B). For these models, the rise of π/π_0 resulting from weakened BGS occurred more quickly than for their counterpart models with an immediate expansion (Figure B.6 in Appendix B). For example, the inflection point at which π under BGS surpassed π under neutrality occurred at -0.305 and -0.848 N_{anc} generations for models 7 and 8, respectively (Figure B.5 in Appendix B). For models 5 and 6, these inflection points occurred later in time at -0.235 and -0.785 N_{anc} generations, respectively. Further, the final π/π_0 values for models 7 and 8 were 0.643 and 0.887 but for models 5 and 6, they were only 0.631 and 0.860. Thus, the sustained lower N_e of models 7 and 8 aided in accelerating the approach to the new equilibrium established by population reduction. This provided further evidence for the dominant role that population bottlenecks have for patterns of diversity under BGS, even when populations expand past their ancestral size.

It is possible that the expansion itself has contributed to the rise in π/π_0 for models 5-8, as was seen for model 4. However, this is unlikely to be the case because the rise in π/π_0 occurred fastest for models 7 and 8, which had a delayed onset of expansion. Second, this rise occurred faster still for models 2 and 3, which had no expansion in size. There, the inflection point at which π under BGS surpassed π under neutrality occurred at -0.44 and -0.865 N_{anc} generations for models 2 and 3, respectively (Figure B.4 in Appendix B). Rather, the population expansion of models 5-8 appeared to retard the approach to equilibrium in response to their size reductions, thus preventing π/π_0 from attaining higher values. When comparing each respective model's maximum π/π_0, we observed that models 2 and 3 both had the highest values given their respective population reductions to 2,000 and 400 individuals (π/π_0 = 0.738 and 0.972, respectively; Figure 3.1). This was followed by models 7 and 8 (π/π_0 = 0.644 and

0.887, respectively) and finally, by models 5 and 6 (π/π_0 = 0.631 and 0.861, respectively; Figure 3.3). Thus, those models experiencing the shortest amount of time at reduced population sizes saw the lowest rises in π/π_0. For models 5-8, though, π/π_0 was still approaching higher values since these models were not at equilibrium after 1.0 N_{anc} generations. It is likely that π/π_0 for these models would attain even higher values if their demographic histories were extended.

Rare variants are more dynamic through time as a function of demography and BGS

In contrast to π, the loss and gain of ψ and change to ψ/ψ_0 in response to the bottleneck-expansion models was much more rapid and dynamic through time. This was expected since rare variants (e.g., singletons) are more likely to be lost during a contraction and during an expansion, injection of new mutations fill these bins in the SFS first. For models 5 and 6, we witnessed a very brief dip in ψ/ψ_0, resulting from a greater relative decrease in ψ under BGS when compared to neutrality (Figure 3.3; Figure B.5 in Appendix B). Following this dip, ψ under BGS increased at a relatively faster rate than ψ under neutrality, resulting in a higher ψ/ψ_0 relative to their initial values (Figure B.5 in Appendix B). Similar patterns were also seen for models 7 and 8 (Figure B.5 and Figure B.6 in Appendix B). Qualitatively, these first directional changes in ψ/ψ_0 matched those of π/π_0, but occurred over a much shorter time span. These changes were likely a consequence of regime change from the dominance of genetic drift immediately following the population reduction to the dominance of weakened BGS from a reduced N_e. This was previously exhibited by models 2 and 3 and additional evidence for this was provided by the observation that changes in the magnitude of ψ/ψ_0 were greater for model 6 and model 8 than for model 5 and model 7 (i.e., greater for models suffering larger reductions in N_e).

Because the dynamics of rare variants are more sensitive to demography, we also witnessed another change in the direction of ψ/ψ_0 that was not observed for π/π_0. For models 5

and 6, following an increase in ψ/ψ_0 above its initial point from weakened BGS, we saw a decrease later in time, with ψ/ψ_0 falling below that initial point by -0.2 N_{anc} generations (Figure 3.3; Figure B.6 in Appendix B). This last decrease in ψ/ψ_0 could not have resulted from an increased sensitivity to drift because the population sizes were larger during this phase of the demography (79,636 N_e and 57,722 N_e at -0.2 N_{anc} generations for models 5 and 6, respectively). Rather, BGS appeared to act more strongly in these later generations and, thus, limited ψ relative to its value under neutrality. Supporting this, we observed a slower rate of increase in ψ under BGS towards the very end of the expansion for models 5-8 (Figure B.5 and Figure B.7 in Appendix B). Finally, we observed that the final ψ/ψ_0 value for model 5 was lower than for model 6 (0.881 vs. 0.893) and the final ψ/ψ_0 value for model 7 was lower than for model 8 (0.879 vs. 0.896) (Figure B.6 in Appendix B). This may have resulted from the fact that models 5 and 7 had higher long term N_e and experienced a concomitantly stronger amount of BGS throughout their history due to their shallower population bottlenecks.

Recent bottlenecks result in opposite patterns of relative diversity compared to longer bottlenecks

We also ran a set of simulations with demographic histories simulating the effects of more recent bottlenecks on patterns of π/π_0 and ψ/ψ_0 (models 9-12; Figure B.2 in Appendix B). These models were similar to models 5-8, with identical starting and ending population sizes and population size reductions. However, the duration of their demographic history lasted only 0.1 N_{anc} generations. For these models, we observed similar patterns in response to the population reductions seen in the previous models. In all cases, π/π_0 suffered a decrease, which was once again in contrast to the expectation given by the Nordborg model (Figure B.6 in Appendix B). Also similar to the previous models, ψ/ψ_0 for models 9-12 suffered a transient decrease followed by a recovery over its initial value. For models 9 and 10, which both had an

immediate expansion following their size reductions, the magnitude of loss for π and ψ was less than for their counterpart models – models 5 and 6 (compare Figure B.7 to Figure B.8 in Appendix B). This result likely stemmed from the higher rate of population growth necessary to end at a size of 200,000 individuals over the course of 0.1 N_{anc} generations for models 9 and 10, which mitigated the greater loss of π and ψ exhibited by models 5 and 6. Additionally, the decrease in π/π_0 was less for models 9 and 10 when compared to models 5 and 6. After 0.1 N_{anc} generations, π/π_0 was 0.545 and 0.492 for models 5 and 6, respectively but 0.573 and 0.541 for models 9 and 10, respectively (Figure B.6 in Appendix B). This demonstrated the effects of the greater rate of expansion on limiting the sensitivity to drift in regions of BGS. Further, for models 11 and 12, which had a delayed expansion, measures of π/π_0 were also lower than for models 9 and 10 after 0.1 N_{anc} generations, exhibiting values of 0.542 and 0.527, respectively (Figure B.6 in Appendix B). These models also clearly demonstrated the feature of ψ under BGS not only declining more quickly in magnitude immediately following the population contraction but also recovering more quickly once it reached its minimum, thus displaying the more rapid behavior characteristic of patterns of diversity under the effects of BGS and demography. Specifically, when comparing ψ under BGS to ψ under neutrality, ψ under BGS in the final generation was relatively higher than its initial value (Figure B.9 in Appendix B). This caused the elevated ψ/ψ_0 exhibited in the final generation of models 9-12 (Figure B.6 in Appendix B).

Because the history of models 9-12 only lasted 0.1 N_{anc} generations, we also observed much more limited dynamics of π/π_0 and ψ/ψ_0. Specifically, π/π_0 did not recover above its initial starting point by the final generation and ψ/ψ_0 did not decrease in response to the population expansion, but rather continued to remain elevated (Figure B.6 in Appendix B). These features are important because they demonstrated that qualitatively similar demographic events, such as the bottleneck-expansion model shared by models 5-8 and 9-12, can yield opposite trends in statistics used as proxies for measuring the intensity of BGS. Thus, the resulting effects on

patterns of relative diversity under BGS depend on how far removed the point of observation is from a particular demographic event. Such patterns also help to reconcile the qualitatively different observations yielded by previous studies [20,51] (discussed in Conclusions).

Diversity patterns under BGS across an extended genetic region

We also measured patterns of π/π_0 across time for the entire 200 kb neutral region. Doing so showed the characteristic "trough" structure of increasing relative diversity as a function of genetic distance from the focal locus under selection (Figure 3.4; Figures B.10-B.20 in Appendix B). For the two-epoch models with a population contraction, we observed that the slope of the trough became more shallow through time, with the difference between the closest 10 kb bin and farthest 10 kb bin from the 2 Mb selected locus decreasing between the initial and final generations (Figures B.10-B.11 and Table B.4 in Appendix B). Thus, the impact of population contractions on mitigating the effects of BGS resulted in larger shifts of π/π_0 in regions of the genome already under the strongest amount of BGS. This makes sense since regions farther removed from loci under purifying selection (i.e., under weaker effects of BGS) have values of π/π_0 closer to the neutral expectation of 1. Thus, the upper bound for change in π/π_0 will be more limited there compared to regions more proximal to a selected locus. However, the decrease in the slope of the trough was initially minimal and only accelerated *after* π/π_0 across the 200 kb region reached its minimum values (Table B.4 in Appendix B). This provided further evidence for the dominant effects of drift and allelic loss on driving the initial decrease in π/π_0 immediately following a population contraction, which should have unbiased effects across all bins within the trough. During the subsequent recovery to higher values of π/π_0, we saw smaller differences arise between the nearest and farthest 10 kb bins, demonstrating the expected weakening effects of BGS following a population decline. This weakening of the

trough structure was also most apparent for model 3, with patterns of π/π_0 appearing essentially flat across the 200 kb region in the final generation (Figure B.11 in Appendix B).

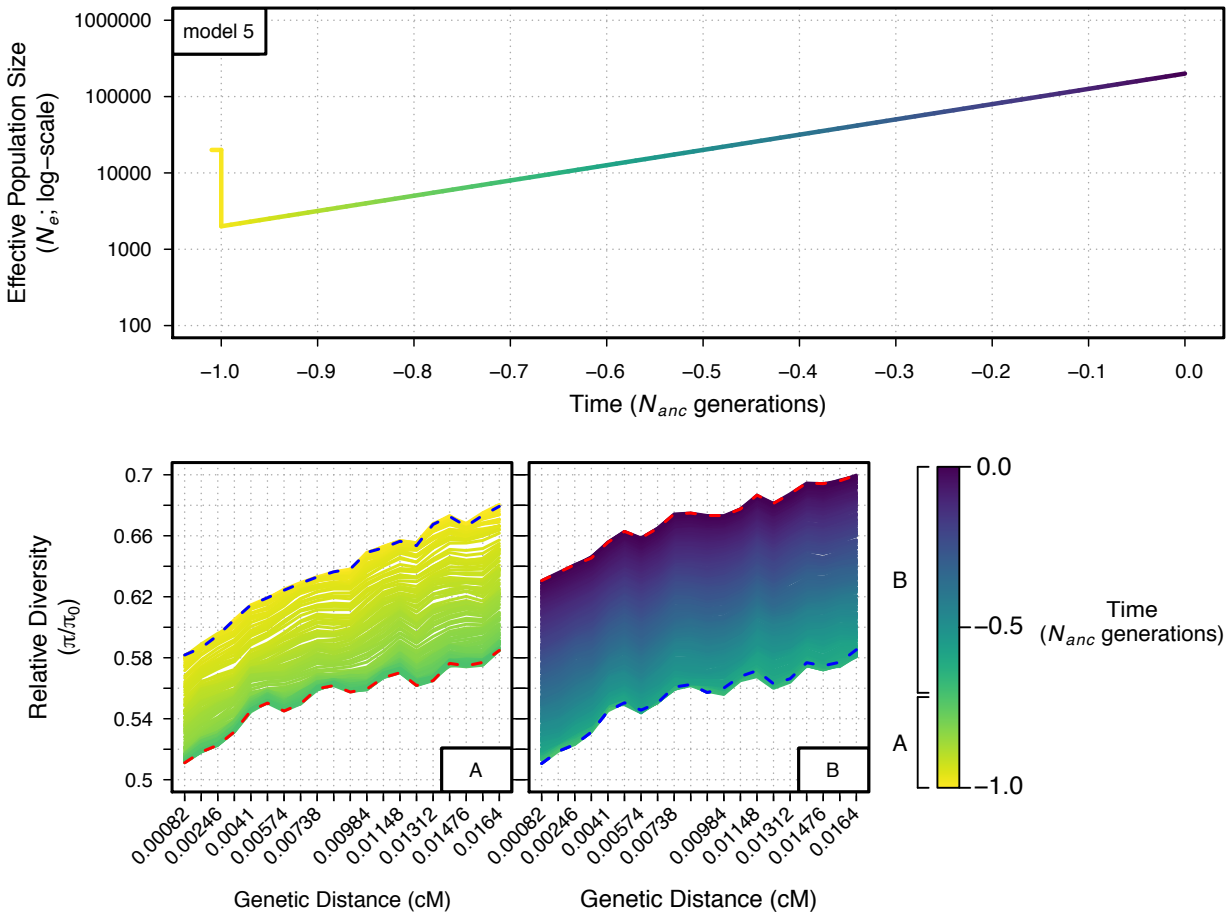


Figure 3.4. Relative diversity (π/π_0) through time for demographic model 5 measured across a neutral 200 kb region under the effects of BGS.
The genetic distance of each 10 kb bin from the selected locus is indicated on the x-axes, with genetic distance increasing from left to right. Each line measuring π/π_0 across the 200 kb neutral region represents a specific generation of the demographic model (401 discrete generations total), which is indicated by the color of the demographic model at the top of the figure (time proceeds forward from left to right; time is scaled by the N_e of the population at the initial generation (N_{anc})) and in the figure legend. Multiple plots are given for π/π_0 to prevent overlap of the measurements between generations (see legend for specific generations covered in each plot). Blue dashed lines and red dashed lines indicate the first generation and last generation measured, respectively, for each specific plot.

We observed similar patterns for the bottleneck-expansion models that lasted 1 N_{anc} generations (Figures B.13-B.16 in Appendix B). Notably, among those models, the slope of the trough became more shallow for models 6 and 8 which were also the models suffering the deepest reductions in size (Figures B.14, B16 and Table B.4 in Appendix B). The fact that the trough structure for models 5 and 7 was better maintained showed that the population expansion following the reduction in size kept BGS stronger through time relative to models with the same decline in size but without a recovery (e.g., model 2). For models lasting 0.1 N_{anc} only captured the decrease of π/π_0 due to drift and saw very little difference develop across their troughs (Figures B.17-B.20 in Appendix B). Similarly, for model 4, the trough structure remained unchanged throughout its demographic history (Figure B.12 in Appendix B).

Finally, repeating the same analysis across the 200 kb regions for ψ/ψ_0 yielded no discernable patterns. Troughs were slightly apparent for the final generations of some models (i.e., models 5 and 7), but the stochasticity between windows for ψ/ψ_0 swamped any obvious patterns across the 200 kb region through time. Since ψ/ψ_0 is already less affected (and, thus, closer to 1) than π/π_0 because BGS perturbs common frequency bins of the SFS more than rare ones [54], any signal using rare frequency bins will be inherently more difficult to capture across differing magnitudes of BGS. However, more extensive simulations may help to uncover such patterns.

Conclusions

Recently, two empirical investigations into the joint impacts of demography and selection at linked sites in the context of BGS yielded interesting and intuitive, albeit contradictory, observations. Beissinger et al. 2016 [20] conducted a study across 36 samples from teosinte and its domesticated counterpart, maize. They found that patterns of relative genetic diversity (i.e., π/π_0) across regions experiencing linked selection in maize were higher than in teosinte.

They attributed this to the historically larger N_e of teosinte, which led to more efficient natural selection and, thus, a greater removal of neutral genetic diversity than has occurred in maize (maize suffered a bottleneck during the course of its domestication). However, the contemporary population size of maize, a staple food crop grown world-wide, is now much larger than teosinte and should be experiencing stronger selection in its recent history. Supporting this hypothesis, relative singleton density of neutral sites in maize, which should reveal more recent signals of evolutionary history, was lower compared to teosinte. However, Torres et al. 2018 [51] (Chapter 2) revealed opposite patterns in humans. There, through a comprehensive analysis of over 2,500 human genomes, they observed that relative genetic diversity (π/π_0) was lower in non-Africans, a population that has undergone a series of extensive population bottlenecks and exhibits a low long-term N_e, when compared to Africans. Additionally, relative singleton density was also higher in non-Africans. In conclusion, the authors attributed these patterns to a higher sensitivity to demography and drift in regions of selection at linked sites, thus yielding lower relative diversity in bottlenecked populations. They also concluded that the greater long-term N_e, and thus more effective purifying selection and greater BGS, of Africans has led to their observed lower relative singleton density.

While these patterns observed in maize and humans are seemingly in disagreement, important demographic details, such as the length of the population bottleneck and the time since the post-bottleneck population expansion began, may also be significant contributors to these results. As our simulations demonstrated, it is possible for two qualitatively similar demographic histories to yield opposite patterns if the window of time in which those patterns are observed are different. In the case of maize compared to teosinte, observations are being made at a time in which BGS is operating less effectively on removing average pairwise diversity (due to its lower long-term N_e) but more effectively on singletons (due to its higher contemporary N_e). But if this population had been sampled more closely to its population

bottleneck event, observations of relative diversity may have been more aligned with what is currently observed for humans. The approximate number of generations removed from the domestication bottleneck event for maize is about 15,000 generations [20]. For humans, the approximate number of generations removed from the out-of-Africa bottleneck event is only 6,000 generations [51]. Therefore, it is not unreasonable to suspect that these different timespans contribute to the qualitatively different observations now being observed. Importantly, these two studies provide striking examples of the importance of considering the impact of demography and time on extant patterns of diversity to avoid mis-attributing the underlying forces driving those patterns in regions experiencing selection at linked sites. Since the null expectation of a natural population should be that it is *not* at demographic equilibrium [56], alternative hypothesis testing on selection at linked sites should also include the effects of non-equilibrium demography and how they affect patterns though time. However, in the specific context of maize and humans, we also note that other details, such as the periodic bottlenecks suffered by non-Africans (which may have further accelerated drift) and the differences among the distribution of fitness effects for these two species, are equally important to consider and warrant further investigation as well.

Recent model development incorporating demography into models of BGS holds promise on generating demographically aware models on the effects of selection at linked sites in populations. In particular, results from Zeng 2013 [49], which formulated a simulation-based structured coalescent model of BGS with demography, also showed that demography can perturb levels of genetic diversity under BGS through time. In a separate study, an analytical model which is capable of incorporating changing demography was formulated and will prove more ideal for performing inference of selection at linked sites in dynamic populations [50]. Both of these models, though, are limited in their ability to accurately predict the effects of selection at linked sites when mutation-selection balance breaks down, which typically occurs when the

population scaled selection coefficient, γ, approaches 1. In general, the deterministic approximation implicit for models of BGS may not be suitable for $\gamma \leq 3$ [53]. During the course of a bottleneck, as we have simulated here, γ is likely to fall below these thresholds and patterns of diversity may be more strongly affected by other processes such as genetic drift or the "interference selection" regime described in Good et al. 2014 [57]. For the case in which *s* is drawn from a skewed distribution, such as the gamma distribution, the deterministic approximation is further likely to break down when *s* is small. We attempt to limit these specific issues in the Nordborg model by simply truncating *s* so that predictions better match observed levels of BGS for various population sizes (albeit, under the additional assumption of demographic equilibrium). This simplistic approach may be suitable for other models of BGS, but as our results showed, it will likely provide only a coarse estimate for the prediction of diversity under BGS (Figure B.21 in Appendix B).

Finally, our results extend the recent debate on patterns of diversity in selected sites in non-equilibrium populations (especially in humans [58–61]) to patterns of diversity across neutral sites. For the specific case of selected sites, sites under strong selection are more sensitive to demographic change and will reach new equilibrium frequency levels more quickly than neutral sites or weakly selected sites [62–64]. As we have shown here in the context of neutral sites, because the underlying equilibrium frequency of neutral sites depends on the strength of selection at linked sites, demographic change will also result in distinct responses to their change in frequency. In addition, the rate of change will also depend on which bins of the SFS diversity is being measured with. Together, this results in the complex change of π/π_0 and ψ/ψ_0 through time that we observed from our simulations. This insight should provide caution, however, for studies attempting to uncover the action of natural selection by comparing sites within the genome since, even when controlling for the strength of BGS itself, frequencies of

neutral sites may still be at different relative levels depending on the recent demographic history of the population.

MATERIALS AND METHODS

Simulation model

We simulated a diploid and randomly mating population using fwdpy11 v1.2a (https://github.com/molpopgen/fwdpy11), a Python package using the fwdpp library [65]. Selection parameters for simulating BGS followed those of Torres et al. 2018 [51] (Chapter 2), with deleterious variation occurring at 20% of sites across a 2 Mb locus and the selection coefficient, *s*, drawn from two distributions of fitness effects (DFE). Specifically, thirteen percent of deleterious sites were drawn from a gamma distribution (parameters: mean = α/β, variance = α/β^2) parameterized Gamma(α = 0.0415, β = 80.11) and seven percent from a distribution parameterized Gamma(α = 0.184, β = 6.25). These distributions mimic the DFEs inferred across non-coding and coding sites within the human genome [66,67]. Fitness followed a purely additive model in which the fitness effect of an allele was 0, 0.5*s*, and *s* for homozygous ancestral, heterozygous, and homozygous derived genotypes, respectively. Per base pair mutation and recombination rates also followed those of Torres et al. 2018 [51] (Chapter 2) and were 1.66×10^{-8} and 8.2×10^{-10}, respectively. We also included a 200 kb neutral locus directly flanking the 2 Mb deleterious locus in order to observe the effects of BGS on neutral diversity. For all simulations, we simulated a burn-in period for 10N generations with an initial population size of 20,000 individuals before simulating under 12 specific demographic models. The demographic models included one demographic model of a constant sized population (model 1) and eleven non-equilibrium demographic models (models 2-12; Figures B.1, B.2 and Table B.1 in Appendix B). The non-equilibrium demographic models modeled an instantaneous population contraction only, an instantaneous population expansion only, or an instantaneous population

contraction followed by a period of exponential growth (i.e., a bottleneck-expansion model). For the bottleneck-expansion models (models 5-12), we varied the time lengths of the contraction and expansion events and the size of the contraction event (Figure B.2 and Table B.1 in Appendix B). For each demographic model, we also conducted an identical set of neutral simulations without BGS by simulating only the 200 kb neutral locus. Each model scenario was simulated 5,000 times.

Diversity statistics and bootstrapping

After the burn-in period, we measured genetic diversity (π) and singleton density (ψ; the number of singletons observed within a locus) within 10 kb windows across the 200 kb neutral locus every 50 generations using a random sample of 400 chromosomes. π and ψ was measured for each demographic model by taking the mean of these values across each set of 5,000 replicate simulations. In the context of measuring these statistics under neutral simulations, we annotated π and ψ as π_0 and ψ_0, respectively. We took the ratio of these statistics (i.e., π/π_0 and ψ/ψ_0) in order to measure the relative impact of BGS within each demographic model. We bootstrapped the diversity statistics by sampling with replication the 5,000 simulated replicates of each demographic model to generate a new set of 5,000 simulations, taking the mean of π and ψ across each new bootstrapped set. 10,000 bootstrap iterations were conducted and we generated confidence intervals from the middle 95% of the resulting bootstrapped distribution.

Calculations of expected BGS

To calculate the predicted π/π_0 given the specific N_e at each time point for each demographic model, we used equation 14 of Nordborg et al. 1996 [7], but modified it accordingly to incorporate two gamma distributions of fitness effects. Additionally, in order to

properly model our simulations, we only calculated the effects of BGS on one side of the selected locus. This resulted in the following modified equation:

$$\frac{N_e}{N} \equiv \frac{\pi}{\pi_0} = \exp\left(-\frac{U_{torg}}{2R}\int_{trunc}^{\infty}\frac{1}{s}\left\{\int_0^R \frac{dz}{[1+r(z)(1-s)/s]^2}\right\} Gamma\left(s,\ \alpha_{torg},\ \beta_{torg}\right)ds\right) \times$$

$$\exp\left(-\frac{U_{boyko}}{2R}\int_{trunc}^{\infty}\frac{1}{s}\left\{\int_0^R \frac{dz}{[1+r(z)(1-s)/s]^2}\right\} Gamma\left(s,\ \alpha_{boyko},\ \beta_{boyko}\right)ds\right)$$

Here, R is the total length of the selected locus, U is the total deleterious mutation rate across the selected locus, $r(z)$ is the genetic map distance between a neutral site and a deleterious mutation, and s is the selection coefficient of a deleterious mutation. The left side of the equation models the effects of BGS according to the gamma DFE inferred by Ref. [66] (represented by $Gamma\left(s,\ \alpha_{torg},\ \beta_{torg}\right)$) and the right side of the equation models the effects of BGS according to the gamma DFE inferred by Ref. [67] (represented by $Gamma\left(s,\ \alpha_{boyko},\ \beta_{boyko}\right)$).

Because N_e is not explicitly included in this model of BGS, we truncated selection at some value $trunc$, such that $trunc = \gamma/2N_e$ (represented in the integral $\int_{trunc}^{\infty} ds$). Here, $trunc$ represents the minimum selection coefficient which is treated as deleterious for the model and γ represents the population scaled selection coefficient ($\gamma = 2N_e s$) that determines the value of $trunc$. Thus, this step excludes effectively neutral mutations from the model that should not contribute to BGS. This truncation step also affects the values used for U in the above equation, resulting in specific values of U for each DFE. This truncation method was also used in previous studies measuring the expectation of BGS across the genome [12,68]. We simulated different population sizes under our BGS simulation model to see how well the modified Nordborg model fit populations of different N_e for different values of γ (Figure B.21 in Appendix B). We used a γ

= 0.15 because this provided the best estimate of π/π_0 for the starting N_e of our demographic models (i.e., N_e = 20,000). While this value provides a coarse estimate for the effects of BGS on π/π_0 for a particular N_e, it will overestimate the effects of BGS for smaller N_e (Figure B.21 in Appendix B).

REFERENCES

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

2. Nei M, Maruyama T, Chakraborty R. The bottleneck effect and genetic variability in populations. Evolution. 1975;29: 1–10. doi:10.2307/2407137

3. Maruyama T, Fuerst PA. Population bottlenecks and nonequilibrium models in population genetics. I. Allele numbers when populations evolve from zero variability. Genetics. 1984;108: 745–763.

4. Maruyama T, Fuerst PA. Population bottlenecks and nonequilibrium models in population genetics. II. Number of alleles in a small population that was formed by a recent bottleneck. Genetics. 1985;111: 675–689.

5. Kaplan NL, Hudson RR, Langley CH. The “hitchhiking effect” revisited. Genetics. 1989;123: 887–899.

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

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

8. Hudson RR, Kaplan NL. Deleterious background selection with recombination. Genetics. 1995;141: 1605–1617.

9. Tajima F. The effect of change in population size on DNA polymorphism. Genetics. 1989;123: 597–601.

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

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

12. Comeron JM. Background selection as baseline for nucleotide variation across the drosophila genome. PLoS Genet. 2014;10: e1004434. doi:10.1371/journal.pgen.1004434

13. Charlesworth B. Background selection and patterns of genetic diversity in Drosophila melanogaster. Genet Res. 1996;68: 131–149. doi:10.1017/S0016672300034029

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

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

16. 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. doi:10.1371/journal.pgen.1006130

17. 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. doi:10.1093/molbev/msr225

18. 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. doi:10.1038/nbt.2050

19. Williamson RJ, Josephs EB, Platts AE, Hazzouri KM, Haudry A, Blanchette M, et al. Evidence for widespread positive and negative selection in coding and conserved voncoding regions of capsella grandiflora. PLoS Genet. 2014;10: e1004622. doi:10.1371/journal.pgen.1004622

20. 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. doi:10.1038/nplants.2016.84

21. 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. doi:10.1038/nature01027.1.

22. Reed FA, Akey JM, Aquadro CF. Fitting background-selection predictions to levels of nucleotide variation and divergence along the human autosomes. Genome Res. 2005;15: 1211–1221. doi:10.1101/gr.3413205

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

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

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

26. 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. doi:10.1126/science.1198878

27. Lohmueller KE, Albrechtsen A, Li Y, Kim SY, Korneliussen T, Vinckenbosch N, et al. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. PLoS Genet. 2011;7: e1002326. doi:10.1371/journal.pgen.1002326

28. Kimura M. The neutral theory of molecular evolution. Cambridge: Cambridge University Press; 1983. doi:10.1017/CBO9780511623486

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

30. 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. doi:10.1371/journal.pbio.1001388

31. Doebley JF, Gaut BS, Smith BD. The molecular genetics of crop domestication. Cell. 2006;127: 1309–1321. doi:10.1016/j.cell.2006.12.006

32. Tang H, Sezen U, Paterson AH. Domestication and plant genomes. Curr Opin Plant Biol. 2010;13: 160–166. doi:10.1016/J.PBI.2009.10.008

33. Wiener P, Wilkinson S. Deciphering the genetic basis of animal domestication. Proc R Soc B. 2011;278: 3161–3170. doi:10.1098/rspb.2011.1376

34. Gaut BS, Seymour DK, Liu Q, Zhou Y. Demography and its effects on genomic variation in crop domestication. Nat Plants. 2018;4: 512–520. doi:10.1038/s41477-018-0210-1

35. Elton CS. Periodic Fluctuations in the Numbers of Animals: Their Causes and Effects. J Exp Biol. 1924;2: 119–163.

36. Ives PT. Further Genetic Studies of the South Amherst Population of Drosophila melanogaster. Evolution. 1970;24: 507–518. doi:10.2307/2406830

37. Itoh M, Nanba N, Hasegawa M, Inomata N, Kondo R, Oshima M, et al. Seasonal Changes in the Long-Distance Linkage Disequilibrium in Drosophila melanogaster. J Hered. 2010;101: 26–32. doi:10.1093/jhered/esp079

38. Norén K, Angerbjörn A. Genetic perspectives on northern population cycles: bridging the gap between theory and empirical studies. Biol Rev. 2014;89: 493–510. doi:10.1111/brv.12070

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

40. Dlugosch KM, Parker IM. Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. Mol Ecol. 2008;17: 431–449. doi:10.1111/j.1365-294X.2007.03538.x

41. David JR, Capy P. Genetic variation of Drosophila melanogaster natural populations. Trends Genet. 1988;4: 106–111.

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

43. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015;526: 68–74. doi:10.1038/nature15393

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

45. Gillespie JH. Is the population size of a species relevant to its evolution? Evolution. 2001;55: 2161–2169.

46. Santiago E, Caballero A. Joint prediction of the effective population size and the rate of fixation of deleterious mutations. Genetics. Genetics; 2016;204: 1267–1279. doi:10.1534/genetics.116.188250

47. Gillespie JH. Genetic drift in an infinite population: The pseudohitchhiking model. Genetics. 2000;155: 909–919.

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

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

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

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

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

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

54. Cvijović I, Good BH, Desai MM. The Effect of Strong Purifying Selection on Genetic Diversity. Genetics. Genetics; 2018;209: 1235–1278. doi:10.1534/genetics.118.301058

55. Crow JF, Kimura M. An introduction to population genetics theory. New York: Harper and Row; 1970.

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

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

58. Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, Hubisz MJ, et al. Proportionally more deleterious genetic variation in European than in African populations. Nature. 2008;451: 994–997. doi:10.1038/nature06611

59. Simons YB, Turchin MC, Pritchard JK, Sella G. The deleterious mutation load is insensitive to recent population history. Nat Genet. 2014;46: 220–4. doi:10.1038/ng.2896

60. 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. doi:10.1073/pnas.1510805112

61. 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. doi:10.1016/j.gde.2016.09.006

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

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

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

65. Thornton KR. A C++ template library for efficient forward-time population genetic simulation of large populations. Genetics. 2014;198: 157–166. doi:10.1534/genetics.114.165019

66. 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. doi:10.1371/journal.pgen.1000592

67. Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, et al. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. 2008;4: e1000083. doi:10.1371/journal.pgen.1000083

68. Charlesworth B. The role of background selection in shaping patterns of molecular evolution and variation: evidence from variability on the Drosophila X chromosome. Genetics. 2012;191: 233–246. doi:10.1534/genetics.111.138073

Chapter 4:

Distortions to the site-frequency spectrum under background selection and its impact on demographic inference

INTRODUCTION

Uncovering the evolutionary and demographic history of modern humans by observing patterns of genetic variation has been a long-standing pursuit in the field of human population genetics ([1–10]). Much of the early work in this effort has relied on data captured from common variation. Because much of this common variation is old, it has had limited ability to help researchers explore the most recent epochs of human history. In contrast, rare variants, which are on average younger than common variants [11,12], may offer more signal for uncovering recent evolutionary and demographic events. However, because these variants are by nature rare, their ascertainment requires large samples sizes. With recent studies now sequencing thousands of human genomes [13–15], human genetic variation can be investigated at finer levels to uncover events in the more recent periods of human history. This includes the fine scale relationships between human populations [16,17], natural selection over the past few thousand years [18], and the age distribution of rare alleles [12,19,20].

The rapid population growth of humans in their recent history has injected a bevy of rare genetic variation into the genome necessitating the sequencing of thousands of samples in order to uncover much of it [21–23]. But because such rare variation is a signature of population growth, the frequency distribution of mutations (known as the site-frequency spectrum [SFS]) in the human genome can also be leveraged to measure the rate and size of recent human expansion. By fitting models of demography to the SFS, early studies conducted on limited sample sizes found evidence in the human genome for recent population growth [22,24–27]. With recent studies sequencing thousands of samples with the ability to detect even greater amounts of rare variation, demographic inference has yielded much stronger evidence for the explosive growth of humans [28–30]. However, these studies also ascertained neutral polymorphisms from genes and exomes and may be plagued by biases introduced from selection at linked sites (i.e., background selection [BGS] [31] and genetic hitchhiking [32]). While standard practice limits the

use of sites for demographic inference to those that are selectively neutral (because natural selection can heavily skew the SFS towards rare variants and confound meaningful inference), the effects of natural selection can still be felt at nearby neutral variants through physical linkage. Notably, this process of selection at linked sites can also generate skews in the SFS towards rare variants [31,33], in effect mimicking population expansions. This has motivated simulation studies demonstrating that demographic inference can be significantly biased by selection at linked sites [34,35] and provides caution for interpreting results generated from sites likely to be influenced by this process.

In an effort to avoid biases generated by selection at linked sites, studies from the past few years have focused on performing demographic inference in regions of the genome where selection at linked sites is expected to be minimal [36–39]. Although these studies were performed at sample sizes much lower than the previous studies using genes or the exome [28,29], they found that results can be significantly biased when garnered from regions affected by selection at linked sites. For example, the studies of Torres et al. 2018 [37] (Chapter 2) and Ragsdale et al. 2018 [38] both fit an exponential growth model to human data from fourfold degenerate synonymous sites and compared results to those from regions under weak selection at linked sites. Both found a greater inferred population growth for fourfold degenerate sites, recapitulating the biased results expected from BGS [34].

While the biases that selection at linked sites introduces to demographic inference has received recent attention, there has not been thorough investigation about how such biases may scale with sample size. Theoretically, sample size should have strong relevance for how the proportion of rare variants changes as a function of selection at linked sites, especially in the context of BGS. While the classical result of BGS predicts a decrease in genetic diversity in an equilibrium population, this main result does not account for the fact that younger, rare mutations will be less affected by BGS relative to older, common ones. This is because selection itself is a

time dependent process and there is a higher probability that newly arisen mutations will not be affected by selection. As these mutations become older, though, this is less likely to be true. Thus, the external branches of a gene genealogy under BGS will be relatively long when compared to older internal branches [40,41]. In essence, the number of young mutations existing at the lowest frequencies in regions of BGS are likely to be similar to the number of young mutations in regions absent of BGS [42]. We expect the opposite, though, for older, common mutations. This result of a dearth of common mutations but a relative increase in rare mutations skews the SFS in a way that leads to the biases in demographic inference previously described. However, in a demographic scenario where the vast majority of mutations are expected to be young and rare (such as from a recent population expansion) the disparity in the proportion of rare alleles between regions of strong and weak BGS may be less apparent. Thus, with increasing sample sizes from a population experiencing rapid growth, the proportion of rare variants in the lowest frequency bins in regions with and without BGS may become more similar to one another. If this result is true, then biases in demographic inference, especially for those models where rare frequency bins are expected to be most important, may also become less apparent.

In order to test whether changes in sample size affect differences in the SFS as a function of BGS, we analyzed the genomes of 2,416 Europeans sequenced to high coverage. We also conducted demographic inference by fitting a simple model of exponential growth utilizing either fourfold degenerate sites or sites inferred to be under the weakest effects of BGS. Doing so, we were able to observe that increasing the sample size for demographic inference affects apparent biases in inferred growth. Interestingly, our results indicate that the largest biases in demographic inference exist for smaller sample sizes, not larger ones. Across regions of varying strength of BGS, we also observed that the proportion of variants in the rarest bins of the SFS become more similar as sample size increases, thus displaying an important consequence of the recent explosive growth of the human population.

RESULTS

Sample size changes the relative effect of BGS on the site-frequency spectrum

To test how increasing sample size changes the shape of the SFS as a function of BGS, we split the genome into percentile bins of *B* (an inferred measure of the strength of BGS) and measured the unfolded SFS within each bin (see Materials and Methods). As expected, we found that BGS increased the frequency of rare variants in the genome. Specifically, among singletons, doubletons, and tripletons in the SFS, we observed a distinct trend of increasing frequency as *B* decreased (i.e., as BGS strength increased) regardless of sample size (Figure 4.1). In addition, larger sample sizes yielded overall greater frequencies of these rare variants across all bins of *B.* The latter result was not unexpected given the recent accelerated exponential growth of humans [43]. However, when observing how the effect of BGS on the SFS varied for different sample sizes, we found that the largest differences between low *B* (strong BGS) and high *B* (weak BGS) occurred when the sample size was low (Figure 4.1). For a sample size of 100 chromosomes, the proportion of singletons in the lowest 1% *B* bin was 0.349 and the proportion of singletons in the highest 1% *B* bin was 0.22, a difference in proportion of 0.129. However, when making the same comparison using 1000 chromosomes and 4832 chromosomes, the differences between the extreme *B* bins were less (0.114 and 0.076, respectively). Similar, albeit much smaller, trends were also observed among doubletons and tripletons. The difference between the extreme percentile bins of *B* for doubletons for sample sizes of 100 and 4832 chromosomes was 0.0198 and 0.00578, respectively. For tripletons, this difference was only 0.0047 and 0.0019, respectively.

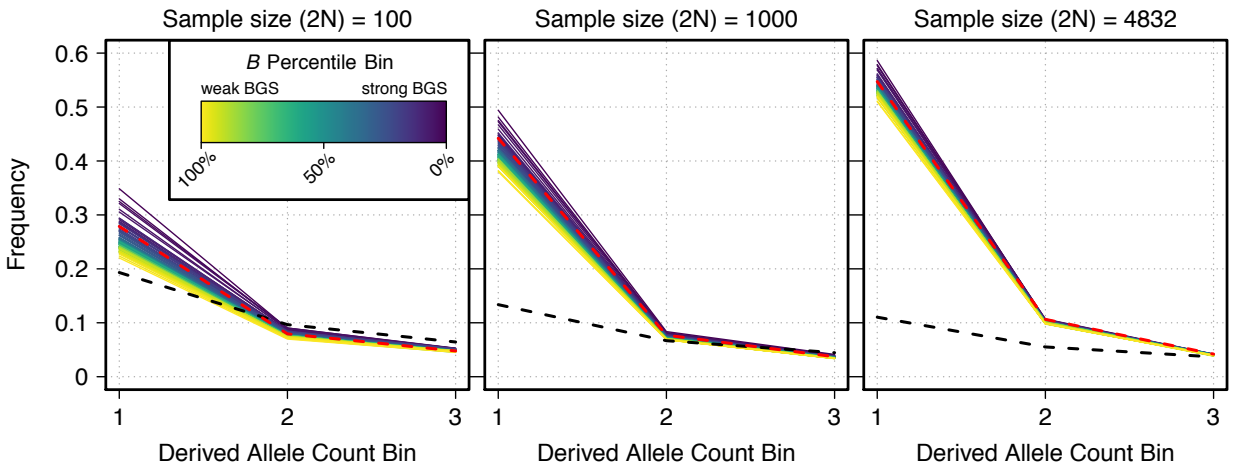


Figure 4.1. Site-frequency spectrum (SFS) for different sample sizes and *B* for the first three derived allele counts.
SFS data is shown for each of 100 percentile bins of *B* (higher percentiles indicate weaker BGS). Each separate plot shows a different sample size from which the SFS was made. Dashed red lines show the SFS from fourfold degenerate sites. Dashed black lines show the SFS from a standard neutral model for the given sample size.

To better illustrate how the relative change in the frequency of these rare variants increased as a function of decreasing sample size, we measured the relative increase of the frequencies of singletons, doubletons, and tripletons across bins of *B* for sample sizes of 100 chromosomes to 4832 chromosomes in increasing sets of 100 chromosomes each (see Materials and Methods). This revealed a striking pattern of a greater enrichment of singletons, doubletons, and tripletons as *B* decreased that became further magnified with lower sample sizes (Figure 4.2). These results demonstrated that the distortions in the rare variant bins of the SFS that are generated by BGS may be exacerbated in humans for small sample sizes but ameliorated as sample sizes increase. Despite these striking patterns among rare variants, when summarizing the entire SFS by measuring average pairwise genetic diversity (π) across the smallest and largest sample sizes, we observed virtually no differences (Figure C.1 in Appendix C). As expected, we did observe a strong association between increasing BGS strength (decreasing *B*) and π. However, this trend was only approximately linear between *B* values of 0.9 and 0.0 (Figure C.1 in

Appendix C), demonstrating that this estimate of BGS is not exact since π should be proportional to *B* for all values [44,45]. But when we observed the full SFS as a function of *B*, two expected results [42] became apparent: 1) a characteristic J shape in the SFS and 2) a slight increase in the proportion of variants in the largest derived allele count bin (Figure C.2 in Appendix C).

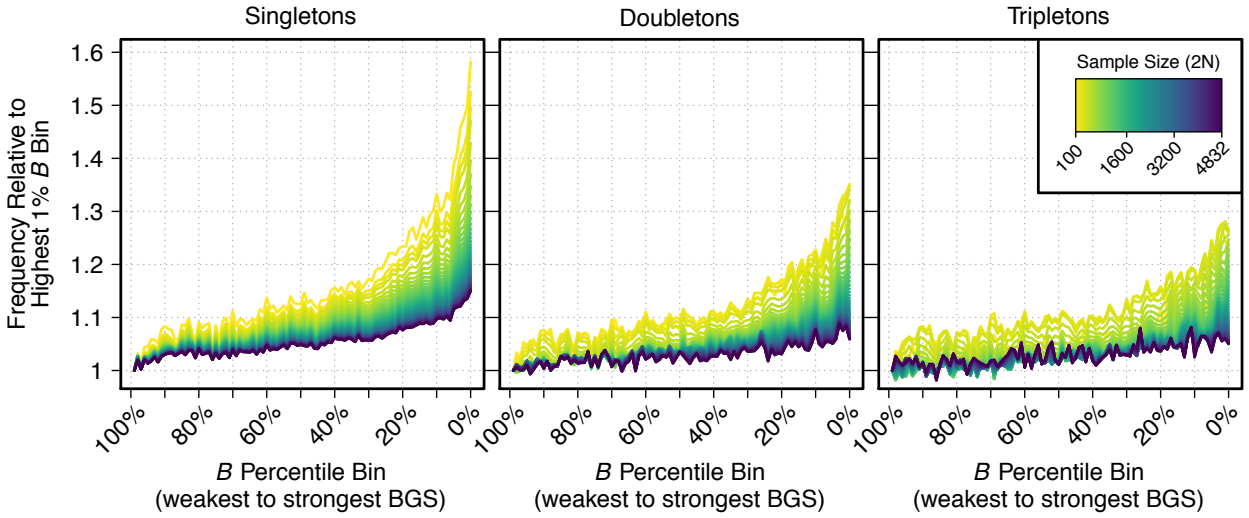


Figure 4.2. Relative increase in singletons, doubletons, and tripletons across *B*.
The relative increase in singletons, doubletons, and tripletons was measured for decreasing percentile bins of *B* (increasing BGS) relative to the highest 1% *B* bin (the weakest BGS bin). These relative increases are plotted for different sample sizes, which are shown in the legend.

Biases in demographic inference due to BGS decrease with sample size

To observe how the change in the frequency of rare variants as a function of both *B* and sample size affects demographic inference, we performed inference by fitting the SFS to a model of exponential growth by using the program *moments* [46] (see Materials and Methods). Inference was run using sample sizes of 1000, 2000, 3000, 4000, and 4832 chromosomes. For performing the inference, we used two sets of sites: filtered sequence data from the highest 1% *B* bin (where BGS is expected to be weakest) and fourfold degenerate sites (see Materials and Methods). Because they are expected to be neutral, fourfold degenerate sites are commonly used sites for performing demographic inference [22,29,47]. However, they are also expected to be under a

significant amount of BGS since they lie within coding regions. Thus, their use may generate biased results.

When we performed inference on 1000 chromosomes, we observed differences in the inferred growth between the two sets of sites. The ending population sizes after exponential growth (denoted as the parameter N_{Eur}) were inferred to be 1,163,554 and 755,334 for the highest 1% *B* bin sites and fourfold degenerate sites, respectively. In addition, the respective 95% confidence intervals for these two sets of sites did not overlap (Figure 4.3; Figure C.3 and Table C.1 in Appendix C). However, as the sample size used for performing inference increased, the inferred ending population sizes became more similar between the two sets of sites; at sample sizes of 2000 chromosomes or greater, 95% confidence intervals for N_{Eur} overlapped. The differences in the SFS used to fit the model also diminished with increasing sample sizes (Figures C.3-C.5 in Appendix C). Interestingly, when using sites in the highest 1% *B* bin, the relative change in N_{Eur} diminished after increasing the sample size for performing inference to 2000 chromosomes or greater, suggesting that more samples are unlikely to change parameter estimates for the most neutral sites in the genome. However, for fourfold degenerate sites, the inferred growth continued to increase, albeit at a diminishing rate, as sample size also increased.

The time for when growth started was also inferred to be more recent when using sites in the highest 1% *B* bin compared to fourfold degenerate sites, regardless of the sample size used. But this inferred time also remained very stable for sites in the highest 1% *B* bin across the different samples sizes (Table C.1 in Appendix C). In contrast, with increasing samples sizes, inference utilizing fourfold degenerate sites generated more recent estimates for the starting time of exponential growth. The pattern of both higher inferred growth and more recent inferred growth as sample sizes increased also increased the growth rate when using fourfold degenerate sites (Table C.1 in Appendix C). These growth rates, though, never exceeded those when using sites in the highest 1% *B* bin (which also remained stable in growth rate as sample size increased). But

for the largest sample sizes, the growth rates from using the two types of sites became more similar.

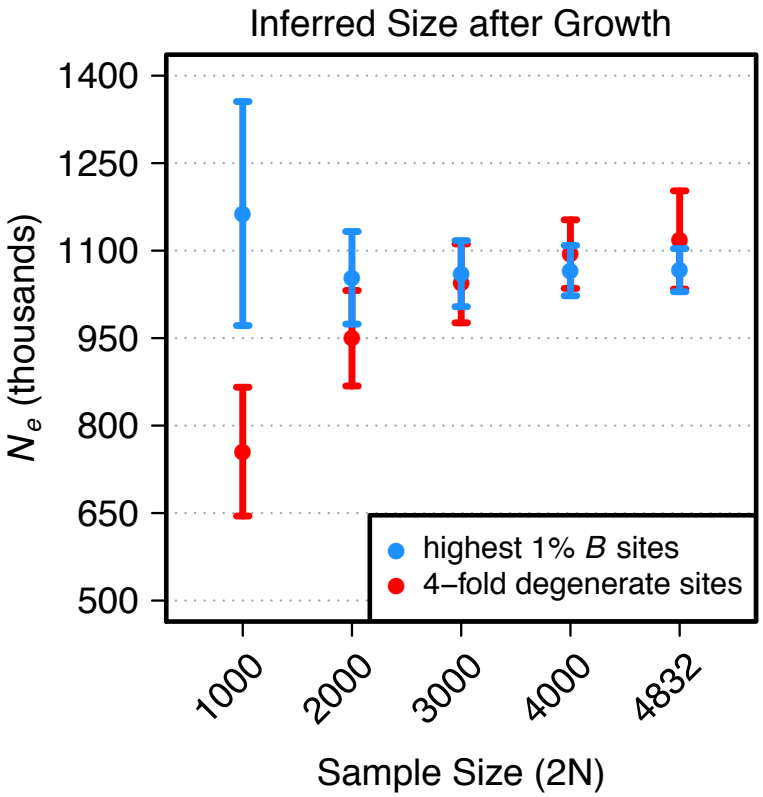


Figure 4.3. Inferred population size after exponential growth from demographic inference. Inferred population size after exponential growth (N_{Eur}) calculated using the inferred θ and a mutation rate of 1.66 x 10^{-8} (see Materials and Methods). Demographic inference was conducted with different sample sizes and using different parts of the genome (see legends). Whiskers show 95% confidence intervals. See Table C.1 in Appendix C for parameter values.

The results described thus far on inferred growth account for the fact that the inferred θ (population scaled mutation rate) is smaller for fourfold degenerate sites. This is expected since selection at linked sites depletes genetic diversity in the genome. Yet, when controlling for θ by comparing the relative starting and ending population sizes, we found that as sample size increased, the relative increase inferred using fourfold degenerate sites surpasses the relative increase inferred using sites in the highest 1% *B* bin (Table C.2 in Appendix C). This result agrees

with the result of Ewing et al. 2016 [34]. However, the rate of increase for fourfold degenerate sites also diminished as the sample size increased.

DISCUSSION

Our work provides an interesting result about the effects of increasing sample size on observing distortions in the SFS under BGS in humans. The expected skew towards rare variants in the SFS under BGS motivated earlier work on understanding how biases are introduced to demographic inference [34,37,38]. Since a skew towards rare variants is better detected at larger samples sizes, on first principles it is not unreasonable to assume that larger samples sizes should introduce even greater distortions to the SFS and contribute to even stronger biases during inference. For a constant sized population, such an assumption is likely valid. However, as demonstrated here, for populations experiencing recent rapid growth, this is not necessarily the case. Rather, large sample sizes are capturing a bevy of young, rare variation that make up a much larger proportion of the SFS than expected under a constant sized population model. In effect, this minimizes the differences in the SFS between regions of weak and strong BGS. In contrast, with smaller sample sizes much of this recently arisen rare variation will not be captured and contributions to the SFS will largely consist of older variants that existed prior to the most recent epoch of human growth. As we showed in our results, this leads to the stronger effects of BGS causing greater skews towards rare variants. In the context of demographic inference, this makes biases more apparent for small sample sizes. In contrast, with larger sample sizes, differences in the inferred values of recent growth between regions of strong and weak BGS may be small enough that they fall well within the distribution of possible inferred estimates (Figure 4.3). While we do not argue that previously inferred values of recent human growth from four-fold degenerate sites represent true estimates [28–30], those estimates may still lie within the confidence intervals of the inferred values from sites under weak BGS. However, differences in the number of populations fit during inference, the number of model parameters, and the exact

inference procedure may limit the extension of our results to these other studies. Despite this, the growth rates from our model of exponential growth in Europeans (~1.95%; Table C.1 in Appendix C) do fair quite comparably to those inferred by Refs. [29] and [30] (1.95% and 1.7%, respectively).

In order to make results more interpretable, we rely on a simple three parameter demographic model of exponential growth, allowing only for two free parameters: the relative size of change after growth and the time span over which growth occurred. While our exponential growth model is a common model for recent human history and has been used in much of the previous work on demographic inference [22,36,38,48], it ignores more ancient demographic parameters which rely on signal from more common variants in the SFS. Previous studies of human demography also incorporated population bottlenecks to account for the population splits and contractions that non-African populations experienced throughout their history [22,29,36,46,48]. These more ancient parameters are also likely to be biased because BGS shrinks the coalescent histories and the effective population size of a region of the genome through time. Thus, inferred human bottlenecks and other ancient population size changes are likely to be underestimated in the presence of BGS (i.e., the inference of both stronger and more recent bottlenecks). We ignored these ancient parameters because our focus on recent population growth in humans and its patterning of variation as a function of increasing sample size should not have as much of an effect on common variation (as evident from the unchanging values of π with changing sample size; Figure C.1 in Appendix C). Additionally, because we are only making inference using a single population, we are less likely to accurately model ancient parameters that rely on historical migration between Europeans and other human populations (such as Asians and Africans) which can contribute significantly to common variation. In contrast, most of the variation that has recently arisen in human history is population-specific and is less likely to have been impacted by migration between populations [19]. Therefore, we believe that in

the context of the large sample sizes of our study, a simple exponential growth model that fits a single population undergoing recent growth allows for better interpretation and, biologically, is more parsimonious than models incorporating additional ancient parameters.

For demographic inference procedures that rely more on common variation and are able to detect more ancient demographic events, such as coalescent hidden markov models (HMMs) [49–52], biased inference resulting from BGS is also likely. Moreover, in the context of genetic hitchhiking via selective sweeps, evidence for biased inference from coalescent hidden markov models (HMMs) has also been demonstrated [35]. Thus, the effects of BGS on both rare and common variation may plague inference even for those methods that do not rely on the SFS.

Our results are limited by the fact that we only focus on one human population: Europeans. Other humans populations, such as East Asians and Africans, are known to have different rates of growth when compared to Europeans – namely a higher rate for East Asians and a lower rate for Africans [22,29]. Because of this, apparent biases and skews in the SFS under BGS for East Asians may diminish in sample sizes even lower than what is shown here for Europeans. However, the opposite effect may exist for Africans. More research into such effects in other human populations will be needed in order to support the overall conclusions presented here. Finally, a more complex and realistic model of human demography that incorporates other populations (such as the models of Refs. [22,48]) will not only allow for the inference of more ancient parameters, but will yield important information on how BGS impacts migration parameters between populations. Unfortunately, current diffusion-based methods fitting a joint SFS with continuous migration do not scale well since the joint SFS is cubic with sample size for three populations [46,48], although coalescent-based methods using pulse migration may be more tractable for larger sample sizes [53].

In conclusion, we demonstrate that the recent demographic history of humans has underappreciated consequences on the SFS in regions of selection at linked sites, with

implications for demographic inference. These results add to previous ones showing that models of BGS, and their predictions on affecting variation within the genome, do not always apply straightforwardly in populations with a dynamic demographic history [37]. As other populations and species have their genomes sequenced into the thousands, observed patterns of variation may reveal unexpected consequences as well. In general, observations of more genomic data from diverse populations will offer population geneticists new insight into the interplay of natural selection and demography and how they interact to pattern variation along the genome.

MATERIALS AND METHODS

Sample selection

In order to sample individuals with a high percentage of European ancestry and to prevent confounds introduced by population structure, we used two separate ascertainment schemes for selecting individuals in our study. First, the program RFMix [54] was run on 18,436 samples from the TOPMed consortium (freeze 3) using the following parameter settings: PopPhased --num-threads 1 --min-node-size 5. For the reference panel, 938 samples from the Human Genome Diversity Panel (HGDP) was used. The 53 populations of HGDP were condensed into 7 super-populations: 1) Sub-Saharan African (n=104), 2) Central and South Asian (n=200), 3) East Asian (n=229), 4) European (n=154), 5) Native American (n=63), 6) Oceanian (n=28), and 7) Middle Eastern (n=160). After running RFMix, we summed local ancestries assigned for each TOPMed sample to create a vector of global ancestries corresponding to the 7 HGDP super-populations. We then selected individuals that had greater than or equal to 90% global European ancestry (Figure C.6 in Appendix C). In addition, we also performed principal components analyses (PCA) on a set of 18,234 pre-selected individuals from the TOPMed consortium (i.e., agnostic to RFMix global ancestry percentages). We then used k-means clustering to cluster individuals on the first 7 PCs, using a cluster number of k=9. We then ran the clustering algorithm with 250 restarts. Based on the k=9 clustering, we selected all individuals within the cluster consisting of samples having the highest mean proportion of global European ancestry from the RFMix results. We then took the union of the two sets of samples from the two ascertainment steps (i.e., samples greater than or equal to 90% global European ancestry and also belonging to the k-means cluster having the highest mean European ancestry). We further selected individuals that were unrelated and gave consent for performing population genetics research. This resulted in 2,416 total individuals for use in our study. When measuring the site-frequency spectrum across these samples as a function of

sample size, we sampled progressively larger random samples of 50 individuals (100 chromosomes) each, generating 49 discrete sample sizes (2N=100, 200, 300…4800, 4832).

Site filtering/ascertainment

In order to perform inference using a high-quality set of neutral sites that are least influenced by the direct effects of natural selection and putative selective sweeps and to avoid potential sequence/mapping error, we performed several steps to filter the genome. Many of these filtering steps were based off of the ascertainment scheme used by Torres et al. 2018 [37] (Chapter 2). Specifically, the following filters were applied (all filters are in hg19 and only autosomes were kept for analyses):

1. Coding regions: coding exons annotated in the UCSC known genes track (table: knownGene, track :UCSC Genes) were removed.
2. phyloP: Sites with phyloP [55] 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) [56] 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 [57] were removed.
6. Simple repeats: Regions in the UCSC simple repeats track were removed due to potential misalignments with outgroups and/or being under natural selection.
7. Gaps/centromeres/telomeres: Regions in the UCSC gap track were removed,

including centromeres and telomeres.

8. Segmental duplications: Regions in the UCSC segmental dups track [58] were removed to limit potential effects of natural selection and/or misalignments with rhesus macaque.
9. Transposons: Active transposons (HERVK retrotransposons, the AluY subfamily of Alu elements, SVA elements, and L1Ta/L1pre-Ta LINEs) in the human genome were removed.
10. Recent positive selection: Regions inferred to be under hard and soft selective sweeps (using iHS and iHH12 regions from selscan v1.2.0 [37,59]); within Thousand Genomes phase 3 [13] European and African populations were removed.
11. Non-coding transcripts: Non-coding transcripts from the UCSC genes track were removed to limit potential effects of natural selection.
12. GC-biased gene conversion (gBGC): Regions in UCSC phastBias track [60] from UCSC genome browser were removed to limit regions inferred to be under strong GC-biased gene conversion.
13. 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 [61] were removed. 1.5 kb flanking regions surrounding the center of hotspots identified by Ref. [62] (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).

Positions in the genome were then annotated for background selection (BGS) by using the background selection coefficient, *B* [45] (downloaded from http://www.phrap.org/othersoftware.html). *B* varies between 0 and 1, with BGS increasing in strength as values approach 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 used all sites annotated with a *B* value for performing general analyses. However, when performing demographic inference, we only focused our analyses on those regions of the genome within the top 1% of the genome-wide distribution of *B* (*B* >= 0.994). These sites correspond to regions of the genome inferred to be under the weakest amount of BGS.

Sites in the genome were also polarized to ancestral and derived states using ancestral annotations called with high-confidence from the GRCh37 e71 ancestral sequence (downloaded from: ftp://ftp.ensembl.org/pub/release-71/fasta/ancestral_alleles/homo_sapiens_ancestor_GRCh37_e71.tar.bz2) from Ensembl [63], which used a multiple species alignment of 6 primates to infer the ancestral state using the Enredo-Pecan-Ortheus (EPO) pipeline [64,65]. All of the filtering steps described, including the annotation for *B* and polarization for ancestral/derived state, left 1,377,691,456 sites within the genome for use in our study, including 10,977,437 sites with *B* >= 0.994. Finally, we filtered polymorphic sites within the filtered genome on being di-allelic only. This left a total 20,324,704 polymorphic sites across the 2,416 European samples, including 191,631 polymorphic sites that had B >= 0.994.

To generate a set of four-fold degenerate synonymous sites, all coding sites within the genome were annotated using the program ANNOVAR [66] using Gencode V19 annotations. This resulted in 5,188,972 total sites. 4,718,653 sites were left after filtering for high-confidence

ancestral/derived states, of which 91,177 were polymorphic (di-allelic) across the 2,416 European samples.

Demographic inference

We performed demographic inference using the program *moments* [46], which fits a specified demographic model to an observed site-frequency spectrum. For our study, we specified a model of exponential growth with three total parameters (N_{Eur0}, N_{Eur}, T_{Eur}). This included two free parameters: the starting time of exponential growth (T_{Eur}) and the ending population size after growth (N_{Eur}). The ancestral size parameter (i.e, the population size when growth begins), N_{Eur0}, was kept constant in our model such that the relative starting size of the population was always 1. We applied the inference procedure to the 2,416 European samples using either fourfold degenerate sites or sites where $B >= 0.994$. The site-frequency spectrum used for inference was unfolded based on the polarization step described above. The inference procedure was fit using sample sizes (2N) of 1000, 2000, 3000, 4000, and 4832 samples (i.e., chromosomes). The inference procedure was run from different initial starting points hundreds of times for each sample size and dataset to ensure convergence on a global optimum. Attempts at using samples sizes smaller than 2N=1000 for inference resulted in convergence issues, likely because of poor model fit.

To convert the scaled genetic parameters output by the inference procedure *moments* to physical units, we used the resulting θ (population scaled mutation rate; also inferred by *moments*) and a mutation rate of 1.66×10^{-8} [67] to generate corresponding effective population sizes. In order to account for the fact that fourfold degenerate sites and sites from regions within the highest 1% of *B* are ascertained from different effective sequence lengths, we had to first normalize θ by their corresponding lengths. These lengths were 4,718,653 sites and 10,977,437

sites for fourfold degenerates sites and highest 1% *B* sites, respectively. To convert time to years, we used a generation time of 25 years. 95% confidence intervals were generated by resampling the SFS 1,000 times and using the Godambe Information Matrix to generate parameter uncertainties [68].

REFERENCES

1. Lewontin RC, Krakauer J. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics. 1973;74: 175–195.

2. Cann RL, Stoneking M, Wilson AC. Mitochondrial DNA and human evolution. Nature. 1987;325: 31–36. doi:10.1038/325031a0

3. Hughes AL, Nei M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature. 1988;335: 167–170. doi:10.1038/335167a0

4. Cavalli-Sforza LL, Menozzi P, Piazza A. The History and Geography of Human Genes. Princeton, NJ: Princeton University Press; 1994.

5. Harding RM, Fullerton SM, Griffiths RC, Bond J, Cox MJ, Schneider JA, et al. Archaic African and Asian lineages in the genetic ancestry of modern humans. Am J Hum Genet. 1997;60: 772–789.

6. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, et al. Genetic structure of human populations. Science. 2002;298: 2381–385. doi:10.1126/science.1078311

7. 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. doi:10.1038/nature01027.1.

8. Bamshad M, Wooding SP. Signatures of natural selection in the human genome. Nat Rev Genet. 2003;4: 99–110. doi:10.1038/nrg999

9. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Todd Hubisz M, Glanowski S, et al. Natural selection on protein-coding genes in the human genome. Nature. 2005;437: 1153–1157. doi:10.1038/nature04240

10. Garrigan D, Hammer MF. Reconstructing human origins in the genomic era. Nat Rev Genet. 2006;7: 669–680. doi:10.1038/nrg1941

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

12. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2013;493: 216–220. doi:10.1038/nature11690

13. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015;526: 68–74. doi:10.1038/nature15393

14. Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, et al. The UK10K project identifies rare variants in health and disease. Nature. 2015;526: 82–90. doi:10.1038/nature14962

15. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, et al. Large-scale whole-genome sequencing of the Icelandic population. Nat Genet. Nature Publishing Group; 2015;47: 435–444. doi:10.1038/ng.3247

16. Mathieson I, McVean G. Demography and the age of rare variants. PLoS Genet. 2014;10: e1004528. doi:10.1371/journal.pgen.1004528

17. O'Connor TD, Fu W, Mychaleckyj JC, Logsdon B, Auer P, Carlson CS, et al. Rare variation facilitates inferences of fine-scale population structure in humans. Mol Biol Evol. 2015;32: 653–660. doi:10.1093/molbev/msu326

18. Field Y, Boyle EA, Telis N, Gao Z, Gaulton KJ, Golan D, et al. Detection of human adaptation during the past 2000 years. Science. 2016;354: 760–764. doi:10.1126/science.aag0776

19. Albers PK, McVean G. Dating genomic variants and shared ancestry in population-scale sequencing data. bioRxiv. 2018; 416610. doi:10.1101/416610

20. Platt A, Hey J. Age distributions of rare lineages reveal recent demographic history and selection. bioRxiv. 2018; 265900. doi:10.1101/265900

21. Keinan A, Clark AG. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science. 2012;336: 740–743. doi:10.1126/science.1217283

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. doi:10.1073/pnas.1019276108

23. Gao F, Keinan A. High burden of private mutations due to explosive human population growth and purifying selection. BMC Genomics. 2014;15: S3. doi:10.1186/1471-2164-15-S4-S3

24. Polanski A, Kimmel M. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics. 2003;165: 427–436.

25. Adams AM, Hudson RR. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. Genetics. 2004;168: 1699–1712. doi:10.1534/genetics.104.030171

26. 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.

27. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 2009;5: e1000695. doi:10.1371/journal.pgen.1000695

28. Coventry A, Bull-Otterson LM, Liu X, Clark AG, Maxwell TJ, Crosby J, et al. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nat Commun. 2010;1: 131. doi:10.1038/ncomms1130

29. 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. doi:10.1126/science.1219240

30. Nelson MR, Wegmann D, Ehm MG, Kessner D, St. Jean P, Verzilli C, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science. 2012;337: 100–104. doi:10.1126/science.1217876

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

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

33. 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.

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

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

36. Gazave E, Ma L, Chang D, Coventry A, Gao F, Muzny D, et al. Neutral genomic regions refine models of recent rapid human population growth. Proc Natl Acad Sci. 2014;111: 757–762. doi:10.1073/pnas.1310398110

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

38. Ragsdale AP, Moreau C, Gravel S. Genomic inference using diffusion models and the allele frequency spectrum. Curr Opin Genet Dev. 2018;53: 140–147. doi:10.1016/J.GDE.2018.10.001

39. Pouyet F, Aeschbacher S, Thiéry A, Excoffier L. Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. Elife. 2018;7: e36317. doi:10.7554/eLife.36317

40. Zeng K, Charlesworth B. The joint effects of background selection and genetic recombination on local gene genealogies. Genetics. 2011;189: 251–266. doi:10.1534/genetics.111.130575

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

42. Cvijović I, Good BH, Desai MM. The Effect of Strong Purifying Selection on Genetic Diversity. Genetics. Genetics; 2018;209: 1235–1278. doi:10.1534/genetics.118.301058

43. Keinan A, Clark AG. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science. 2012;336: 740–743. doi:10.1126/science.1217283

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

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

46. Jouganous J, Long W, Ragsdale AP, Gravel S. Inferring the joint demographic history of multiple populations: beyond the diffusion approximation. Genetics. 2017;206: 1549–1567. doi:10.1534/genetics.117.200493

47. Gao F, Keinan A. Inference of super-exponential human population growth via efficient computation of the site frequency spectrum for generalized models. Genetics. 2016;202: 235–245. doi:10.1534/genetics.115.180570

48. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 2009;5: e1000695. doi:10.1371/journal.pgen.1000695

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

50. Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. Nat Genet. 2014;46: 919–925. doi:10.1038/ng.3015

51. Sheehan S, Harris K, Song YS. Estimating variable effective population sizes from multiple genomes: A sequentially markov conditional sampling distribution approach. Genetics. 2013;194: 647–661. doi:10.1534/genetics.112.149096/-/DC1

52. Spence JP, Steinrücken M, Terhorst J, Song YS. Inference of population history using coalescent HMMs: review and outlook. Curr Opin Genet Dev. 2018;53: 70–76. doi:10.1016/J.GDE.2018.07.002

53. Kamm JA, Terhorst J, Durbin R, Song YS. Efficiently inferring the demographic history of many populations with allele count data. bioRxiv. 2018; 287268. doi:10.1101/287268

54. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. Am J Hum Genet. 2013;93: 278–288. doi:10.1016/j.ajhg.2013.06.020

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

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

57. 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. doi:10.1038/nature11247

58. 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. doi:10.1101/gr.187101

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

60. 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. doi:10.1371/journal.pgen.1003684

61. 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. doi:10.1038/nature06258

62. 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. doi:10.1126/science.1256442

63. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, et al. Ensembl 2013. Nucleic Acids Res. 2012;41: D48–D55. doi:10.1093/nar/gks1236

64. Paten B, Herrero J, Fitzgerald S, Beal K, Flicek P, Holmes I, et al. Genome-wide nucleotide-level mammalian ancestor reconstruction. Genome Res. 2008;18: 1829–1843. doi:10.1101/gr.076521.108

65. Paten B, Herrero J, Beal K, Fitzgerald S, Birney E. Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs. Genome Res. 2008;18: 1814–1828. doi:10.1101/gr.076554.108

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

67. 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. The American Society of Human Genetics; 2015;97: 775–789. doi:10.1016/j.ajhg.2015.10.006

68. Coffman AJ, Hsieh PH, Gravel S, Gutenkunst RN. Computationally efficient composite likelihood statistics for demographic inference. Mol Biol Evol. 2016;33: 591–593. doi:10.1093/molbev/msv255

APPENDIX A: Supplemental Material to Chapter 2

Recent admixture has not altered the impact of selection at linked sites

We investigated whether the effects of selection at linked sites have remained consistent across human populations that have experienced recent admixture. To do so, we measured normalized and relative diversity (more precisely, heterozygosity) as a function of *B* in the 6 admixed TGP populations (ASW, ACB, CLM, MXL, PEL, and PUR). We first used local ancestry to divide up admixed samples into genomic segments that are homozygous for a specific local ancestry (i.e., African, European, or Native American). These homozygous ancestral segments are simply regions of the genome in which both maternal and paternal copies of an individual's chromosomes were inferred to have the same ancestral label. To do this, we used the ancestry deconvolution results generated by the 1000 Genomes Project Consortium (see ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20140818_ancestry_deconvolution/README_20140721_phase3_ancestry_deconvolution). Briefly, the local-ancestry inference tool, RFMix [1], was run across the ACB, ASW, CLM, MXL, PEL, and PUR phase 3 TGP samples. For the reference panel, 50 unrelated shapeit2 [2] trio-phased YRI and CEU samples each (from phase 3 TGP) and 43 shapeit2 population-phased Native American samples (from Ref. [3]) were used. We utilized local ancestry tracks that were inferred by RFMix using "trio-phased" mode.

Admixed samples were then parsed for all genomic segments homozygous for each particular ancestry (i.e., African, European, or Native American). These homozygous segments were also filtered according to the 13-filter set described in the "Filtering and ascertainment scheme" section of Materials and Methods in Chapter 2. Heterozygosity was then calculated across admixed samples for each set of homozygous ancestries and *B* quantile bins. Samples were included in this analysis only if the total length of their genome that passed all filters for the particular ancestry and *B* quantile bin was greater than 1 Mb. Additionally, per-site

heterozygosity estimates for each ancestry and *B* quantile set were averaged across all admixed samples, regardless of their TGP population of origin. Heterozygosity was also normalized by divergence with Rhesus macaque (see Materials and Methods in Chapter 2). See Table A.11 for total number of Mb used in these analyses. For comparison, heterozygosity was also calculated across the 4 continental groups using the same 13-filter set and as a function of the same *B* quantile bins.

Across all *B* quantile bins, normalized diversity (heterozygosity/divergence) in African and European ancestry segments closely matched the values observed in their non-admixed counterparts (Figure A.5). However, normalized diversity was significantly lower in the Native American ancestry segments of admixed individuals than in the East Asian continental group (Figure A.5). This was expected given the more recent divergence of Native American populations and the strong population bottleneck they experienced migrating into the Americas [4-6].

Overall, patterns of relative diversity across local ancestries were similar to the broader analyses of the 20 non-admixed populations, with a consistent rank order of decreasing relative diversity observed for African, European, and Native American ancestral segments. However, for relative diversity calculated using the lowest 1% *B* quantile bin (i.e., where selection at linked sites is expected to be strongest), relative diversity in Native American ancestry segments was observed to be greater than for the European continental group or European local ancestry segments, which was inconsistent with the other *B* quantile bins.

Linear regression of F_{ST} on recombination rate and multiple linear regression of F_{ST} on recombination rate and *B*

F_{ST} calculations were performed as a function of 2% recombination rate quantile bins between every pair of non-admixed phase 3 TGP populations in an identical fashion as was

done for *B* (see Materials and Methods in Chapter 2). To do so, we annotated sites based on the recombination rate estimates from the HapMap II GRCh37 genetic map. To annotate sites in phase 3 that were not in HapMap II, recombination rates were interpolated to the midway point between the preceding and following positions in HapMap II. If the difference between successive HapMap II positions was greater than 18,848 base pairs (the first standard deviation for the distribution of distances between positions in HapMap II), then the recombination rate was only extended out 9,424 base pairs beyond the focal position. Positions beyond this distance were then ignored during analysis in which the recombination rate was used. Recombination rate quantiles were calculated using the genome-wide distribution of recombination rates (i.e., the distribution of recombination rates across all sites, including those that are not polymorphic in the data set) resulting from the procedure described above.

Simple linear regression was then conducted using the linear model $F_{ST} = \beta_0 + \beta_1\rho + \varepsilon$ (where ρ is recombination rate). Recombination rate was scaled to be between 0 and 1 (the minimum and maximum observed recombination rate was 0.0 cM/Mb and 126.88 cM/Mb, respectively) to aid in the comparison of the regression coefficient with *B*. Earlier studies using SNP array data have shown that F_{ST} and recombination rate are correlated in humans [7]. We could only partially replicate these findings when we conducted linear regression with the model $F_{ST} = \beta_0 + \beta_1\rho + \varepsilon$. We observed that recombination rate only significantly predicts a change in F_{ST} across the genome for comparisons between South Asian and East Asian populations (Figure A.7, Table A.3). This result remained unchanged when performing robust linear regression for the model (Table A.4).

Since the correlation between F_{ST} and recombination rate was previously documented as being strongest in coding regions [7], where the effects of selection at linked sites are also expected to be strongest, we investigated whether recombination rate provides added value, in

addition to B, as an explanatory variable for predicting F_{ST} by using multiple linear regression. To do so, we first split the genome into 2% recombination rate quantile bins and further subdivided each of these bins into 4% B quantile bins (50×25 = 1,250 bins total). We then measured F_{ST} within each bin. We also partitioned sites in the reverse order (2% B bins followed by 4% recombination rate bins) and repeated all analyses. Our choice of total number of bins resulted in a minimum of 320 SNPs per bin for estimating F_{ST} between any two populations, which should be sufficient to avoid errors when estimating F_{ST} across multiple loci [8]. As with the simple linear regression step, recombination rate was scaled to be between 0 and 1 and the mean of the bounds defining each quantile bin was used when defining the explanatory variables. After performing multiple linear regression of F_{ST} on B, recombination rate (ρ), and an interaction term between the two ($B\rho$) with the linear model $F_{ST} = \beta_0 + \beta_1 B + \beta_2 \rho + \beta_3 B\rho + \varepsilon$, we observed that B was a statistically significant predictor ($p < 1e\text{-}04$) for F_{ST} across all population comparisons regardless of how we partitioned sites (Table A.5). This result remained unchanged when performing robust regression. In contrast, recombination rate exhibited sporadic significance as an explanatory variable for F_{ST} across population comparisons and was dependent upon how sites were partitioned (i.e., whether we first partitioned by B or by recombination rate; Table A.5). Furthermore, strong differences between the two binning schemes were observed for the magnitude of the recombination rate regression coefficient for certain population comparisons (e.g., African vs. East Asian and South Asian vs. East Asian), while the coefficients for B were consistently similar across binning schemes. The direction in which recombination rate explained F_{ST} was also inconsistent across different population comparisons, with European vs. South Asian and European vs. East Asian comparisons showing a significant positive change in F_{ST} as a function of increasing recombination rate. This result was contrary to an expectation of decreasing F_{ST} as a function of increasing

recombination rate [7]. We also failed to observe consistent effects from the interaction term for *B* and recombination rate on F_{ST} across population comparisons or binning schemes (Table A.5). Performing robust regression on the model did not change these results. However, in contrast to recombination rate, when the model was performed utilizing all TGP populations (i.e., the "Global" estimate), the interaction term was significant in explaining F_{ST} across both types of binning schemes.

To aid in visualizing the results of our multidimensional linear model, we plotted F_{ST} for each population comparison as a function of recombination rate (across 4% quantile bins) while conditioning on *B* (Figure A.8). We also plotted points in the reciprocal direction, with F_{ST} being plotted as a function of *B* while conditioning on recombination rate (Figure A.8). These data points were derived from the same points used as input for the multiple linear regression model described above. These specific results for F_{ST} between African and South Asian populations showed that *B* separated different levels of F_{ST} across most recombination rate bins (Figure A.8, Table A.6). Furthermore, regardless of how *B* was conditioned on recombination rate, it still exhibited a strong trend of increasing F_{ST} as it decreased (i.e., in the direction of stronger BGS) (Figure A.8, Table A.7). These patterns were imperfect though, and statistical significance was not always attained, especially for comparisons between non-African populations (Figure A.9, Table A.7). However, greater separation in F_{ST} was generally achieved when conditioning recombination rate on *B* and the slope was always negative when plotting F_{ST} against *B*, regardless of which recombination rate percentile bin *B* was conditioned on.

SFS_CODE command line example

Below is a representative SFS_CODE command for running a simulation of BGS and human demography with 20.46% of sites experiencing deleterious mutation in two 1Mb flanking regions surrounding a neutral 30kb central region. Note that we simulate two distributions of

fitness effects for purifying selection here (see Materials and Methods in Chapter 2 for details). More specifically, this is given by the command -W 2 -0.3394 0.184 0.00040244 0.0415 0.00515625 (see below) where the '-' in front of 0.3394 allows us to draw from a negative gamma distribution with parameters (0.184, 0.00040244) for 33.94% of selected sites and from a negative gamma distribution with parameters (0.0415, 0.00515625) for 66.06% of selected sites. This ability to draw from two negative gamma distributions of fitness effects is a special option not available in the general distribution of SFS_CODE. It is available for download at https://doi.org/10.1371/journal.pgen.1007387.s001.tar.

```
sfs_code 3 1 -A -r 6.0443e-05 -N 18449 -s 1100 -n 100 -TS 0.437017 0 1
-TS 0.546498 1 2 -TE 0.5994242 -Td 0 P 0 2.10709 -Td 0.437017 P 1
0.152957396219 -Td 0.546498 P 1 0.573964845871 -Tg 0.546498 P 1
60.0453856768 -Td 0.546498 P 2 0.221523138739 -Tg 0.546498 P 2
95.5344964867 -Tm 0.437017 P 0 1 6.0846016512 -Tm 0.437017 P 1 0
0.9306848256 -Tm 0.546498 L 0.39374558703 0.104413684315
0.034567778862 0.067228907022 0.0035379117753 0.0259471614066 -t
0.0002650345 -L 11 200000 200000 200000 200000 200000 30000 200000
200000 200000 200000 200000 -v L A 40920 -v L 5 30000 -W 2 -0.3394
0.184 0.00040244 0.0415 0.00515625 -W L 5 0 --printLocus 5 -a N -Tn
0.437017 100 -Tn 0.546498 100 -Tn 0 R 0.00271017399317 100
```

REFERENCES

1. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. Am J Hum Genet. 2013;93: 278–288. doi:10.1016/j.ajhg.2013.06.020

2. O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, et al. A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genet. 2014;10: e1004234. doi:10.1371/journal.pgen.1004234

3. Mao X, Bigham AW, Mei R, Gutierrez G, Weiss KM, Brutsaert TD, et al. A genomewide admixture mapping panel for Hispanic/Latino populations. Am J Hum Genet. 2007;80: 1171–1178. doi:10.1086/518564

4. 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. doi:10.1073/pnas.0507611102

5. Moreno-Estrada A, Gignoux CR, Fernández-López JC, Zakharia F, Sikora M, Contreras AV, et al. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. Science. 2014;344: 1280–1285. doi:10.1126/science.1251688

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

7. Keinan A, Reich D. Human population differentiation is strongly correlated with local recombination rate. PLoS Genet. 2010;6: e1000886. doi:10.1371/journal.pgen.1000886

8. Willing EM, Dreyer C, van Oosterhout C. Estimates of genetic differentiation measured by F_{ST} do not necessarily require large sample sizes when using many SNP markers. PLoS One. 2012;7: e42649. doi:10.1371/journal.pone.0042649

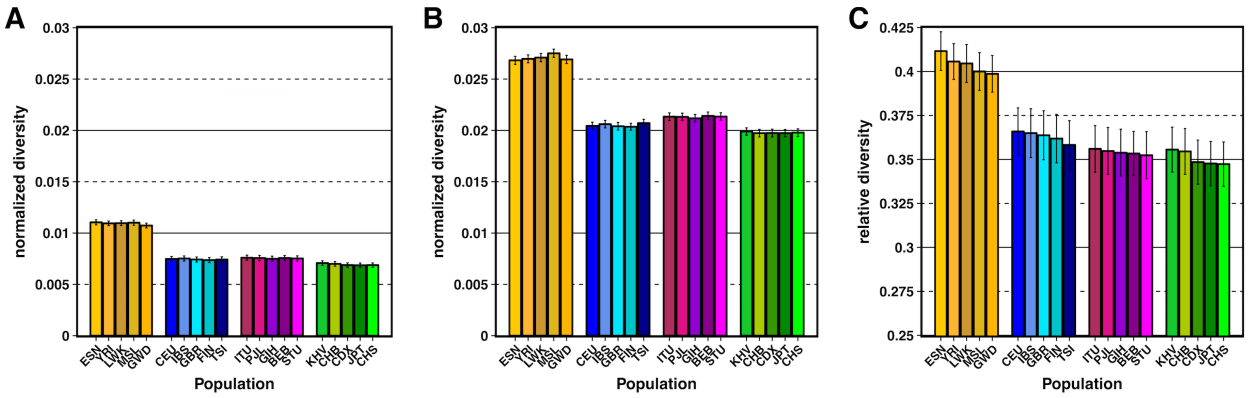


Figure A.1. Diversity for TGP non-admixed populations while controlling for GC-biased gene conversion and recombination hotspots.
(A) Normalized diversity (π/divergence) measured across the lowest 1% *B* quantile bin (strong BGS). (B) Normalized diversity measured across the highest 1% *B* quantile bin (weak BGS). (C) Relative diversity: the ratio of normalized diversity for the lowest 1% *B* bin to normalized diversity for the highest 1% *B* bin (π/π_{min}). Error bars represent ±1 SEM calculated from 1,000 bootstrapped datasets.

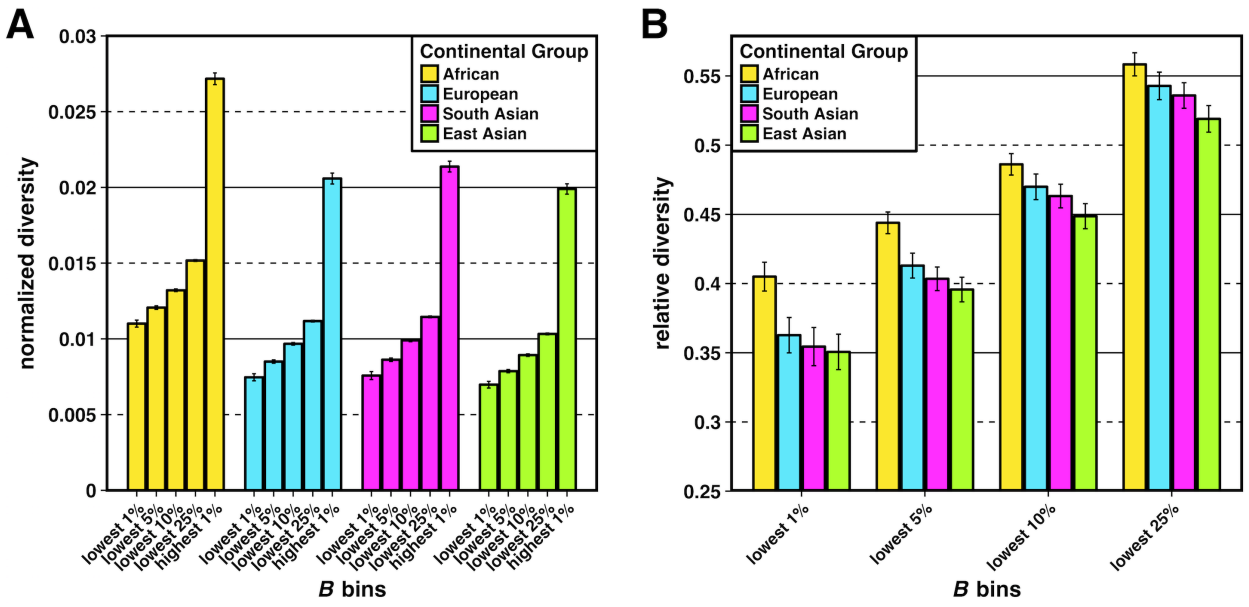


Figure A.2. Diversity for TGP continental groups while controlling for GC-biased gene conversion and recombination hotspots.
(A) Normalized diversity (π/divergence) measured across the lowest 1%, 5%, 10% and 25% *B* quantile bins (strong BGS) and the highest 1% *B* quantile bin (weak BGS). (B) Relative diversity (π/π_{min}) for the lowest 1%, 5%, 10%, and 25% *B* bins. Error bars represent ±1 SEM calculated from 1,000 bootstrapped datasets.

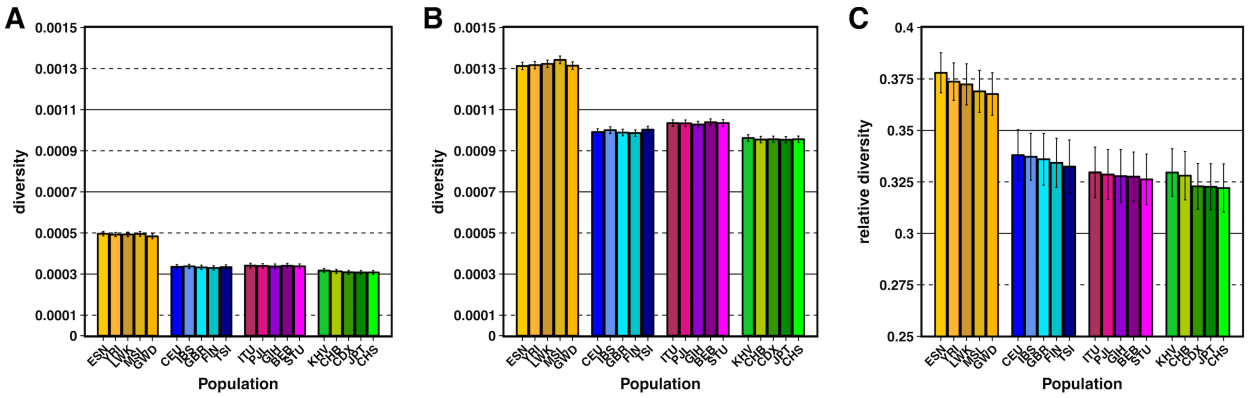


Figure A.3. Diversity for TGP non-admixed populations without normalizing by divergence with Rhesus macaque.
(A) Diversity (π) measured across the lowest 1% *B* quantile bin (strong BGS). (B) Diversity measured across the highest 1% *B* quantile bin (weak BGS). (C) Relative diversity: the ratio of diversity for the lowest 1% *B* bin to diversity for the highest 1% *B* bin (π/π_{min}). Error bars represent ±1 SEM calculated from 1,000 bootstrapped datasets.

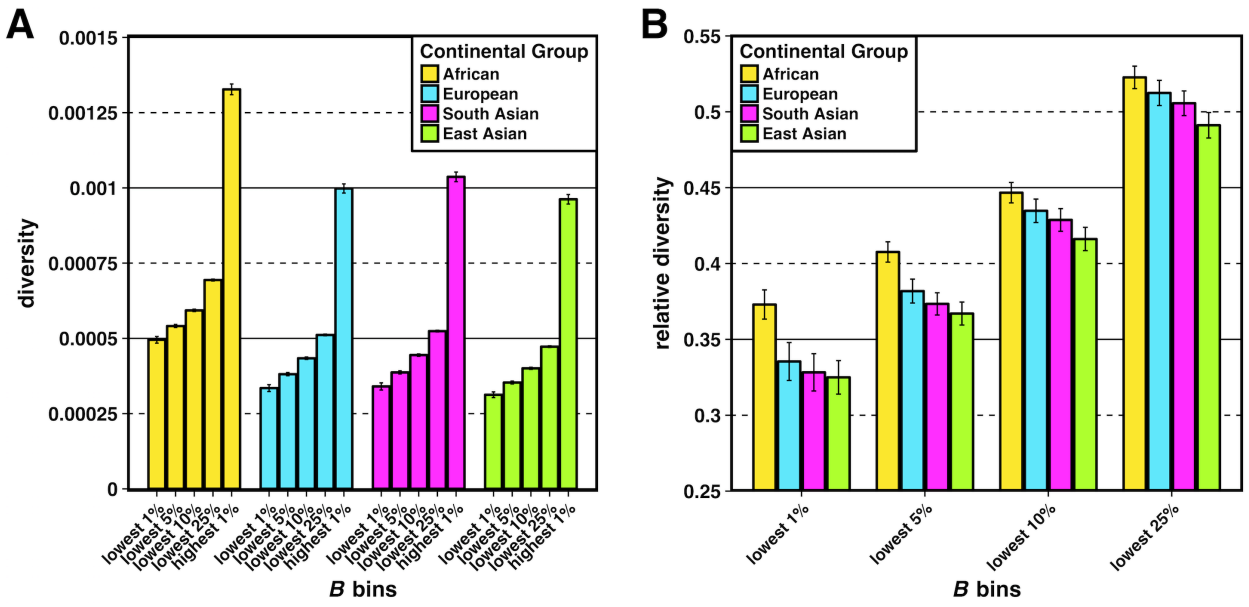


Figure A.4. Diversity for TGP continental groups without normalizing by divergence with Rhesus macaque.
(A) Diversity (π) measured across the lowest 1%, 5%, 10% and 25% *B* quantile bins (strong BGS) and the highest 1% *B* quantile bin (weak BGS). (B) Relative diversity (π/π_{min}) for the lowest 1%, 5%, 10%, and 25% *B* bins. Error bars represent ±1 SEM calculated from 1,000 bootstrapped datasets.

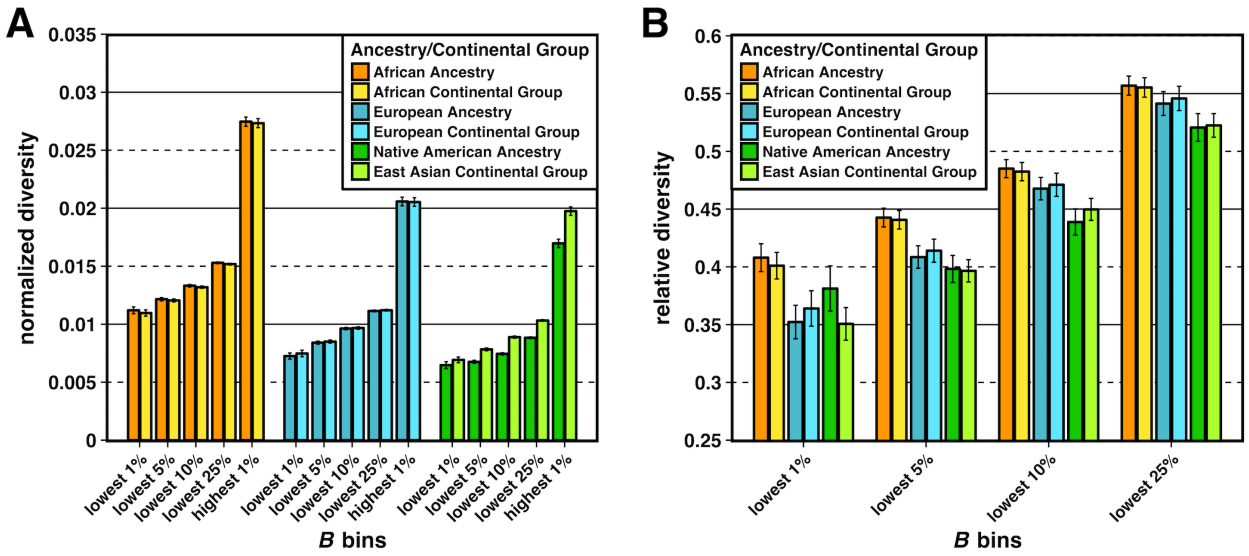


Figure A.5. Comparing patterns of diversity between local ancestry segments of admixed samples and continental groups.

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

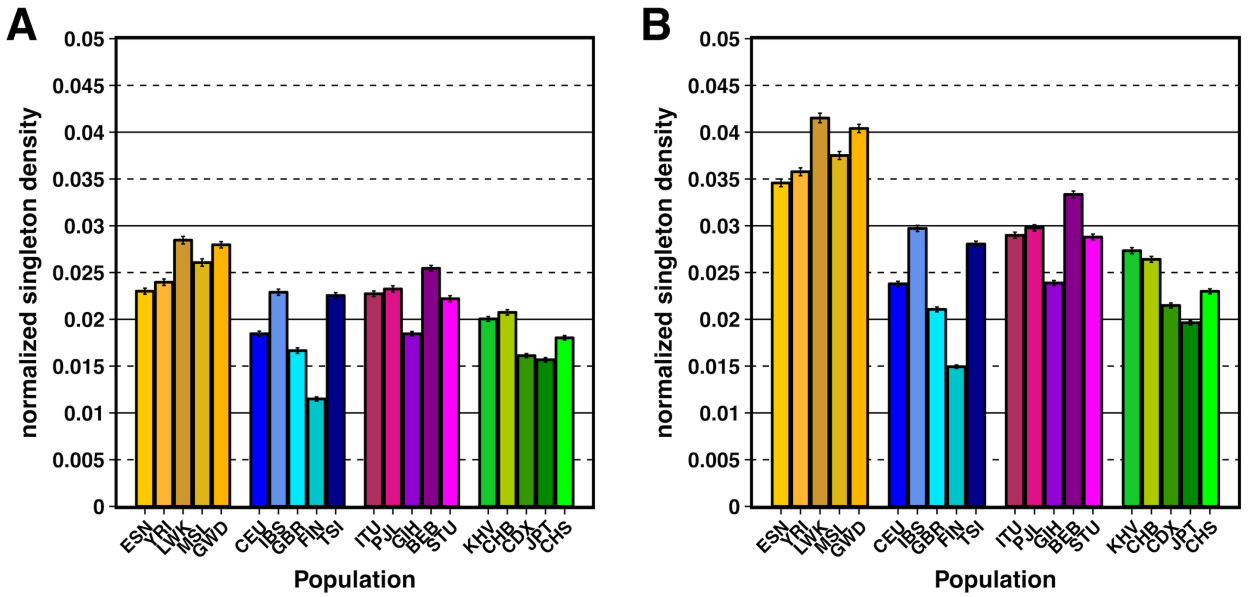


Figure A.6. Singleton density for the lowest and highest 1% *B* quantile bins for non-admixed populations of the Thousand Genomes Project (TGP).
(A) Normalized singleton density (ψ/divergence) measured across the lowest 1% *B* quantile bin (strong BGS). (B) Normalized singleton density measured across the highest 1% *B* quantile bin (weak BGS). TGP population labels are indicated below each bar (see Table A.1 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.

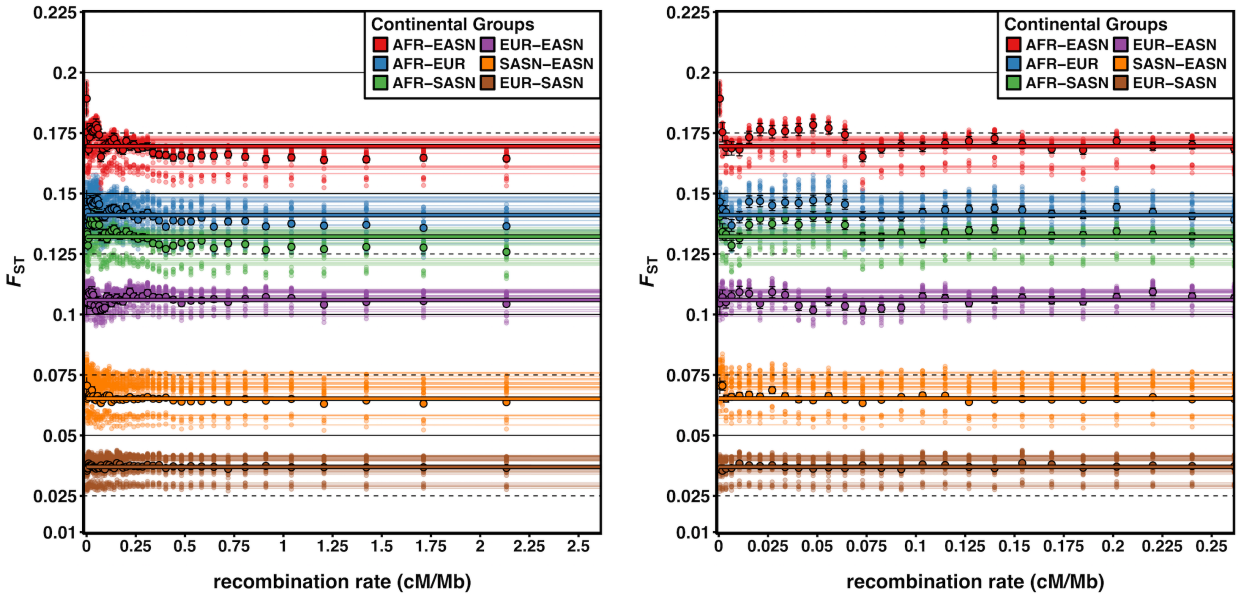


Figure A.7. F_{ST} is not correlated with recombination rate.
F_{ST} between TGP populations measured across 2% recombination rate quantile bins. The right panel displays a narrower range of recombination rates to show detail. Smaller transparent points and lines show the estimates and corresponding lines of best fit (using linear regression) for F_{ST} between every pairwise population comparison 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.

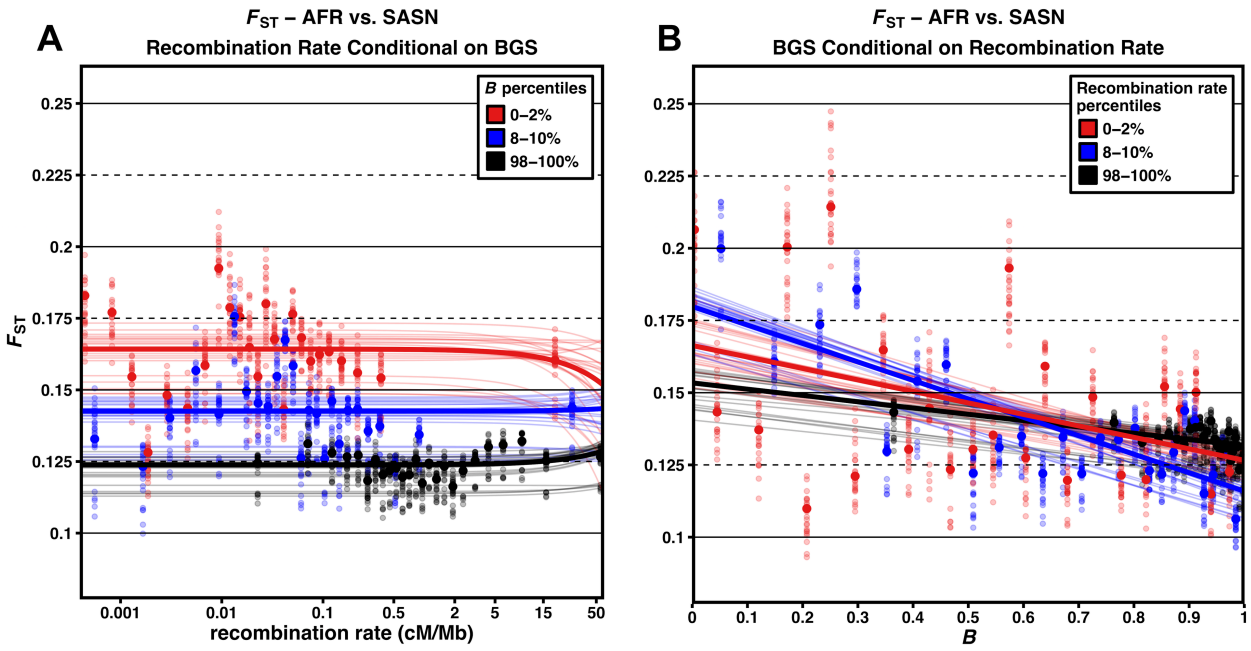


Figure A.8. F_{ST} between African (AFR) and South Asian (SASN) populations jointly across *B* and recombination rate.
(A) F_{ST} as a function of 25 recombination rate bins (4% quantile bins) conditional on three different 2% *B* quantile bins (note log scale of x-axis for recombination rate). (B) F_{ST} as a function of 25 *B* bins (4% quantile bins) conditional on three different 2% recombination rate quantile bins. Smaller transparent points and lines show the F_{ST} estimates and corresponding lines of best fit (using linear regression) for each of the pairwise comparisons of AFR vs. SASN Thousand Genomes Project (TGP) populations (25 comparisons total). Larger opaque points are mean F_{ST} estimates across all pairwise comparisons of AFR vs. SASN TGP populations (with bold lines showing their corresponding lines of best fit).

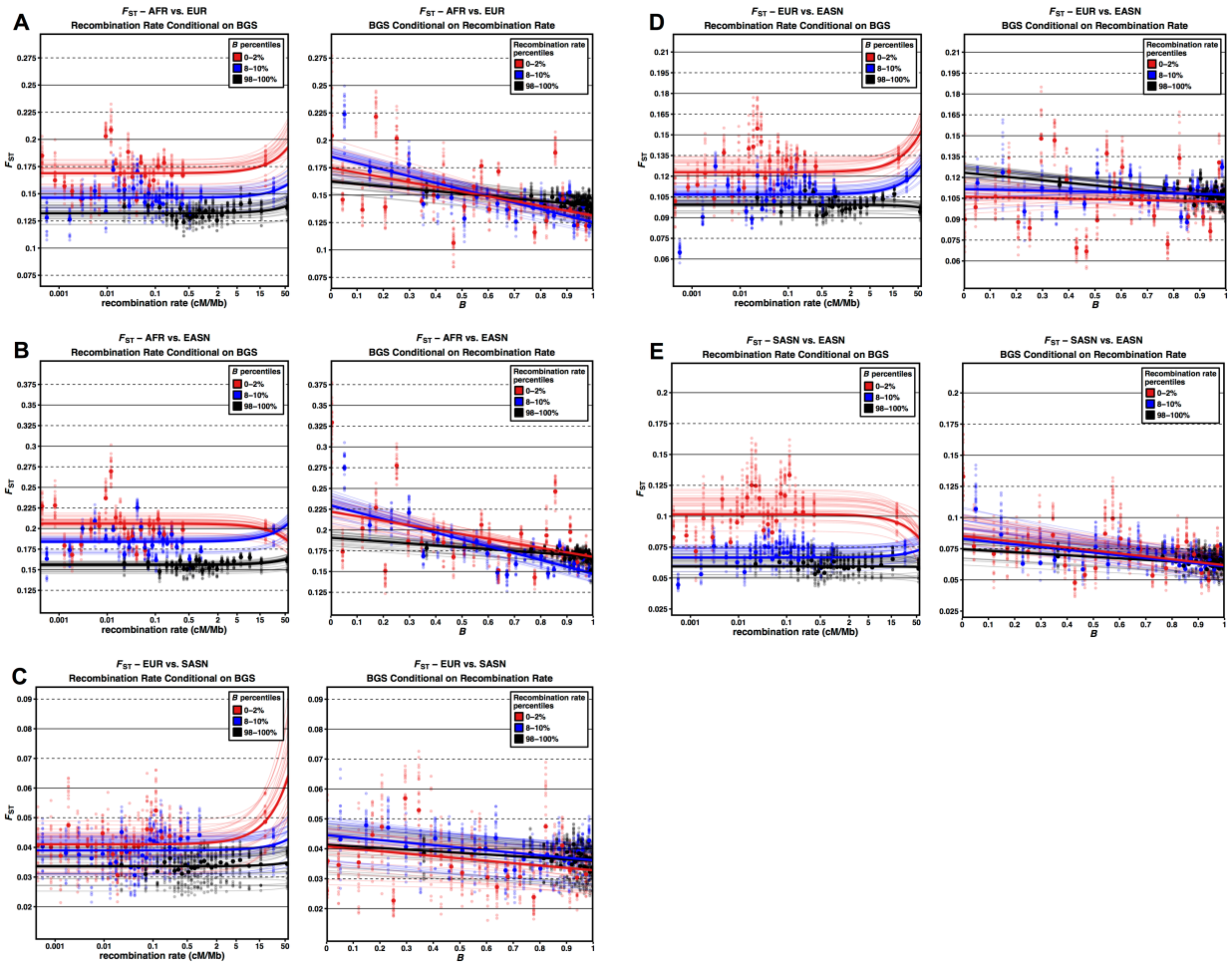


Figure A.9. F_{ST} measured across joint bins of *B* and recombination rate for different TGP continental groups.

The left panels of Figure A.9 A-E show F_{ST} measured as a function of 25 4% recombination rate quantile bins conditional on three 2% *B* quantile bins (note log scale of x-axis for recombination rate). The right panels of Figure A.9 A-E show F_{ST} measured as a function of 25 4% *B* quantile bins conditional on three 2% recombination rate quantile bins. The following continental group comparisons are shown for each plot: (A) African vs. European, (B) African vs. East Asian, (C) European vs. South Asian, (D) European vs. East Asian, (E) South Asian vs. East Asian. Smaller transparent points and lines show the F_{ST} estimates and corresponding lines of best fit (using linear regression) for each of the pairwise population comparisons within a particular pair of continental groups (25 comparisons total). Larger opaque points are mean F_{ST} estimates across all pairwise comparisons within a particular pair of continental groups (bold lines showing their corresponding lines of best fit).

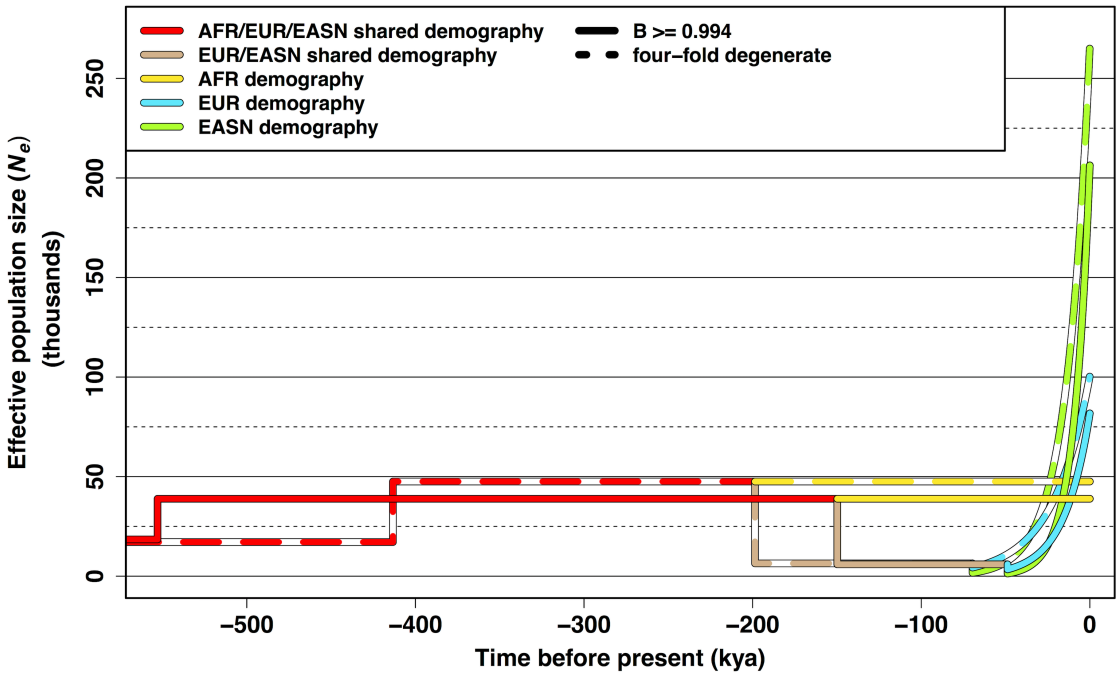


Figure A.10. Inference models inferred from TGP Complete Genomics (CG) high *B* neutral regions and coding four-fold degenerate sites.
Solid lines are the inference results from running dadi on 53 YRI (African), 64 CEU (European), and 62 CHS (East Asian) TGP CG samples (projected down to 106 chromosomes during inference procedure) across neutral regions in the highest 1% *B* bin ($B \geq 0.994$). Broken lines represent the inference results using the same CG samples but with sequence data only from coding four-fold degenerate synonymous sites. See Table A.8 for parameter values.

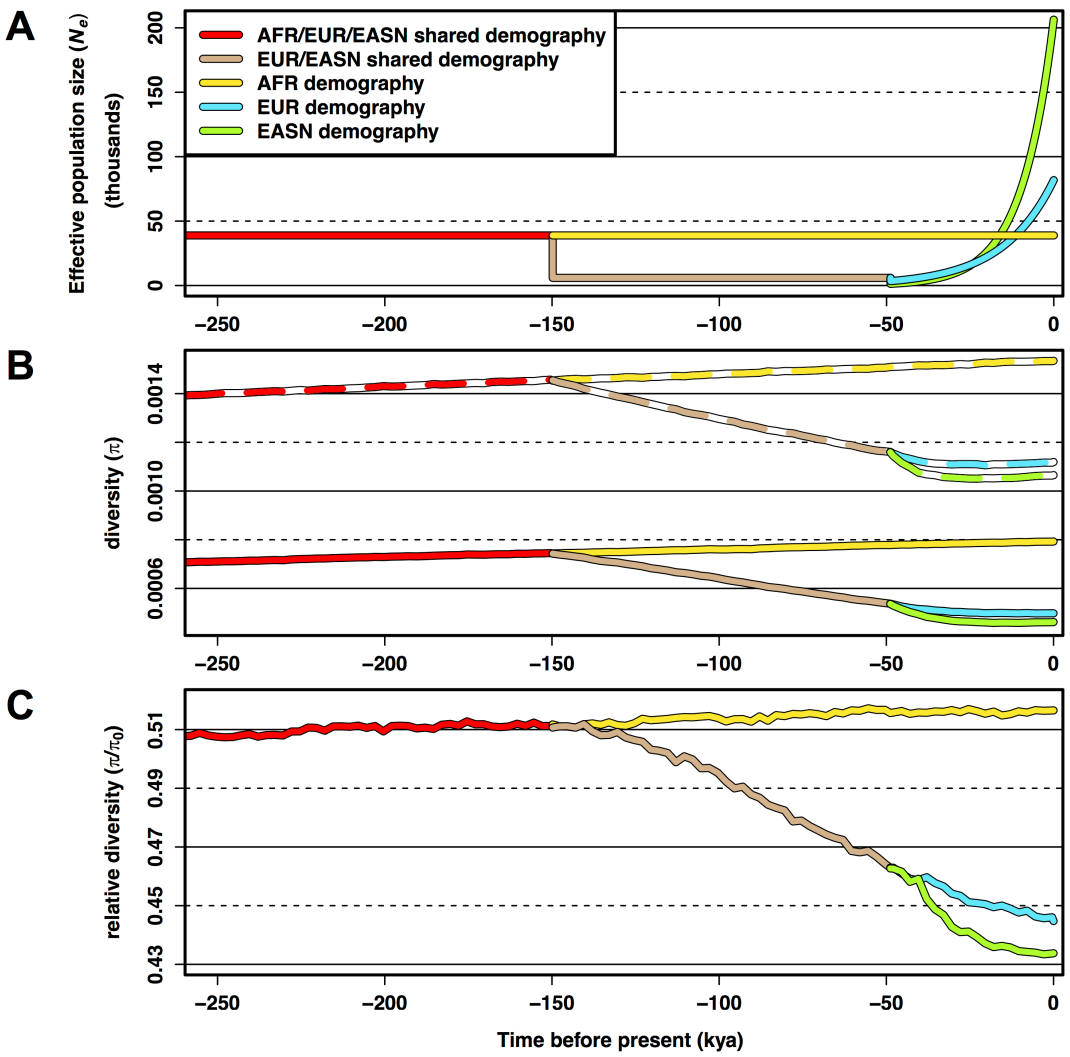


Figure A.11. Simulations of diversity and relative diversity under BGS using a human demographic model without migration.
(A) Inferred demographic model from Complete Genomics TGP data. The demographic model used for the simulations in this figure are identical to those used for Figure 2.5, except that migration parameters between all populations are set to 0. (B) Simulated diversity at neutral sites across populations as a function of time under our inferred demographic model without BGS (π_0 - dashed colored lines) and with BGS (π - solid colored lines). (C) Relative diversity (π/π_0) measured by taking the ratio of diversity with BGS (π) to diversity without BGS (π_0) at each time point. Note that the x-axes in all three figures are on the same scale. Time is scaled using a human generation time of 25 years per generation. Simulation data was sampled every 100 generations.

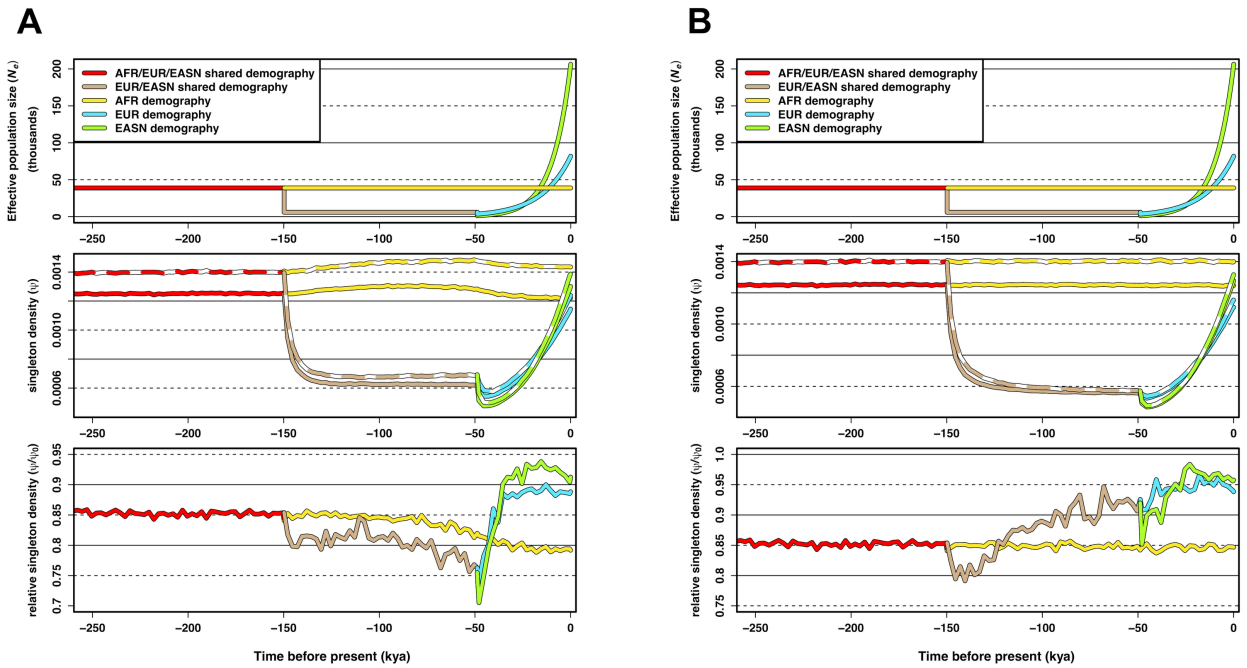
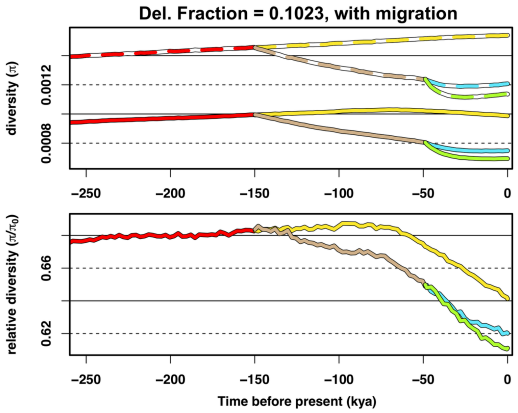
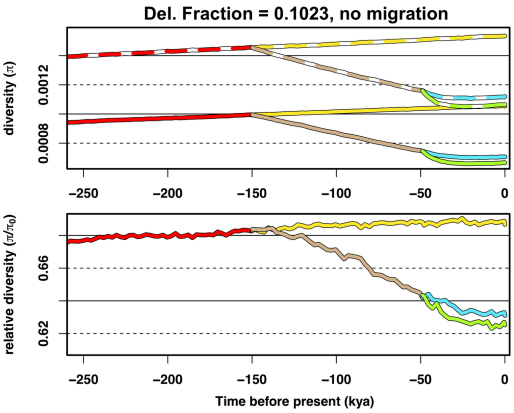
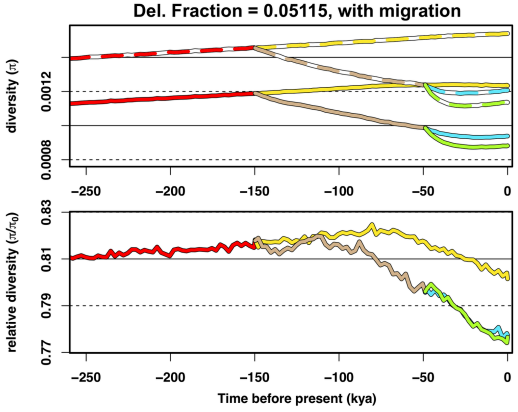
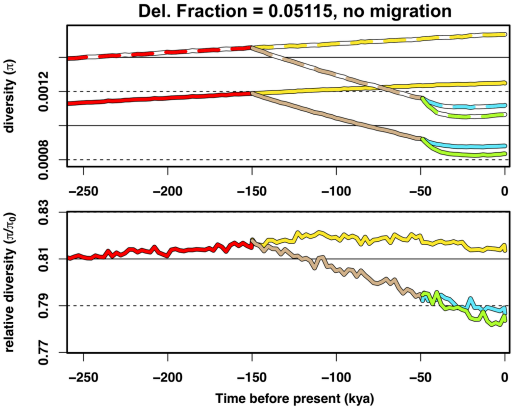


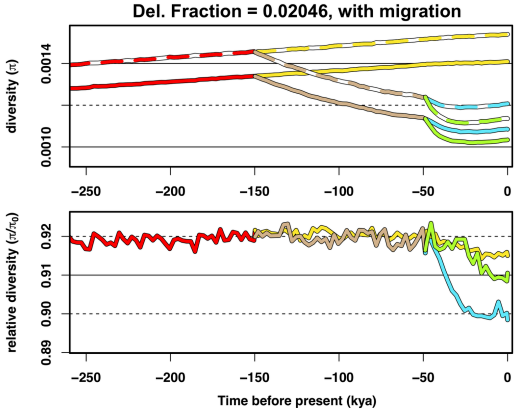
Figure A.12. Simulations of singleton density and relative singleton density.
(A) Results of simulations under a demographic model with migration between all human populations. (B) Results of simulations under a demographic model with no migration. The second row of (A) and (B) shows measurements of singleton density (i.e., number of singletons observed per site) from simulations without BGS (ψ_0 - dashed colored lines) and with BGS (ψ - solid colored lines). The bottom row of (A) and (B) shows corresponding relative singleton density (ψ/ψ_0) measured by taking the ratio of singleton density with BGS (ψ) to singleton density without BGS (ψ_0) at each sampled generation time point. The simulation data used for these measurements is identical to that of Figure 2.5 (for simulations with migration) and Figure A.11 (for simulations without migration).

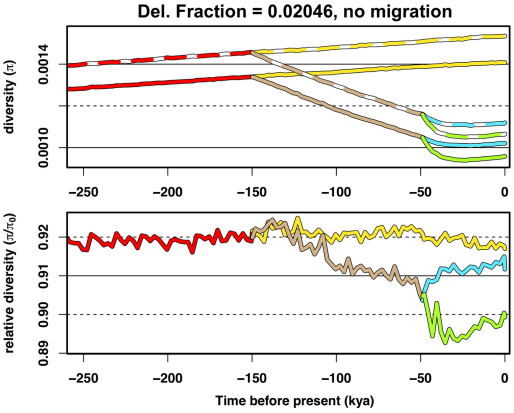

Del. Fraction = 0.1023, with migration
diversity (π)
0.0008
0.0012
−250
−200
−150
−100
−50
0
relative diversity (π/π0)
0.62
0.66
Time before present (kya)

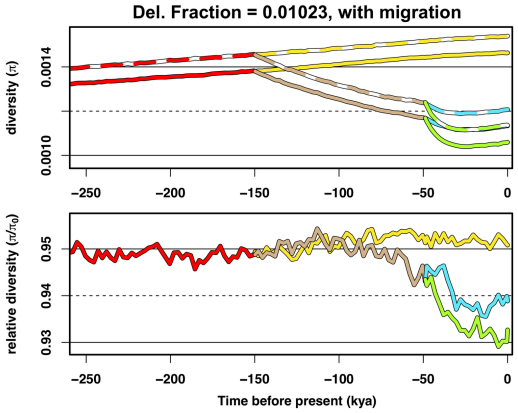

Del. Fraction = 0.1023, no migration
diversity (π)
0.0008
0.0012
−250
−200
−150
−100
−50
0
relative diversity (π/π0)
0.62
0.66
Time before present (kya)


Del. Fraction = 0.05115, with migration
diversity (π)
0.0008
0.0012
−250
−200
−150
−100
−50
0
relative diversity (π/π0)
0.77
0.79
0.81
0.83
Time before present (kya)


Del. Fraction = 0.05115, no migration
diversity (π)
0.0008
0.0012
−250
−200
−150
−100
−50
0
relative diversity (π/π0)
0.77
0.79
0.81
0.83
Time before present (kya)


Del. Fraction = 0.02046, with migration
diversity (π)
0.0010
0.0014
−250
−200
−150
−100
−50
0
relative diversity (π/π0)
0.89
0.90
0.91
0.92
Time before present (kya)


Del. Fraction = 0.02046, no migration
diversity (π)
0.0010
0.0014
−250
−200
−150
−100
−50
0
relative diversity (π/π0)
0.89
0.90
0.91
0.92
Time before present (kya)


Del. Fraction = 0.01023, with migration
diversity (π)
0.0010
0.0014
−250
−200
−150
−100
−50
0
relative diversity (π/π0)
0.93
0.94
0.95
Time before present (kya)

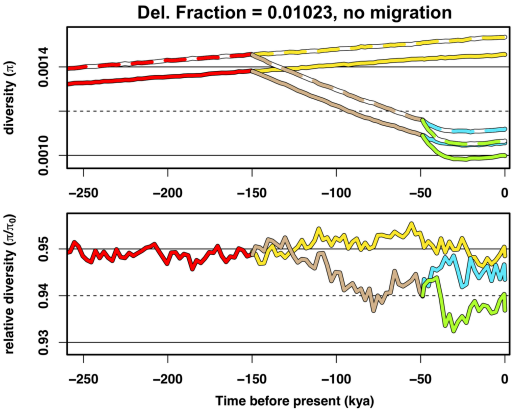

Del. Fraction = 0.01023, no migration
diversity (π)
0.0010
0.0014
−250
−200
−150
−100
−50
0
relative diversity (π/π0)
0.93
0.94
0.95
Time before present (kya)

Figure A.13. Simulations of diversity and relative diversity under BGS using various fractions of sites experiencing deleterious mutation.
Values for the deleterious site fraction are provided in the title for each set of plots. Left column plots show results of simulations under a demographic model with migration between all human populations. Right column plots show results of simulations under a demographic model with no migration. Colored lines represent different populations though time and are identical to those in Figure 2.5 and Figure A.11. The demographic model used is also identical to that in Figure 2.5 (for simulations with migration) and Figure A.11 (for simulations without migration). Simulation data was sampled every 100 generations.

Table A.1. Phase 3 TGP population information and classification by continental group (or admixed).

Population	Ethnic Group/Population	TGP Population Label	Continental Group (or Admixed)	Sample Size
Esan in Nigeria	Esan	ESN	AFR	99
Gambian in Western Division, Mandinka	Gambian	GWD	AFR	113
Luhya in Webuye, Kenya	Luhya	LWK	AFR	99
Mende in Sierra Leone	Mende	MSL	AFR	85
Yoruba in Ibadan, Nigeria	Yoruba	YRI	AFR	108
Utah residents (CEPH) with Northern and Western European ancestry	CEPH	CEU	EUR	99
British in England and Scotland	British	GBR	EUR	91
Finnish in Finland	Finnish	FIN	EUR	99
Iberian Populations in Spain	Spanish	IBS	EUR	107
Toscani in Italia	Tuscan	TSI	EUR	107
Bengali in Bangladesh	Bengali	BEB	SASN	86
Gujarati Indians in Houston, TX, USA	Gujarati	GIH	SASN	103
Indian Telugu in the UK	Telugu	ITU	SASN	102
Punjabi in Lahore, Pakistan	Punjabi	PJL	SASN	96
Sri Lankan Tamil in the UK	Tamil	STU	SASN	102
Chinese Dai in Xishuangbanna, China	Dai Chinese	CDX	EASN	93
Han Chinese in Beijing, China	Han Chinese	CHB	EASN	103
Southern Han Chinese	Southern Han Chinese	CHS	EASN	105
Japanese in Tokyo, Japan	Japanese	JPT	EASN	104
Kinh in Ho Chi Minh City, Vietnam	Kinh Vietnamese	KHV	EASN	99
African Caribbean in Barbados	Barbadian	ACB	Admixed	96
People with African Ancestry in Southwest USA	African-American SW	ASW	Admixed	61
Colombians in Medellin, Colombia	Colombian	CLM	Admixed	94
People with Mexican Ancestry in Los Angeles, CA, USA	Mexican-American	MXL	Admixed	64
Peruvians in Lima, Peru	Peruvian	PEL	Admixed	85
Puerto Ricans in Puerto Rico	Puerto Rican	PUR	Admixed	104

	AFR vs. EASN	AFR vs. EUR	AFR vs. SASN	EUR vs. SASN	EUR vs. EASN	SASN vs. EASN	Global
β_0 ± SEM (p-value)	0.2043 ± 0.0036 (< 1e-04)	0.1724 ± 0.0031 (< 1e-04)	0.1591 ± 0.0028 (< 1e-04)	0.0459 ± 0.0011 (< 1e-04)	0.1214 ± 0.0029 (< 1e-04)	0.0880 ± 0.0021 (< 1e-04)	0.1337 ± 0.0020 (< 1e-04)
β_1 ± SEM (p-value)	-0.0428 ± 0.0042 (< 1e-04)	-0.0363 ± 0.0036 (< 1e-04)	-0.0344 ± 0.0033 (< 1e-04)	-0.0099 ± 0.0013 (< 1e-04)	-0.0168 ± 0.0034 (< 1e-04)	-0.0223 ± 0.0024 (< 1e-04)	-0.0295 ± 0.0023 (< 1e-04)

Table A.2. Regression coefficient estimates for robust linear regression of F_{ST} on B.
To apply an additional test for the relationship between background selection and F_{ST} that is more robust to outlier points or points with high influence, we performed robust linear regression using M-estimation with Huber weighting. Robust linear regression was run on the same data as was used for the linear regression described for Table 2.1 in Chapter 2. Each column gives the regression coefficients for the linear model $F_{ST} = \beta_0 + \beta_1 B$, 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. The final column, "Global", gives the regression coefficients for the linear model applied to all pairwise population comparisons (150 total). Standard errors of the mean (SEM) for β_0 and β_1 were calculated from 1,000 bootstrap iterations (see Materials and Methods in Chapter 2). P-values are derived from a Wald (F-distribution) test on the F-statistic for the corresponding regression coefficient.

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

Table A.3. Regression coefficient estimates for linear regression of F_{ST} on 2% quantile bins of recombination rate.
The first two rows give the regression coefficients for the linear model $F_{ST} = \beta_0 + \beta_1 \rho + \varepsilon$, where ρ represents the mean recombination rate for the bin being tested and F_{ST} is the estimated F_{ST} for all population comparisons within a particular pair of continental groups (given in the column header). The final column, "Global", gives the regression coefficients for the linear model applied to all pairwise population comparisons (150 total). When performing the regression, ρ was first scaled to between 0 and 1, such that 1 represents the maximum observed recombination rate (126.88 cM/Mb) and 0 represents the minimum observed recombination rate (0.0 cM/Mb). The correlation coefficient, r, between ρ and F_{ST} for each comparison is shown in the bottom row. Standard errors of the mean (SEM) for β_0, β_1, and r were calculated from 1,000 bootstrap iterations (see Materials and Methods in Chapter 2). P-values are derived from a two-sided t-test of the t-value for the corresponding regression coefficient.

	AFR vs. EASN	AFR vs. EUR	AFR vs. SASN	EUR vs. SASN	EUR vs. EASN	SASN vs. EASN	Global
β_0 ± SEM (p-value)	0.1688 ± 0.0006 (< 1e-04)	0.1425 ± 0.0006 (< 1e-04)	0.1308 ± 0.0005 (< 1e-04)	0.0376 ± 0.0002 (< 1e-04)	0.1073 ± 0.0006 (< 1e-04)	0.0699 ± 0.0004 (< 1e-04)	0.1093 ± 0.0003 (< 1e-04)
β_1 ± SEM (p-value)	-0.0001 ± 0.0026 (0.9755)	0.0007 ± 0.0022 (0.7591)	0.0008 ± 0.0021 (0.7281)	-0.0015 ± 0.0008 (0.4317)	0.0004 ± 0.0020 (0.7869)	-0.0046 ± 0.0014 (0.0225)	-0.0009 ± 0.0013 (0.9022)

Table A.4. Regression coefficient estimates for robust linear regression of F_{ST} on recombination rate.

To apply an additional test for the relationship between recombination rate and F_{ST} that is more robust to outlier points or points with high influence, we performed robust linear regression using M-estimation with Huber weighting. Robust linear regression was run on the same data as was used for the linear regression described for Table A.3. Each column gives the regression coefficients for the linear model $F_{ST} = \beta_0 + \beta_1\rho$, where ρ represents the mean recombination rate for the bin being tested and F_{ST} is the estimated F_{ST} for all population comparisons within a particular pair of continental groups. The final column, "Global", gives the regression coefficients for the linear model applied to all pairwise population comparisons (150 total). When performing the regression, ρ was first scaled to between 0 and 1, such that 1 represents the maximum observed recombination rate (126.88 cM/Mb) and 0 represents the minimum observed recombination rate (0.0 cM/Mb). Standard errors of the mean (SEM) for β_0 and β_1 were calculated from 1,000 bootstrap iterations (see Materials and Methods in Chapter 2). P-values are derived from a Wald (F-distribution) test on the F-statistic for the corresponding regression coefficient.

	AFR vs. EASN		AFR vs. EUR		AFR vs. SASN		EUR vs. SASN		EUR vs. EASN		SASN vs. EASN		Global	
	ρ by B	B by ρ	ρ by B	B by ρ	ρ by B	B by ρ	ρ by B	B by ρ	ρ by B	B by ρ	ρ by B	B by ρ	ρ by B	B by ρ
β_0 (p-value)	0.2037 (< 1e-04)	0.2054 (< 1e-04)	0.1713 (< 1e-04)	0.1725 (< 1e-04)	0.1594 (< 1e-04)	0.1609 (< 1e-04)	0.0450 (< 1e-04)	0.0451 (< 1e-04)	0.1200 (< 1e-04)	0.1209 (< 1e-04)	0.0888 (< 1e-04)	0.0903 (< 1e-04)	0.1314 (< 1e-04)	0.1325 (< 1e-04)
β_0 (p-value) - robust	0.1995 (< 1e-04)	0.2019 (< 1e-04)	0.1696 (< 1e-04)	0.1709 (< 1e-04)	0.1567 (< 1e-04)	0.1589 (< 1e-04)	0.0450 (< 1e-04)	0.0449 (< 1e-04)	0.1185 (< 1e-04)	0.1199 (< 1e-04)	0.0853 (< 1e-04)	0.0868 (< 1e-04)	0.1314 (< 1e-04)	0.1329 (< 1e-04)
β_1 (p-value)	-0.0427 (< 1e-04)	-0.0448 (< 1e-04)	-0.0355 (< 1e-04)	-0.0368 (< 1e-04)	-0.0353 (< 1e-04)	-0.0371 (< 1e-04)	-0.0093 (< 1e-04)	-0.0094 (< 1e-04)	-0.0157 (< 1e-04)	-0.0168 (< 1e-04)	-0.0245 (< 1e-04)	-0.0262 (< 1e-04)	-0.0272 (< 1e-04)	-0.0285 (< 1e-04)
β_1 (p-value) - robust	-0.0380 (< 1e-04)	-0.0408 (< 1e-04)	-0.0335 (< 1e-04)	-0.0350 (< 1e-04)	-0.0323 (< 1e-04)	-0.0350 (< 1e-04)	-0.0092 (< 1e-04)	-0.0091 (< 1e-04)	-0.0140 (< 1e-04)	-0.0157 (< 1e-04)	-0.0200 (< 1e-04)	-0.0218 (< 1e-04)	-0.0271 (< 1e-04)	-0.0288 (< 1e-04)
β_2 (p-value)	-0.0448 (< 1e-04)	-0.0166 (0.0018)	-0.0300 (< 1e-04)	-0.0285 (< 1e-04)	-0.0332 (< 1e-04)	-0.0324 (< 1e-04)	-0.0039 (0.2212)	0.0095 (0.0002)	0.0146 (0.0046)	0.0142 (0.0008)	-0.0252 (< 1e-04)	-0.0081 (0.0481)	-0.0204 (0.0389)	-0.0103 (0.1963)
β_2 (p-value) - robust	-0.0306 (< 1e-04)	-0.0039 (0.4085)	-0.0225 (< 1e-04)	-0.0230 (< 1e-04)	-0.0226 (< 1e-04)	-0.0243 (< 1e-04)	-0.0035 (0.2843)	0.0104 (< 1e-04)	0.0178 (< 1e-04)	0.0180 (< 1e-04)	-0.0125 (0.0066)	0.0049 (0.2047)	-0.0176 (0.0932)	-0.0096 (0.2576)
β_3 (p-value)	0.0588 (< 1e-04)	0.0345 (< 1e-04)	0.0419 (< 1e-04)	0.0419 (< 1e-04)	0.0458 (< 1e-04)	0.0488 (< 1e-04)	0.0052 (0.1387)	-0.0094 (0.0018)	-0.0107 (0.0585)	-0.0084 (0.0886)	0.0286 (< 1e-04)	0.0127 (0.0083)	0.0283 (0.0092)	0.0200 (0.0322)
β_3 (p-value) - robust	0.0436 (< 1e-04)	0.0201 (0.0002)	0.0338 (< 1e-04)	0.0358 (< 1e-04)	0.0343 (< 1e-04)	0.0400 (< 1e-04)	0.0047 (0.1889)	-0.0103 (0.0009)	-0.0146 (0.0032)	-0.0127 (0.0047)	0.0141 (0.0053)	-0.0023 (0.6046)	0.0254 (0.0269)	0.0196 (0.0484)

Table A.5. Multiple Linear Regression and Robust Regression of F_{ST} on B and recombination rate.

We performed multiple linear regression of B and recombination rate on F_{ST} estimated for 1,250 bins of either 1) 50 2% quantile recombination rate bins by 25 4% quantile B bins ("ρ by B" in the table) or 2) 50 2% quantile B bins by 25 4% quantile recombination rate bins ("B by ρ" in the table). Each column gives the regression coefficients for the linear model $F_{ST} = \beta_0 + \beta_1 B + \beta_2 \rho + \beta_3 B\rho$, where B represents the mean background selection coefficient for the bin being tested, ρ represents the mean recombination rate for the bin being tested, and $B\rho$ is an interaction term for the background selection coefficient and recombination rate. F_{ST} is the estimated F_{ST} for all population comparisons within a particular pair of continental groups except for the final pair of columns, for which the linear model was applied to all pairwise population estimates of F_{ST} (150 total). When performing the regression, ρ was first scaled to between 0 and 1, such that 1 represents the maximum observed recombination rate (126.88 cM/Mb) and 0 represents the minimum observed recombination rate (0.0 cM/Mb). We also performed robust linear regression using M-estimation with Huber weighting on the same linear model, $F_{ST} = \beta_0 + \beta_1 B + \beta_2 \rho + \beta_3 B\rho$. The regression coefficients from performing robust linear regression are indicated by the rows labeled "robust" in the table. P-values for normal linear regression are derived from a two-sided t-test of the t-value for the corresponding regression coefficients. P-values for robust linear regression are derived from a Wald (F-distribution) test on the F-statistic for the corresponding regression coefficients.

	AFR vs. EASN			AFR vs. EUR			AFR vs. SASN			EUR vs. SASN			EUR vs. EASN			SASN vs. EASN		
quantile *B* bin	0-2% *B*	8-10% *B*	98-100% *B*	0-2% *B*	8-10% *B*	98-100% *B*	0-2% *B*	8-10% *B*	98-100% *B*	0-2% *B*	8-10% *B*	98-100% *B*	0-2% *B*	8-10% *B*	98-100% *B*	0-2% *B*	8-10% *B*	98-100% *B*
β_0 (p-value*)	0.2058	0.1838	0.1561	0.1689	0.1463	0.132	0.1643	0.1425	0.1238	0.041	0.039	0.0336	0.1229	0.1067	0.0993	0.1017	0.0666	0.0594
β_1 (p-value)	-0.0464 (0.8263)	0.0489 (0.4957)	0.0157 (0.1663)	0.0529 (0.6589)	0.0279 (0.5937)	0.016 (0.167)	-0.0273 (0.7961)	0.002 (0.9735)	0.0106 (0.3121)	0.0506 (0.1034)	0.0083 (0.6497)	0.0028 (0.4086)	0.0658 (0.5656)	0.0461 (0.4479)	-0.0046 (0.7096)	-0.0434 (0.6934)	0.0136 (0.7103)	-7e-04 (0.9216)
r	-0.0462	0.1429	0.2856	0.0928	0.1121	0.2852	-0.0544	0.007	0.2107	0.3334	0.0955	0.1729	0.1207	0.159	-0.0784	-0.083	0.0782	-0.0207

Table A.6. Linear Regression of F_{ST} on recombination rate but conditioning on *B* quantiles.

This table gives the results of running simple linear regression of F_{ST} on 25 4% quantile bins of recombination rate (ρ), while conditioning on 3 specific quantile bins of *B*. The specific quantile bins of *B* that were conditioned on were 0-2%, 8-10%, and 98-100%. Each column gives the regression coefficients for the linear model $F_{ST} = \beta_0 + \beta_1\rho + \varepsilon$ for each population comparison of F_{ST}. The correlation coefficient, *r*, between ρ and F_{ST} for each population comparison is shown in the bottom row. P-values are derived from a two-sided t-test of the t-value for the corresponding regression coefficient. The values in this table correspond directly to Figure A.8 and Figure A.9.

*all p-values for β_0 are < 1e-4

	AFR vs. EASN			AFR vs. EUR			AFR vs. SASN			EUR vs. SASN			EUR vs. EASN			SASN vs. EASN		
quantile ρ bin	0-2% ρ	8-10% ρ	98-100% ρ	0-2% ρ	8-10% ρ	98-100% ρ	0-2% ρ	8-10% ρ	98-100% ρ	0-2% ρ	8-10% ρ	98-100% ρ	0-2% ρ	8-10% ρ	98-100% ρ	0-2% ρ	8-10% ρ	98-100% ρ
β_0 (p-value*)	0.2222	0.2296	0.1907	0.1749	0.1852	0.1625	0.1663	0.1798	0.1534	0.0408	0.0446	0.0413	0.1062	0.1116	0.1234	0.0854	0.0829	0.0746
β_1 (p-value)	-0.0545 (0.052)	-0.0822 (< 1e-04)	-0.0229 (0.016)	-0.0434 (0.0258)	-0.0601 (< 1e-04)	-0.0222 (4e-04)	-0.0397 (0.0506)	-0.0639 (< 1e-04)	-0.0218 (0.0015)	-0.0079 (0.1767)	-0.0084 (0.0025)	-0.0054 (0.1041)	-0.0036 (0.8313)	-0.0039 (0.6132)	-0.0182 (0.0429)	-0.0236 (0.0551)	-0.0221 (0.0024)	-0.0122 (0.0139)
r	-0.3929	-0.7865	-0.4767	-0.4451	-0.7599	-0.6562	-0.3951	-0.78	-0.6001	-0.2791	-0.578	-0.3328	-0.0449	-0.1062	-0.408	-0.3883	-0.5788	-0.4852

Table A.7. Linear Regression of F_{ST} on *B* but conditioning on recombination rate quantiles.

This table gives the results of running simple linear regression of F_{ST} on 25 4% quantile bins of *B*, while conditioning on 3 specific quantile bins of recombination rate (ρ). The specific quantile bins of ρ that were conditioned on were 0-2%, 8-10%, and 98-100%. Each column gives the regression coefficients for the linear model $F_{ST} = \beta_0 + \beta_1 B + \varepsilon$ for each population comparison of F_{ST}. The correlation coefficient, *r*, between *B* and F_{ST} for each population comparison is shown in the bottom row. P-values are derived from a two-sided t-test of the t-value for the corresponding regression coefficient. The values in this table correspond directly to Figure A.8 and Figure A.9.

*all p-values for β_0 are < 1e-4

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

Table A.8

Inferred parameters from running dadi on TGP CG data across neutral regions in the highest 1% *B* value bin ($B \geq 0.994$) and across four-fold degenerate sites. The demographic model inferred is the Out-of-Africa demographic model of Gutenkunst et al. 2009 (Ref. [7] in Chapter 2). Time parameters, *T*, assume a generation time of 25 years per generation. Growth rates, *r*, and migration rates, *m*, are per generation. Parameters with subscript, "*Bott*", represent parameters inferred for the ancestral European and East Asian out-of-Africa bottleneck population. Time parameters with subscript "*EUR_EASN*" represent the European-East Asian population split.

Henn et al. 2016 samples		
SampleID	**Number of Sites**	**Mean Depth**
HGDP00991	2,207,845	6.96118
HGDP00987	2,229,426	7.19132
HGDP01036	2,373,023	11.6072
HGDP00992	2,452,509	12.1913
HGDP01029	2,415,792	12.3526
HGDP01032	2,407,400	12.8113
Kidd et al. 2014 samples		
SampleID	**Number of Sites**	**Mean Depth**
SA1000A	547,527	2.56481
SA1025A	2,136,905	9.1239
Kim et al. 2014 samples		
SampleID	**Number of Sites**	**Mean Depth**
KB2	2,756,225	27.5951
NB1	2,599,220	28.0148
MD8	2,777,871	38.4532
NB8	2,778,198	40.1789
KB1	2,757,336	50.5629

Table A.9.
Number of polymorphic sites and mean depth coverage of 13 KhoeSan samples used for SNP ascertainment in calculations of F_{ST} from studies of Henn et al. 2016, Kidd et al. 2014, and Kim et al. 2014 (Refs. [95-97] in Chapter 2).

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

Table A.10.
Total number of Mb in the human genome passing the set of 13 filters described in Materials and Methods in Chapter 2 that were used for calculating pairwise genetic diversity (π) for each quantile of *B*. The bottom row is the total number Mb when including the set of filters to remove regions sensitive to GC-biased gene conversion (gBGC) or sites in recombination hotspots. Additionally, these totals only include those 100 kb regions that had a minimum of 10 kb of divergence information for Rhesus macaque (see Materials and Methods in Chapter 2).

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

Table A.11.
Total number of Mb of homozygous ancestry that passed all filters and were used in the analyses of admixed samples in the 6 admixed TGP populations (ACB, ASW, CLM, MXL, PEL, PUR) for each quantile of *B*. Additionally, these totals only include those 100 kb regions that had a minimum of 10 kb of divergence information for Rhesus macaque (see Materials and Methods in Chapter 2).

APPENDIX B: Supplemental Material to Chapter 3

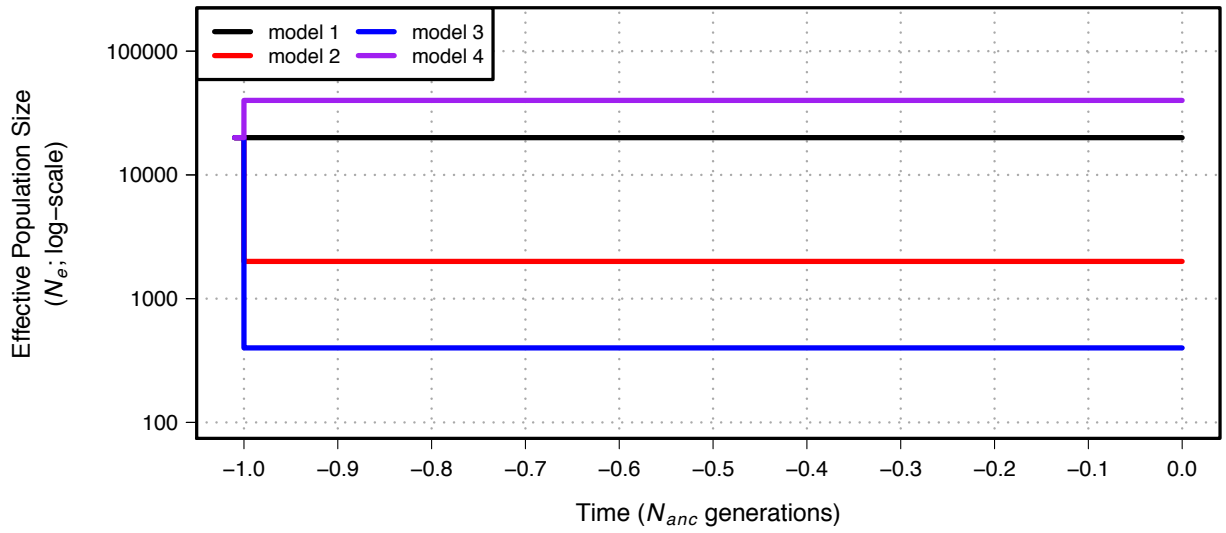


Figure B.1. Demographic models 1-4 simulated in our study.
Time proceeds forward from left to right and is scaled by the N_e of the population at the initial generation (N_{anc}; 20,000 individuals). Demographic model 2 experiences a population contraction to 2000 individuals while demographic model 3 experiences a population contraction to 400 individuals. Demographic model 4 experiences a population expansion to 40,000 individuals. All population size changes are instantaneous for models 2-4. See Table B.1 for additional model parameters.

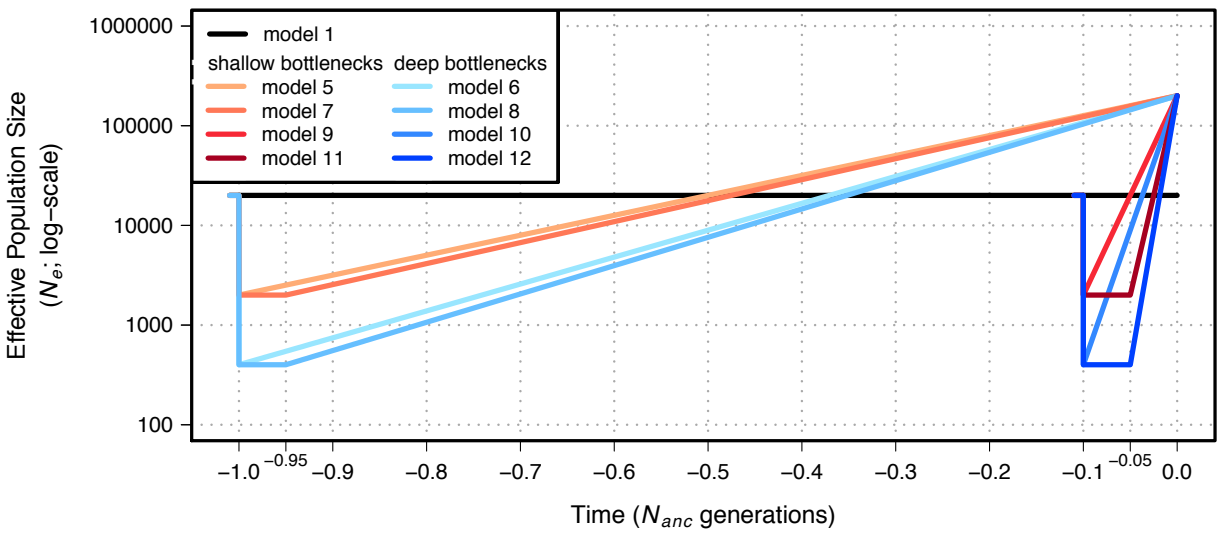


Figure B.2. Demographic models 1 and 5-12 simulated in our study.
Time proceeds forward from left to right and is scaled by the N_e of the population at the initial generation (N_{anc}; 20,000 individuals). Demographic models experiencing a shallow bottleneck (models 5, 7, 9, and 11) experience a population contraction to 2000 individuals while demographic models experiencing a deep bottleneck (models 6, 8, 10, and 12) experience a population contraction to 400 individuals. After contraction, demographic models 5-12 undergo exponential growth to a final population size of 200,000 individuals. See Table B.1 for additional model parameters.

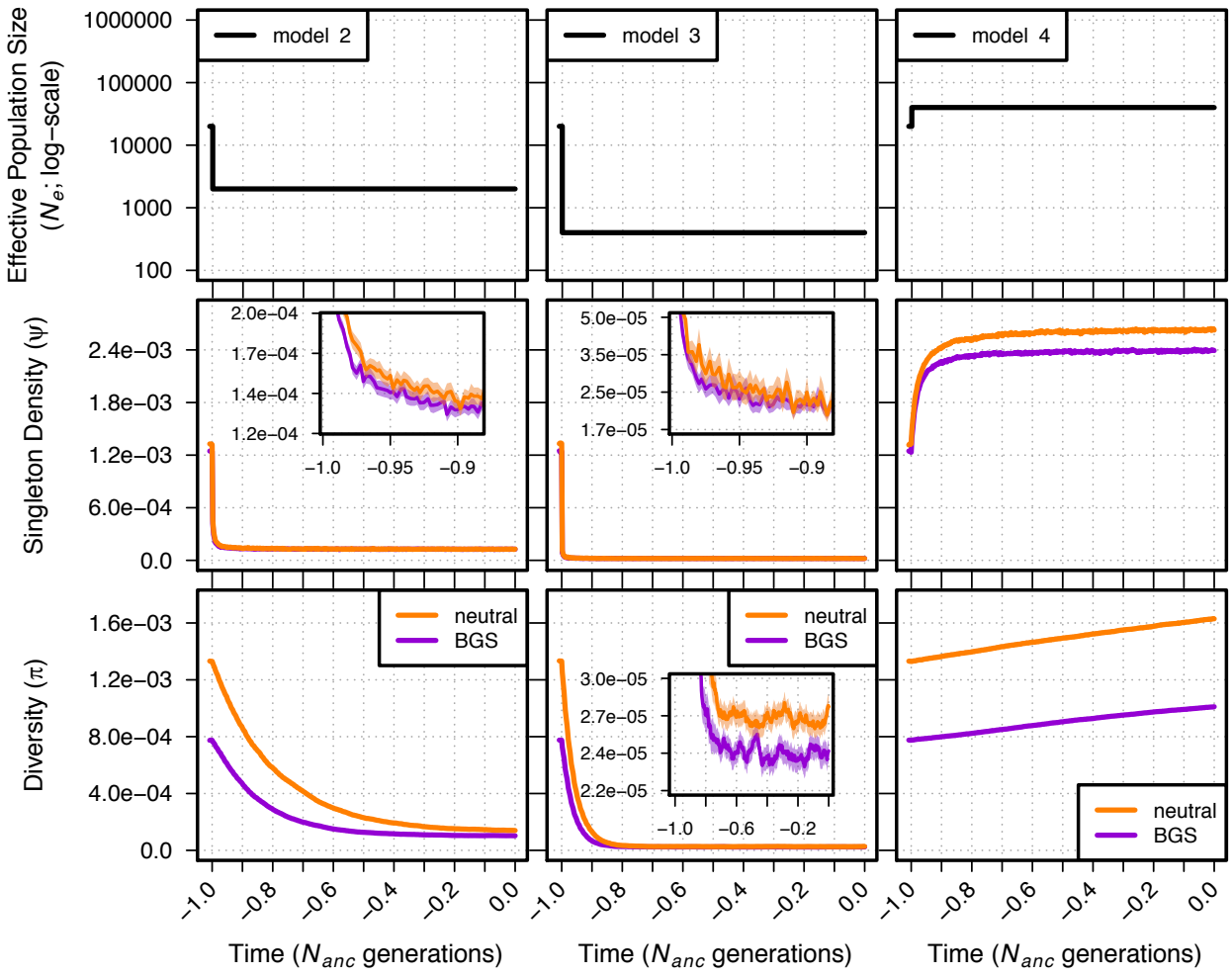


Figure B.3. Singleton density (ψ per site) and diversity (π per site) for models 2-4.
Top panel shows each demographic model (time proceeds forward from left to right; time is scaled by the N_e of the population at the initial generation (N_{anc})). Singleton density and diversity were calculated from simulations of demography with BGS (violet lines) and simulations of demography without BGS (orange lines). Singleton density insets show calculations for generations -1.0 to -0.9 N_{anc} generations in the past (note y-axes for insets are log-scaled). Diversity inset for model 3 shows calculations for all generations but the y-axis is log-scaled to show better detail. Envelopes are 95% CIs calculated from 10,000 bootstraps of the original simulation data (note: only inset plots are small enough to display envelopes).

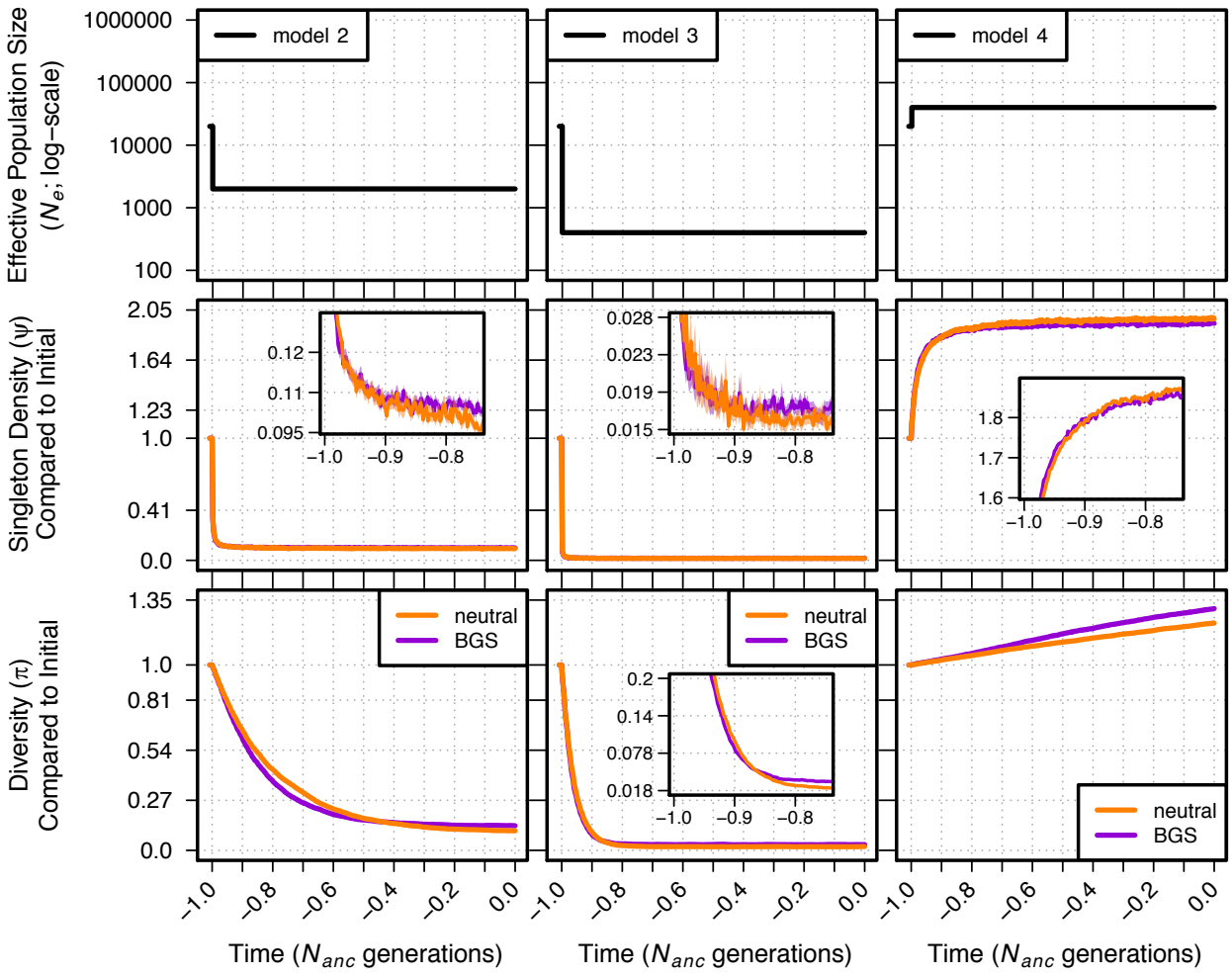


Figure B.4. Singleton density (ψ) and diversity (π) compared to the initial generation for demographic models 2-4.

Top panel shows each demographic model (time proceeds forward from left to right; time is scaled by the N_e of the population at the initial generation (N_{anc})). Singleton density (ψ) and diversity (π) from neutral and BGS simulations were measured in comparison to their initial values in the first generation of the demographic model and are shown in orange and violet lines, respectively. The first value on the y-axis for each demographic model is always 1. For greater detail, insets show data for generations over a smaller time scale and smaller y-axis (note: y-axes for insets are scaled linearly). Envelopes are 95% CIs calculated from 10,000 bootstraps of the original simulation data (note: only inset plots are small enough to display envelopes). The data used for this figure is identical to that of Figure B.3.

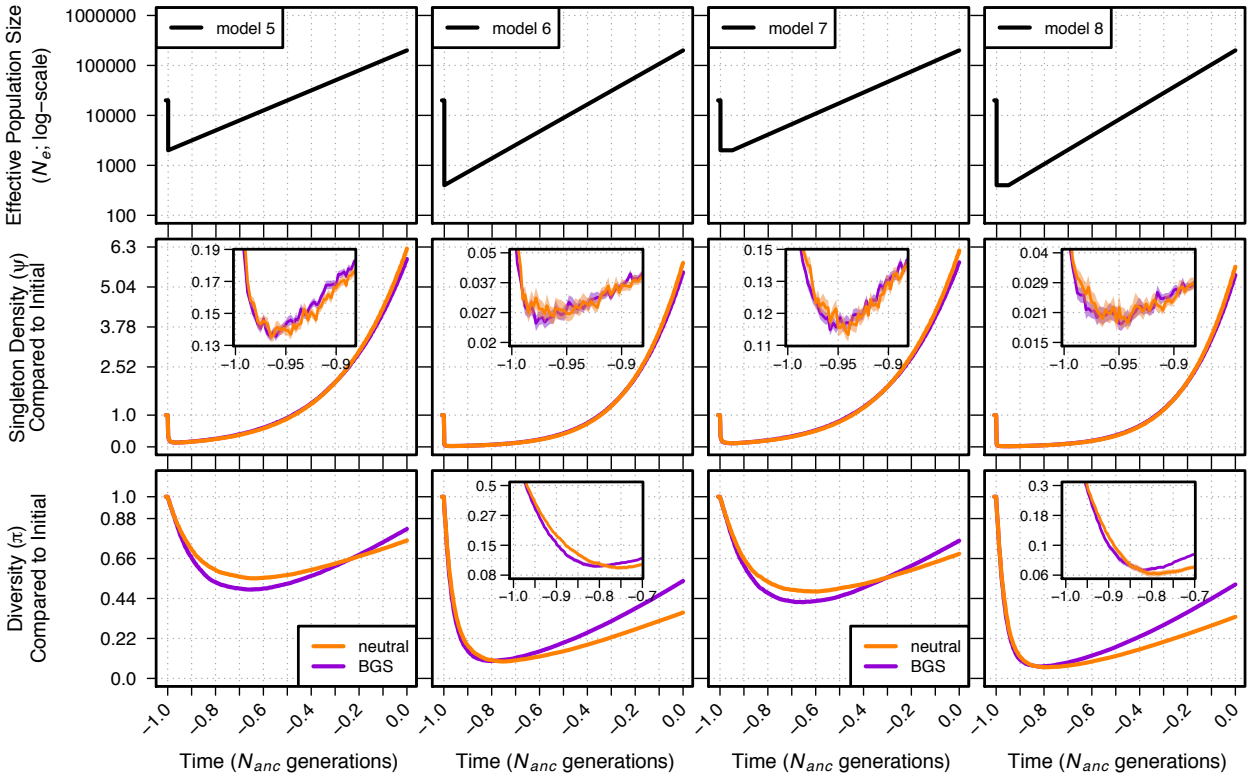


Figure B.5. Singleton density (ψ) and diversity (π) compared to the initial generation for demographic models 5-8.

Top panel shows each demographic model (time proceeds forward from left to right; time is scaled by the N_e of the population at the initial generation (N_{anc})). Singleton density (ψ) and diversity (π) from neutral and BGS simulations were measured in comparison to their initial values in the first generation of the demographic model and are shown in orange and violet lines, respectively. The first value on the y-axis for each demographic model is always 1. For greater detail, insets show data for generations over a smaller time scale and smaller y-axis (note: y-axes for insets are log-scaled). Envelopes are 95% CIs calculated from 10,000 bootstraps of the original simulation data (note: only inset plots are small enough to display envelopes). The data used for this figure is identical to that of Figure B.7.

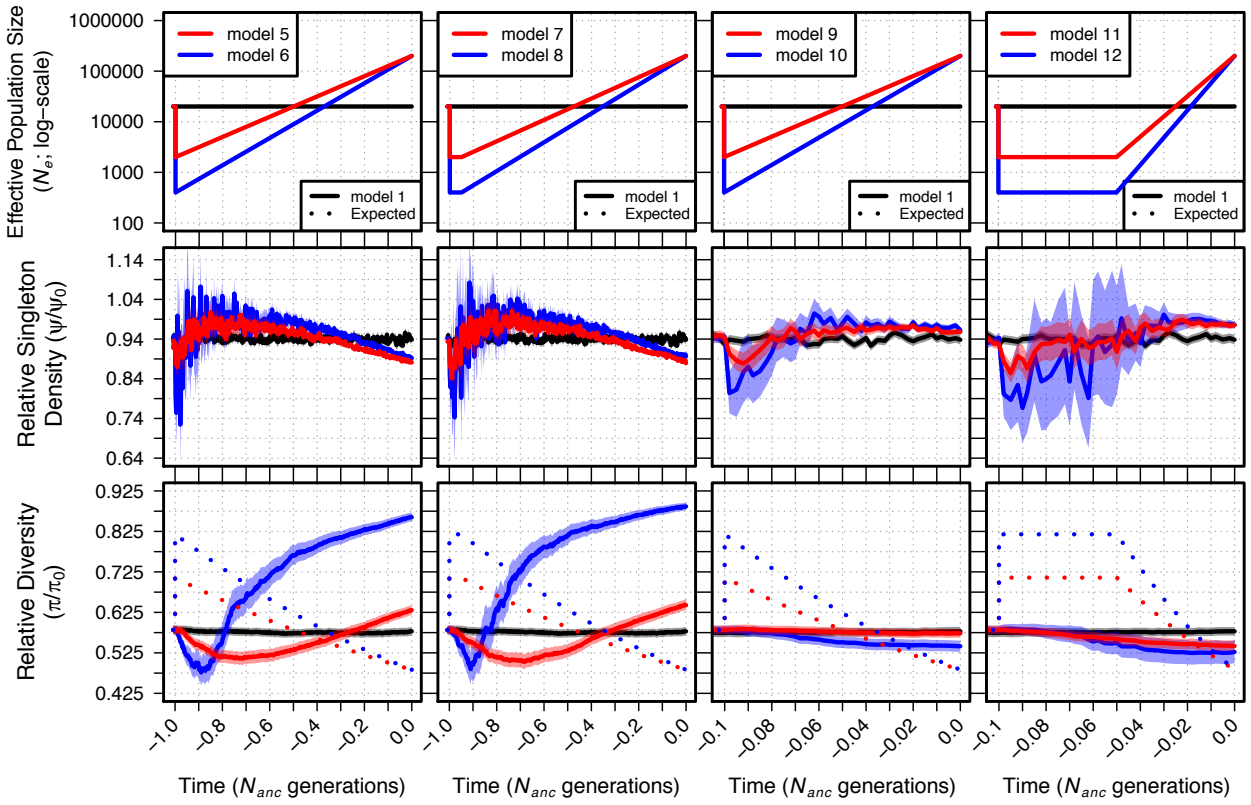


Figure B.6. Relative singleton density (ψ/ψ_0) and relative diversity (π/π_0) across time for demographic models 1 and 5-12.
Top panel shows each demographic model (time proceeds forward from left to right; time is scaled by the N_e of the population at the initial generation (N_{anc})). Black lines show ψ/ψ_0 and π/π_0 from simulations of a constant sized population (model 1). Dotted lines in the bottom panel show the expectation of π/π_0 from Eq. (14) of Nordborg et al. 1996 for each demographic model given the specific selection parameters and N_e at each time point. See Table B.1 for demographic model parameters. Envelopes are 95% CIs calculated from 10,000 bootstraps of the original simulation data.

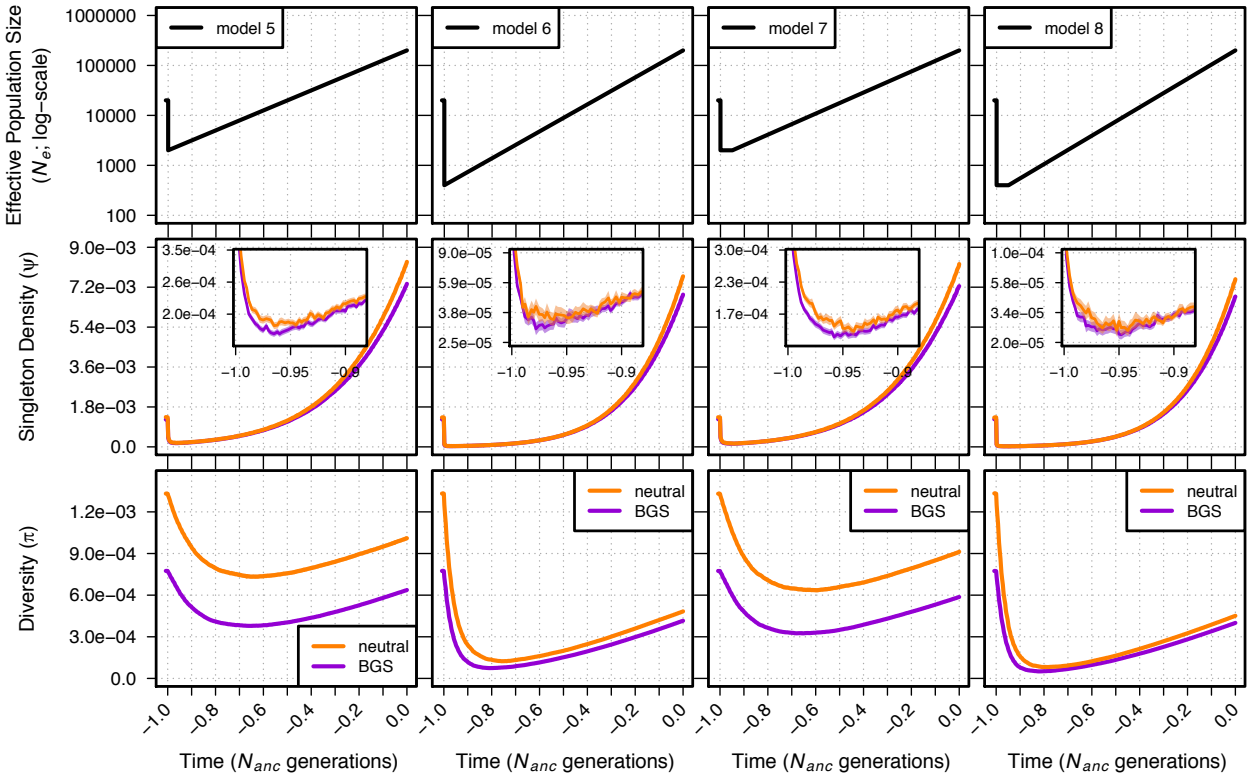


Figure B.7. Singleton density (ψ per site) and diversity (π per site) for models 5-8. Top panel shows each demographic model (time proceeds forward from left to right; time is scaled by the N_e of the population at the initial generation (N_{anc})). Singleton density and diversity were calculated from simulations of demography with BGS (violet lines) and simulations of demography without BGS (orange lines). Insets show calculations of singleton density for generations -1.0 to -0.9 N_{anc} generations in the past (note: y-axes for insets are log-scaled). Envelopes are 95% CIs calculated from 10,000 bootstraps of the original simulation data (note: only inset plots are small enough to display envelopes).

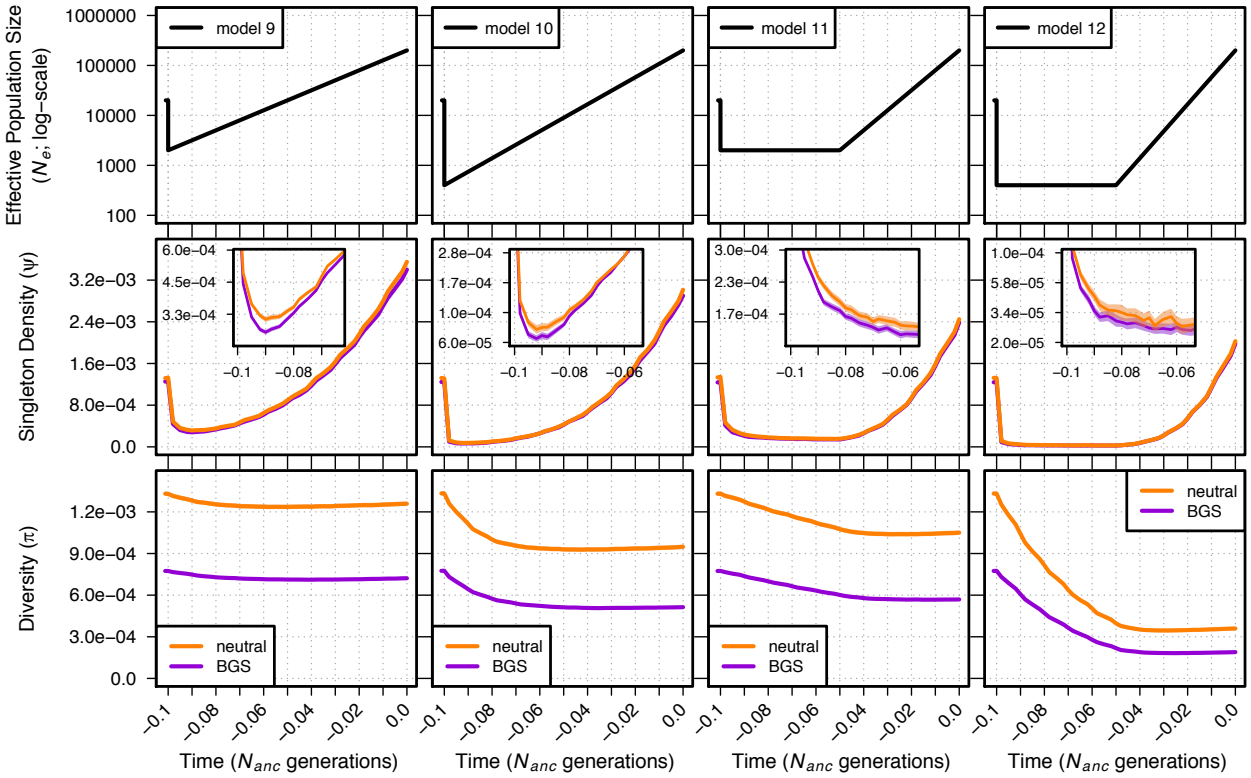


Figure B.8. Singleton density (ψ per site) and diversity (π per site) for models 9-12. Top panel shows each demographic model (time proceeds forward from left to right; time is scaled by the N_e of the population at the initial generation (N_{anc})). Singleton density and diversity were calculated from simulations of demography with BGS (violet lines) and simulations of demography without BGS (orange lines). Insets show calculations of singleton density for generations -0.1 to -0.06 N_{anc} generations in the past (note: y-axes for insets are log-scaled). Envelopes are 95% CIs calculated from 10,000 bootstraps of the original simulation data (note: only inset plots are small enough to display envelopes).

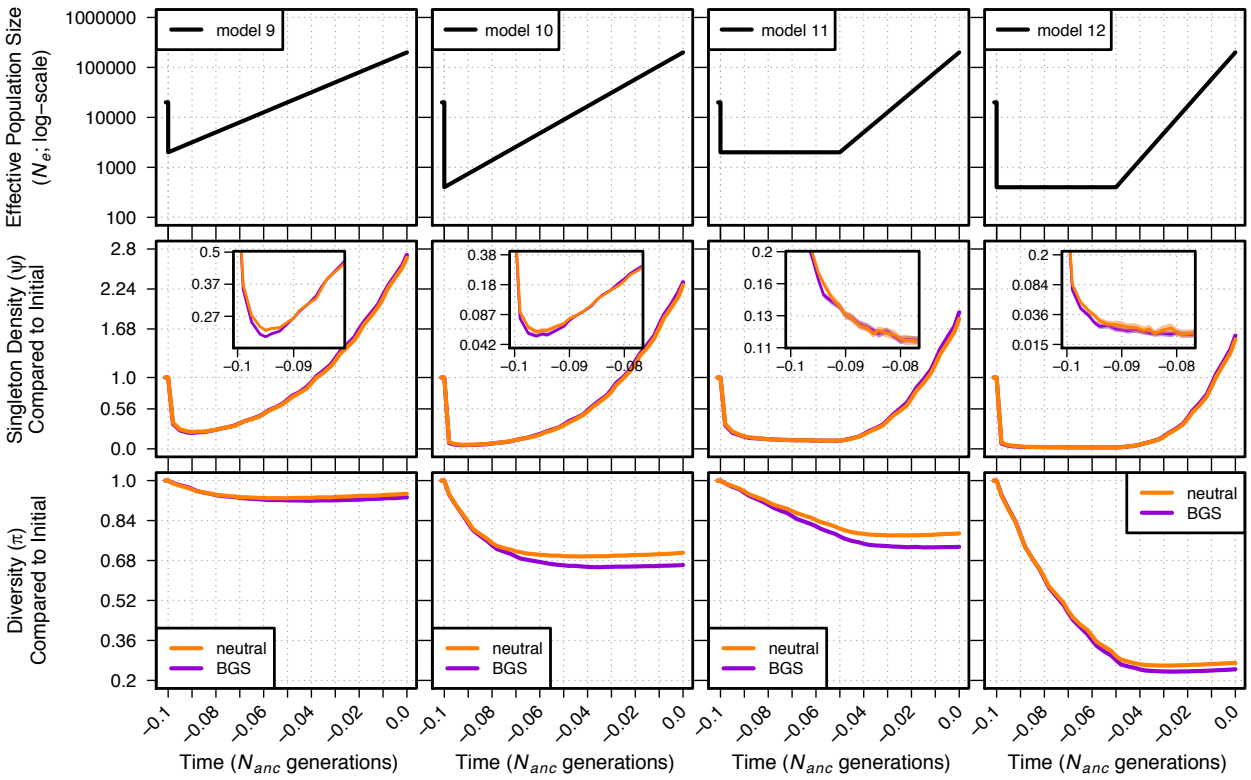


Figure B.9. Singleton density (ψ) and diversity (π) compared to the initial generation for demographic models 9-12.

Top panel shows each demographic model (time proceeds forward from left to right; time is scaled by the N_e of the population at the initial generation (N_{anc})). Singleton density (ψ) and diversity (π) from neutral and BGS simulations were measured in comparison to their initial values in the first generation of the demographic model and are shown in orange and violet lines, respectively. The first value on the y-axis for each demographic model is always 1. For greater detail, insets show data for generations over a smaller time scale and smaller y-axis (note: y-axes for insets are log-scaled). Envelopes are 95% CIs calculated from 10,000 bootstraps of the original simulation data (note: only inset plots are small enough to display envelopes). The data used for this figure is identical to that of Figure B.8.

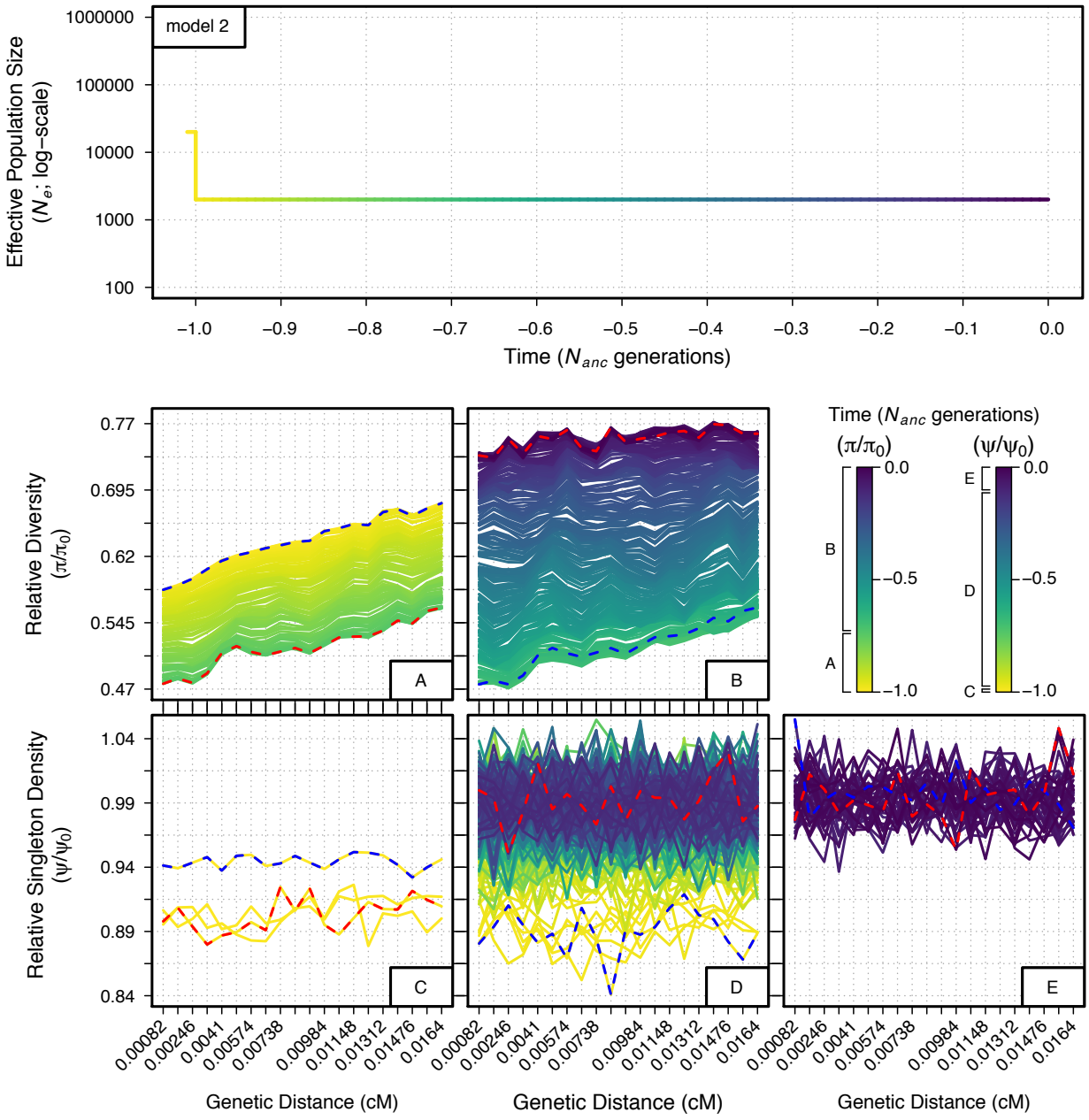

model 2
Effective Population Size
(N_e; log−scale)
1000000
100000
10000
1000
100
−1.0
−0.9
−0.8
−0.7
−0.6
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
Time (N_{anc} generations)
Relative Diversity
(π/π_0)
0.77
0.695
0.62
0.545
0.47
A
B
Time (N_{anc} generations)
(π/π_0)
(ψ/ψ_0)
0.0
−0.5
−1.0
Relative Singleton Density
(ψ/ψ_0)
1.04
0.99
0.94
0.89
0.84
C
D
E
0.00082
0.00246
0.0041
0.00574
0.00738
0.00984
0.01148
0.01312
0.01476
0.0164
Genetic Distance (cM)

Figure B.10

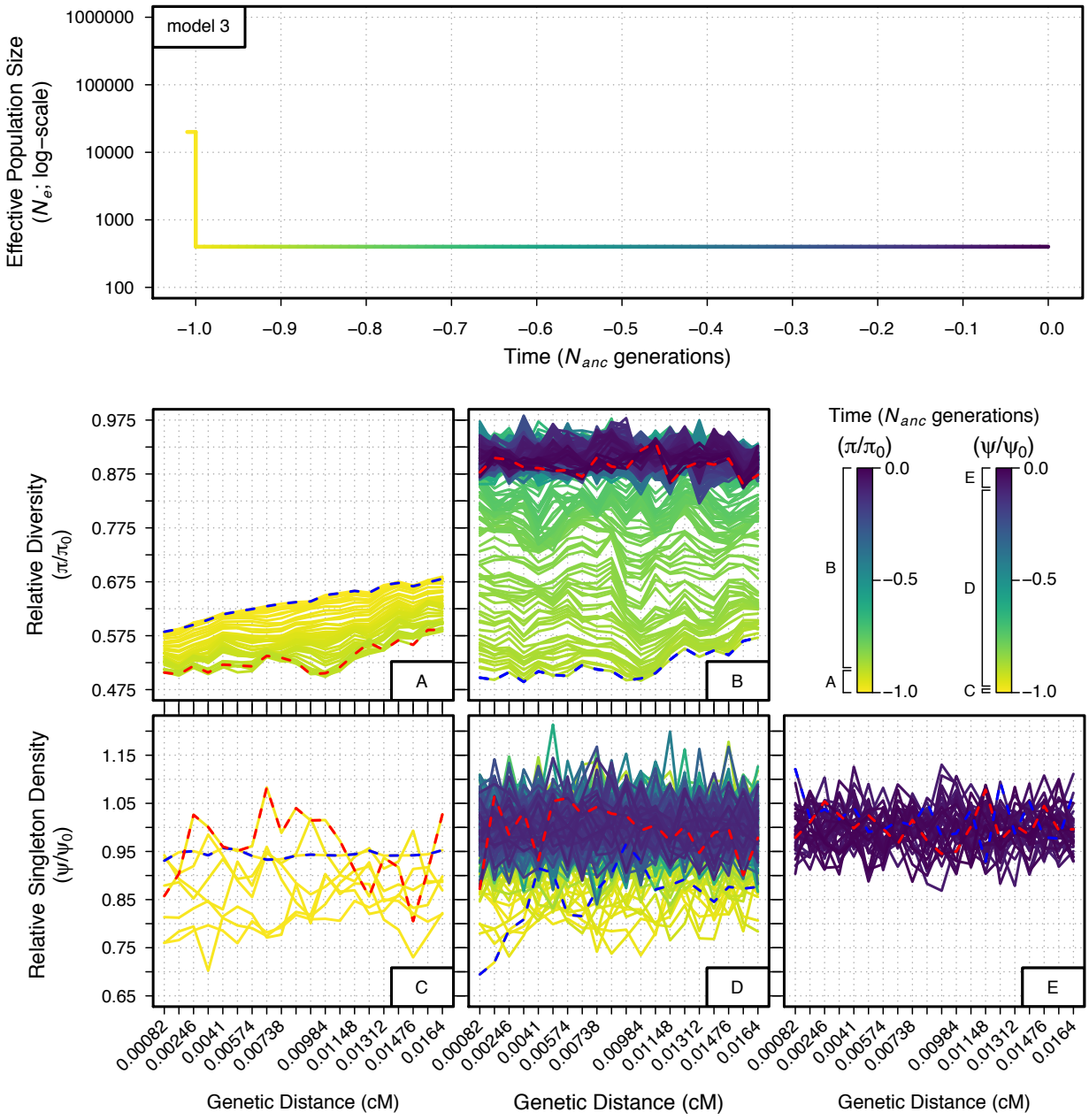

model 3
Effective Population Size
(N_e; log−scale)
1000000
100000
10000
1000
100
−1.0
−0.9
−0.8
−0.7
−0.6
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
Time (N_{anc} generations)
Relative Diversity
(π/π_0)
0.975
0.875
0.775
0.675
0.575
0.475
A
B
Time (N_{anc} generations)
(π/π_0)
(ψ/ψ_0)
0.0
−0.5
−1.0
Relative Singleton Density
(ψ/ψ_0)
1.15
1.05
0.95
0.85
0.75
0.65
C
D
E
0.00082
0.00246
0.0041
0.00574
0.00738
0.00984
0.01148
0.01312
0.01476
0.0164
Genetic Distance (cM)

Figure B.11

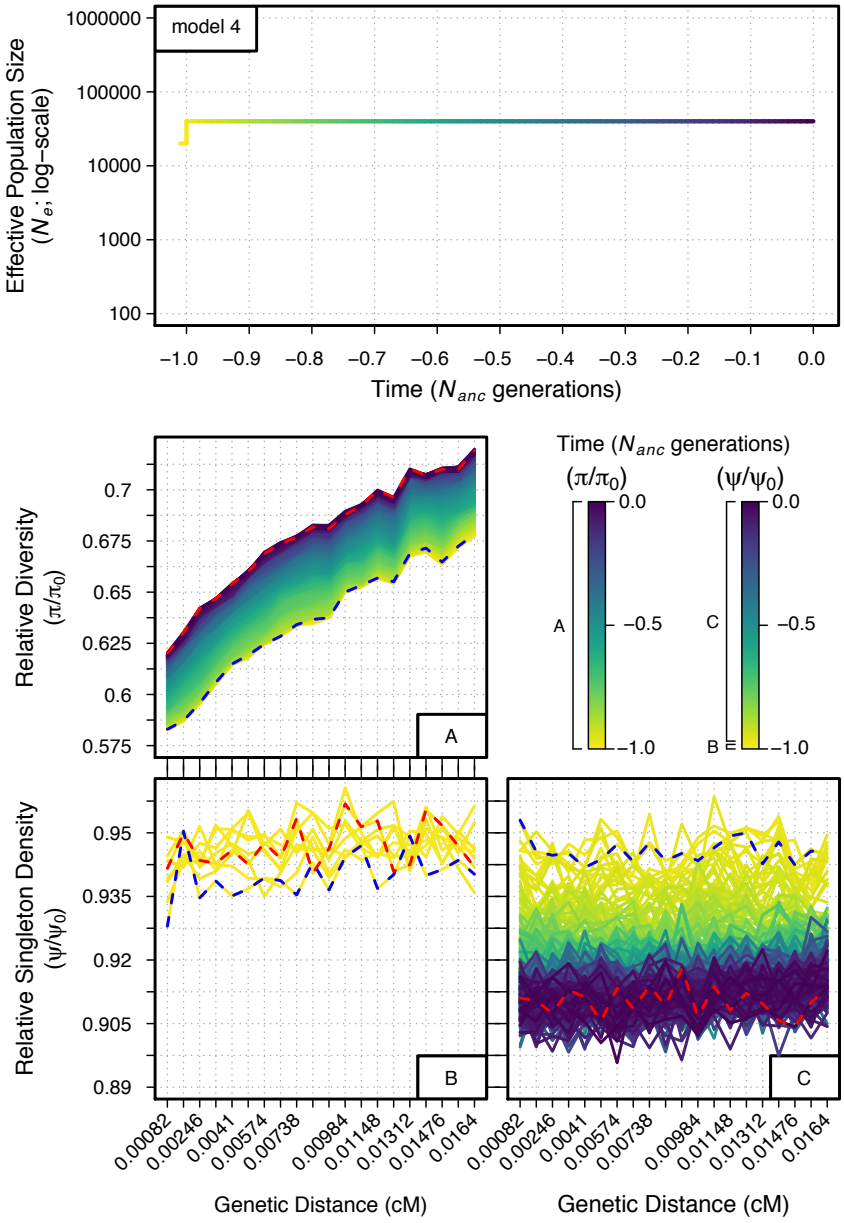

model 4
Effective Population Size
(N_e; log−scale)
1000000
100000
10000
1000
100
−1.0 −0.9 −0.8 −0.7 −0.6 −0.5 −0.4 −0.3 −0.2 −0.1 0.0
Time (N_{anc} generations)
Time (N_{anc} generations)
(π/π_0)
(ψ/ψ_0)
0.0
−0.5
−1.0
A
B
C
Relative Diversity
(π/π_0)
0.7
0.675
0.65
0.625
0.6
0.575
Relative Singleton Density
(ψ/ψ_0)
0.95
0.935
0.92
0.905
0.89
0.00082 0.00246 0.0041 0.00574 0.00738 0.00984 0.01148 0.01312 0.01476 0.0164
Genetic Distance (cM)
Genetic Distance (cM)

Figure B.12

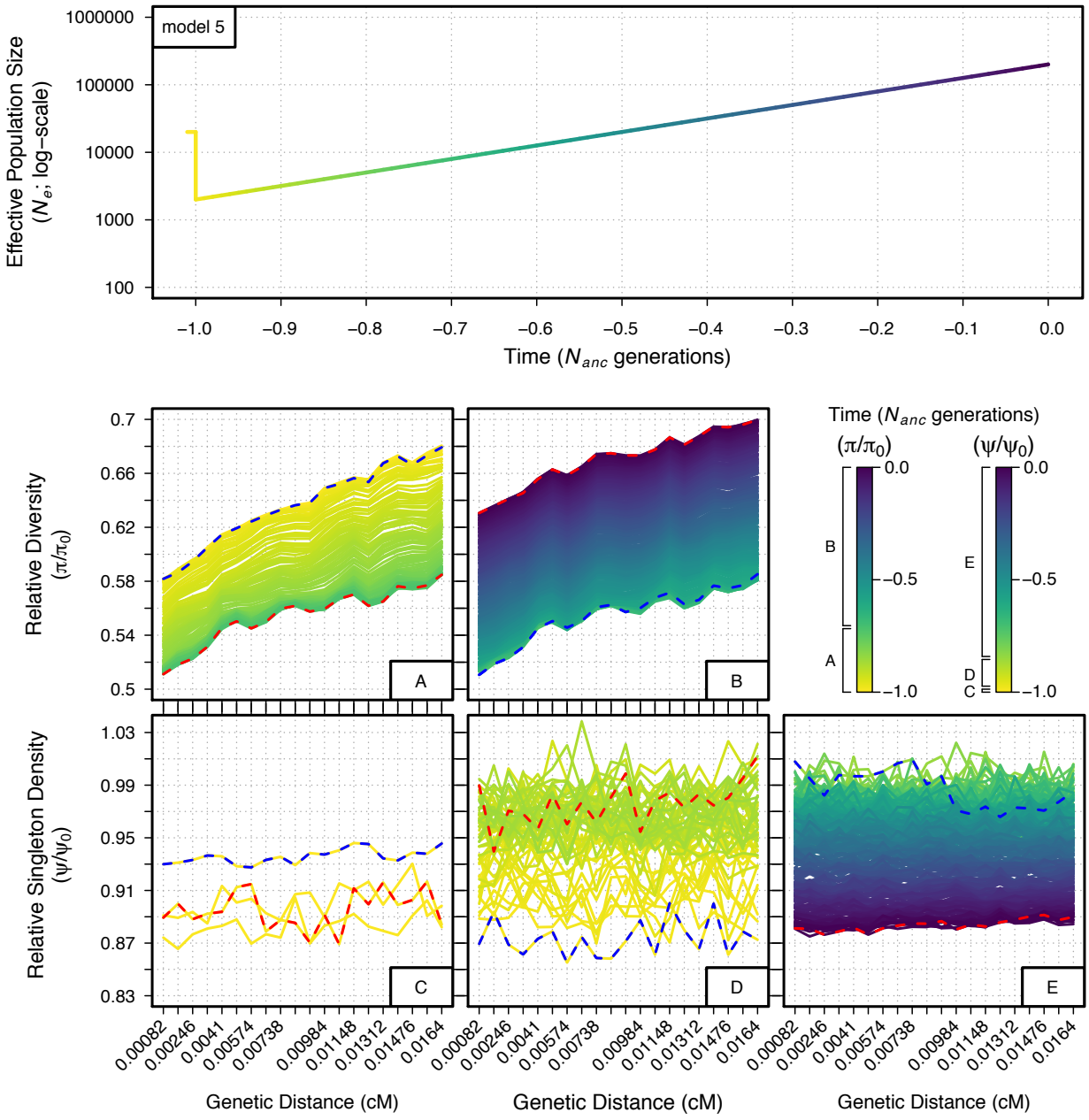

model 5
Effective Population Size
(N_e; log−scale)
1000000
100000
10000
1000
100
−1.0
−0.9
−0.8
−0.7
−0.6
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
Time (N_{anc} generations)
Relative Diversity
(π/π_0)
0.7
0.66
0.62
0.58
0.54
0.5
A
B
Time (N_{anc} generations)
(π/π_0)
(ψ/ψ_0)
0.0
−0.5
−1.0
B
A
E
D
C
Relative Singleton Density
(ψ/ψ_0)
1.03
0.99
0.95
0.91
0.87
0.83
C
D
E
0.00082
0.00246
0.0041
0.00574
0.00738
0.00984
0.01148
0.01312
0.01476
0.0164
Genetic Distance (cM)
Genetic Distance (cM)
Genetic Distance (cM)

Figure B.13

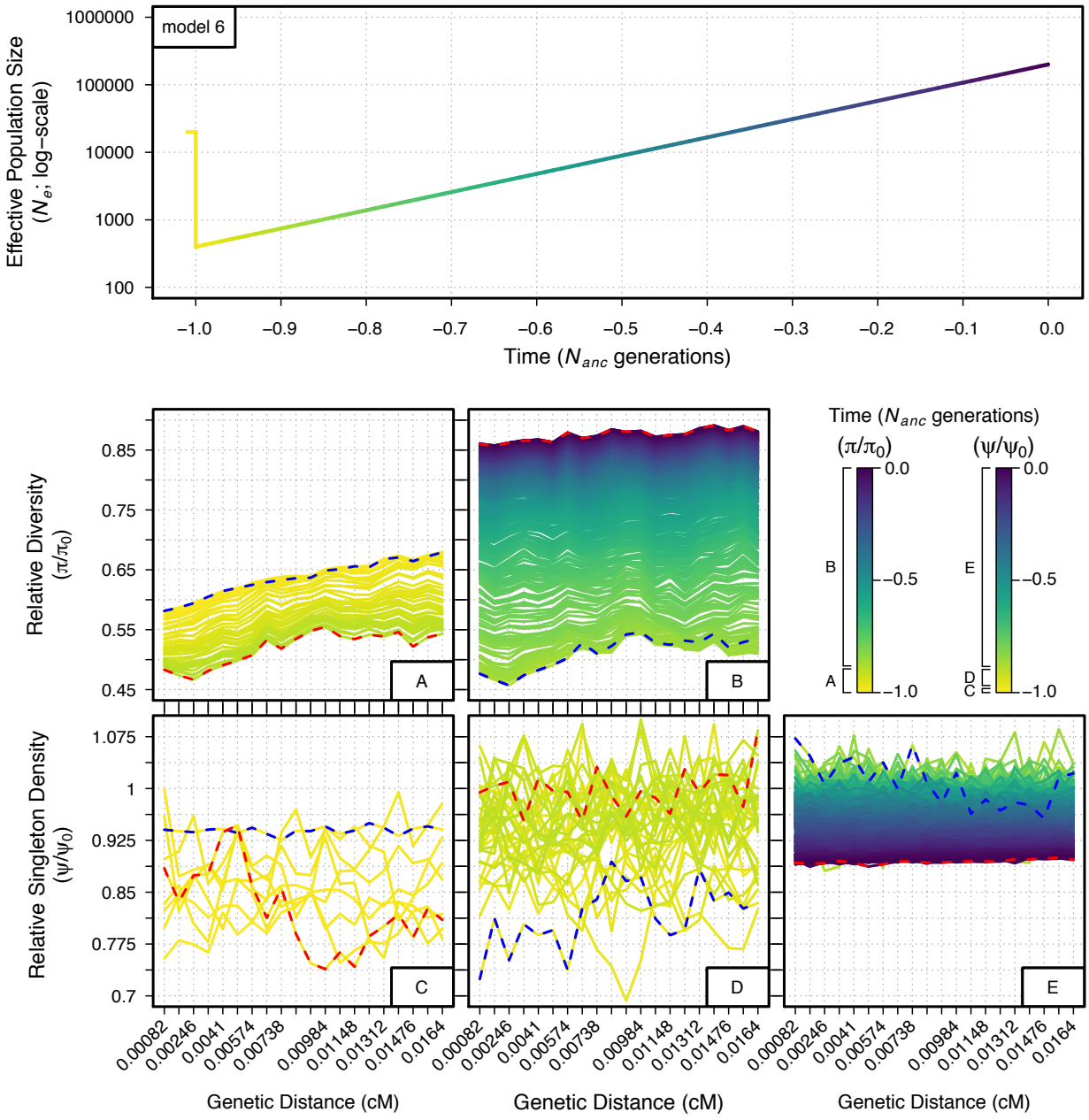

model 6
Effective Population Size
(N_e; log−scale)
1000000
100000
10000
1000
100
−1.0
−0.9
−0.8
−0.7
−0.6
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
Time (N_{anc} generations)
Relative Diversity
(π/π_0)
0.85
0.75
0.65
0.55
0.45
A
B
Time (N_{anc} generations)
(π/π_0)
(ψ/ψ_0)
0.0
−0.5
−1.0
B
A
E
D
C
Relative Singleton Density
(ψ/ψ_0)
1.075
1
0.925
0.85
0.775
0.7
C
D
E
0.00082
0.00246
0.0041
0.00574
0.00738
0.00984
0.01148
0.01312
0.01476
0.0164
Genetic Distance (cM)
Genetic Distance (cM)
Genetic Distance (cM)

Figure B.14

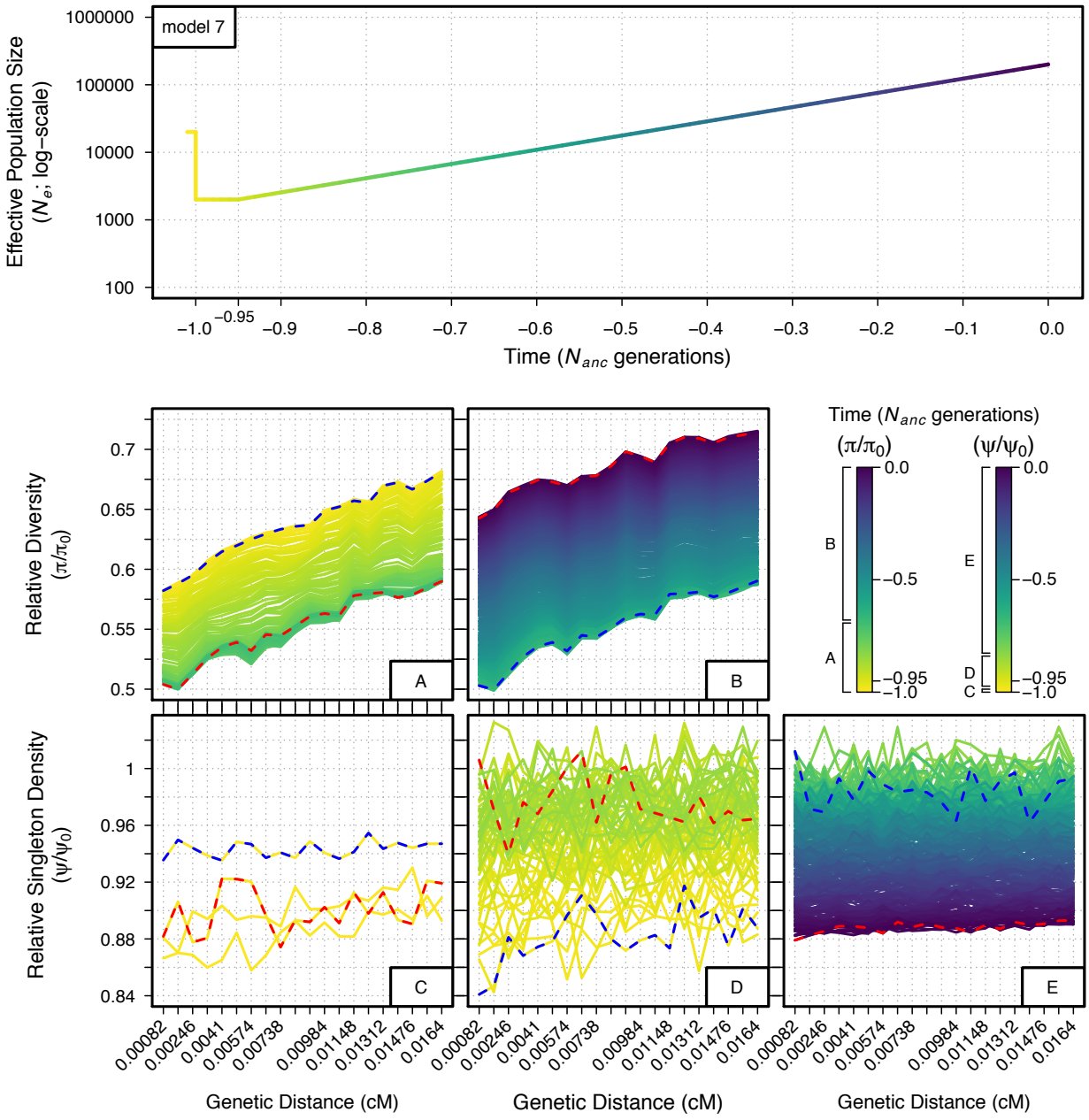

model 7
Effective Population Size
(N_e; log−scale)
1000000
100000
10000
1000
100
−1.0
−0.95
−0.9
−0.8
−0.7
−0.6
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
Time (N_{anc} generations)
Relative Diversity
(π/π_0)
0.7
0.65
0.6
0.55
0.5
A
B
Time (N_{anc} generations)
(π/π_0)
(ψ/ψ_0)
0.0
−0.5
−0.95
−1.0
B
A
E
D
C
Relative Singleton Density
(ψ/ψ_0)
1
0.96
0.92
0.88
0.84
C
D
E
0.00082
0.00246
0.0041
0.00574
0.00738
0.00984
0.01148
0.01312
0.01476
0.0164
Genetic Distance (cM)
Genetic Distance (cM)
Genetic Distance (cM)

Figure B.15

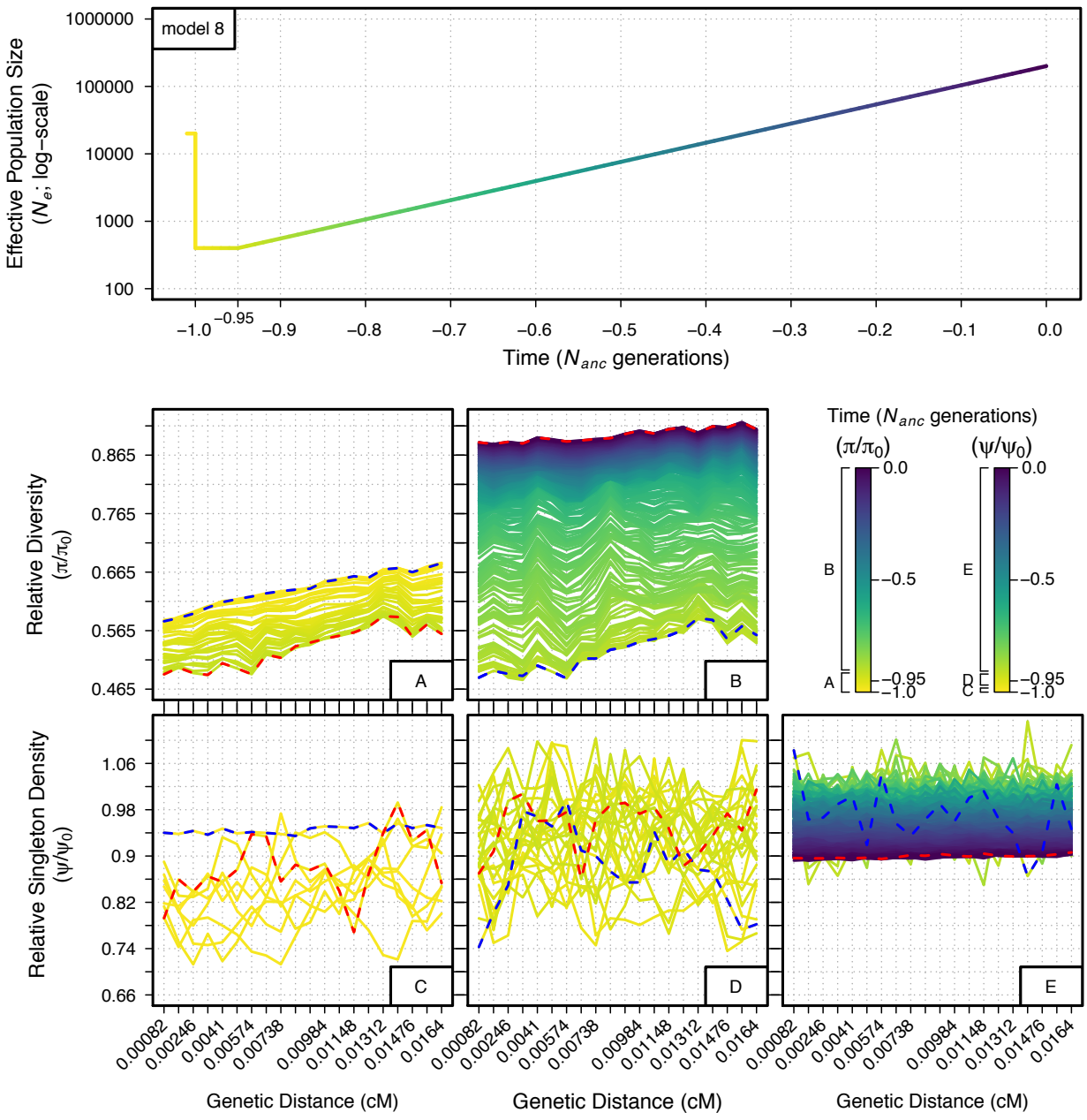

model 8
Effective Population Size
(N_e; log−scale)
1000000
100000
10000
1000
100
−1.0
−0.95
−0.9
−0.8
−0.7
−0.6
−0.5
−0.4
−0.3
−0.2
−0.1
0.0
Time (N_{anc} generations)
Relative Diversity
(π/π_0)
0.865
0.765
0.665
0.565
0.465
A
B
Time (N_{anc} generations)
(π/π_0)
(ψ/ψ_0)
0.0
−0.5
−0.95
−1.0
Relative Singleton Density
(ψ/ψ_0)
1.06
0.98
0.9
0.82
0.74
0.66
C
D
E
0.00082
0.00246
0.0041
0.00574
0.00738
0.00984
0.01148
0.01312
0.01476
0.0164
Genetic Distance (cM)

Figure B.16

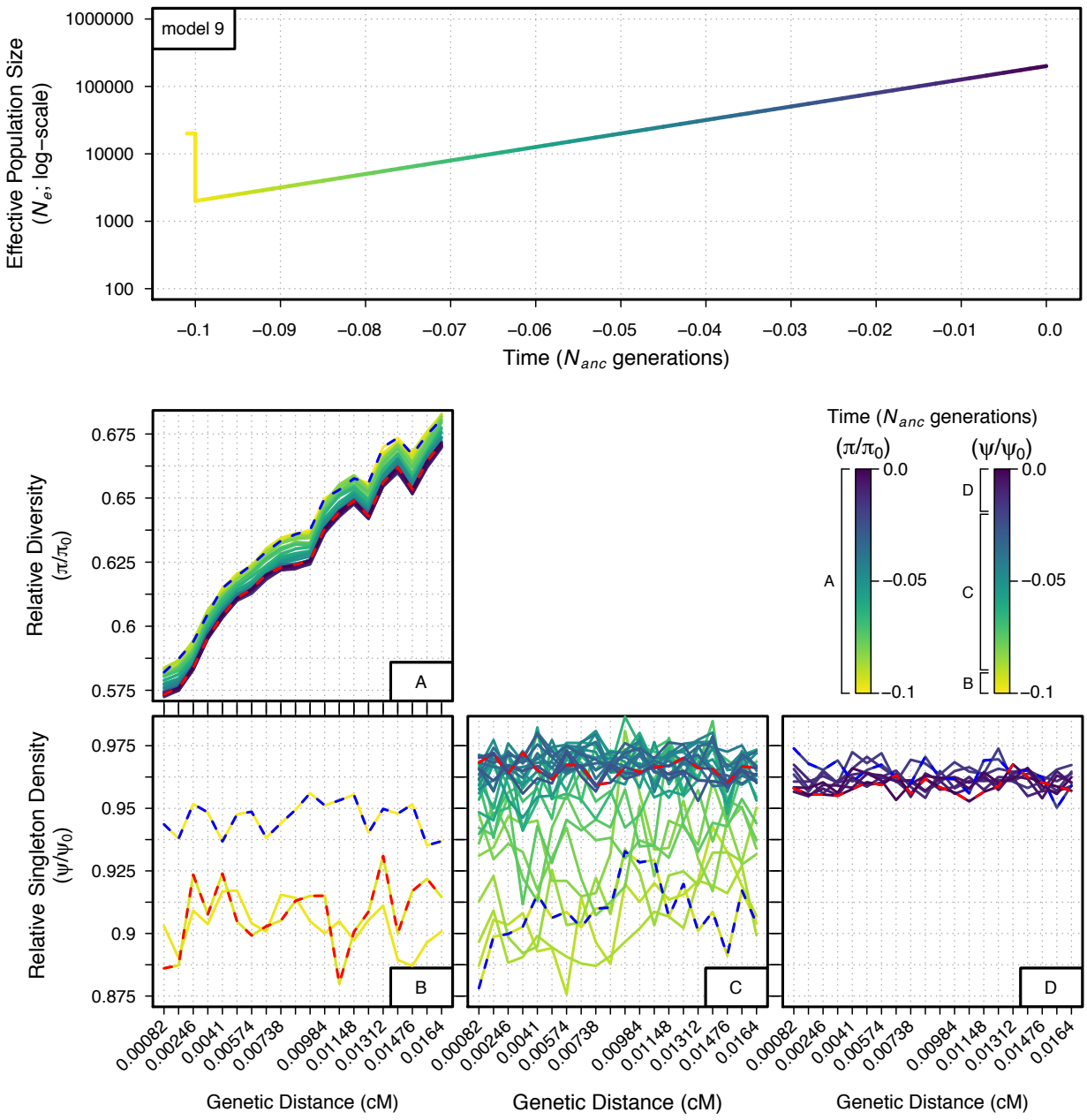

model 9
Effective Population Size
(N_e; log−scale)
1000000
100000
10000
1000
100
−0.1
−0.09
−0.08
−0.07
−0.06
−0.05
−0.04
−0.03
−0.02
−0.01
0.0
Time (N_{anc} generations)
Relative Diversity
(π/π_0)
0.675
0.65
0.625
0.6
0.575
A
Time (N_{anc} generations)
(π/π_0)
(ψ/ψ_0)
0.0
−0.05
−0.1
A
D
C
B
Relative Singleton Density
(ψ/ψ_0)
0.975
0.95
0.925
0.9
0.875
B
C
D
0.00082
0.00246
0.0041
0.00574
0.00738
0.00984
0.01148
0.01312
0.01476
0.0164
Genetic Distance (cM)

Figure B.17

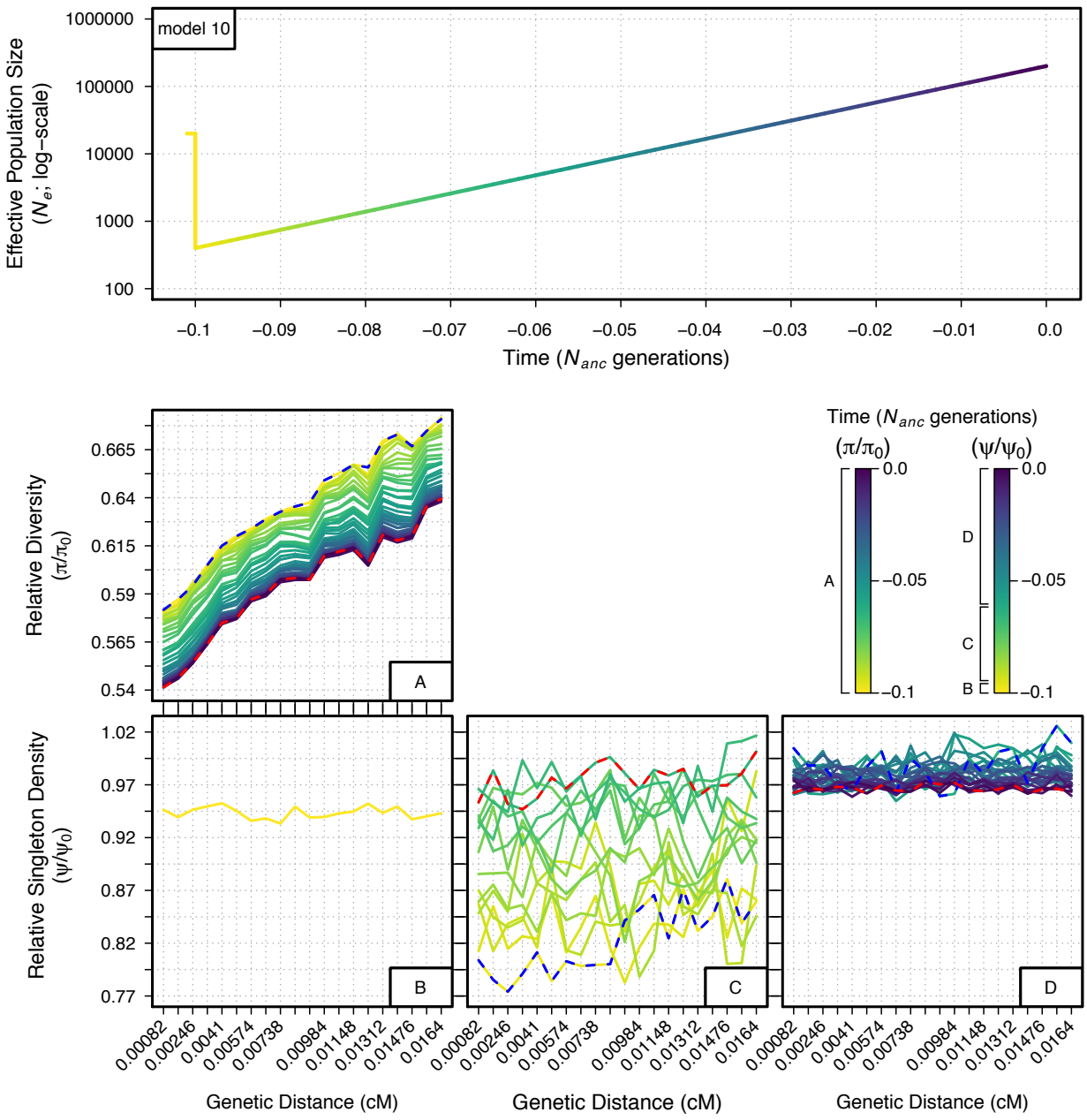

model 10
Effective Population Size
(N_e; log−scale)
1000000
100000
10000
1000
100
−0.1
−0.09
−0.08
−0.07
−0.06
−0.05
−0.04
−0.03
−0.02
−0.01
0.0
Time (N_{anc} generations)
Relative Diversity
(π/π_0)
0.665
0.64
0.615
0.59
0.565
0.54
A
Time (N_{anc} generations)
(π/π_0)
(ψ/ψ_0)
0.0
−0.05
−0.1
A
D
C
B
Relative Singleton Density
(ψ/ψ_0)
1.02
0.97
0.92
0.87
0.82
0.77
B
C
D
0.00082
0.00246
0.0041
0.00574
0.00738
0.00984
0.01148
0.01312
0.01476
0.0164
Genetic Distance (cM)
Genetic Distance (cM)
Genetic Distance (cM)

Figure B.18

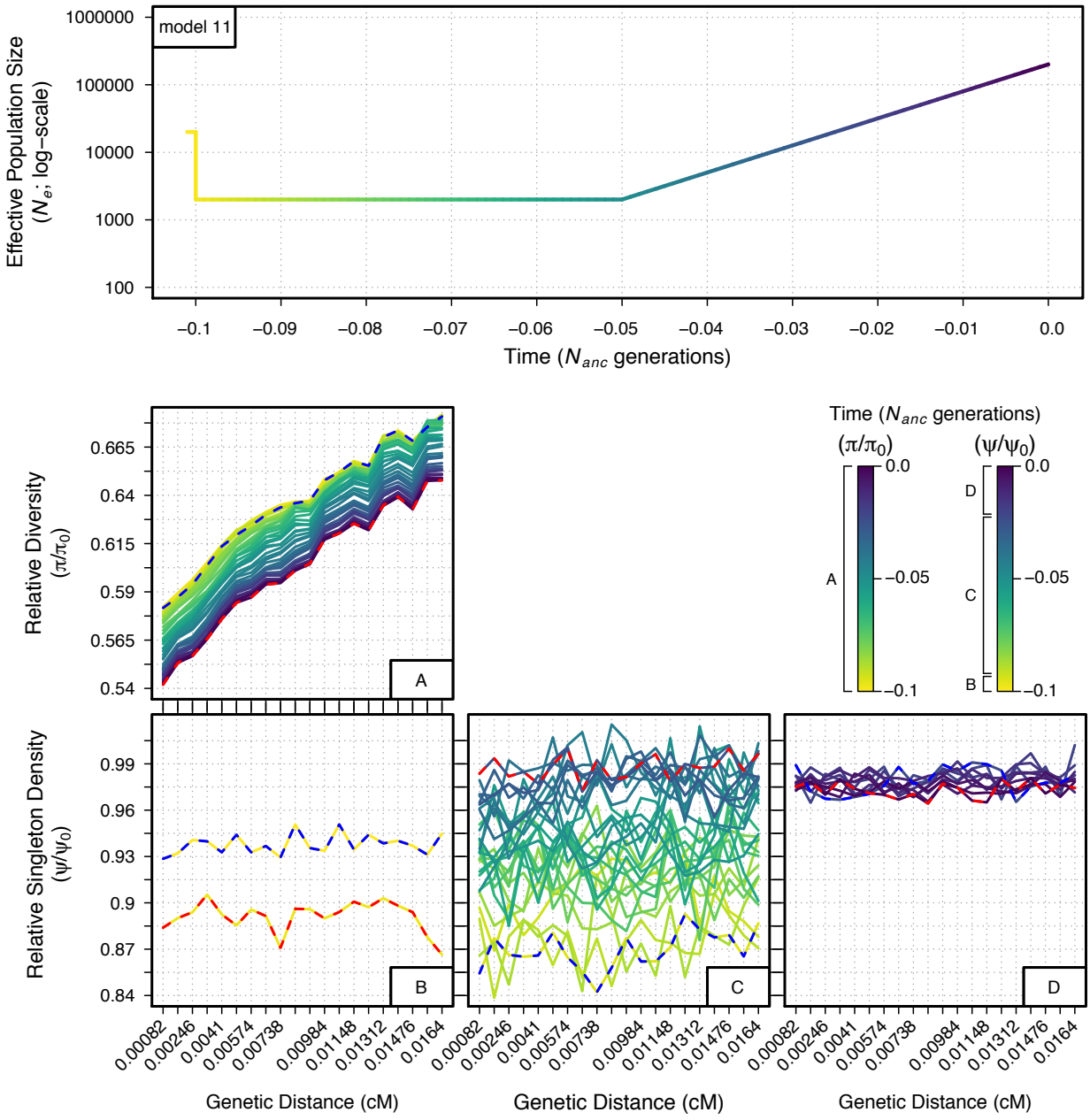

model 11
Effective Population Size
(N_e; log−scale)
1000000
100000
10000
1000
100
−0.1
−0.09
−0.08
−0.07
−0.06
−0.05
−0.04
−0.03
−0.02
−0.01
0.0
Time (N_{anc} generations)
Relative Diversity
(π/π_0)
0.665
0.64
0.615
0.59
0.565
0.54
A
Time (N_{anc} generations)
(π/π_0)
(ψ/ψ_0)
0.0
−0.05
−0.1
A
D
C
B
Relative Singleton Density
(ψ/ψ_0)
0.99
0.96
0.93
0.9
0.87
0.84
B
C
D
0.00082
0.00246
0.0041
0.00574
0.00738
0.00984
0.01148
0.01312
0.01476
0.0164
Genetic Distance (cM)
Genetic Distance (cM)
Genetic Distance (cM)

Figure B.19

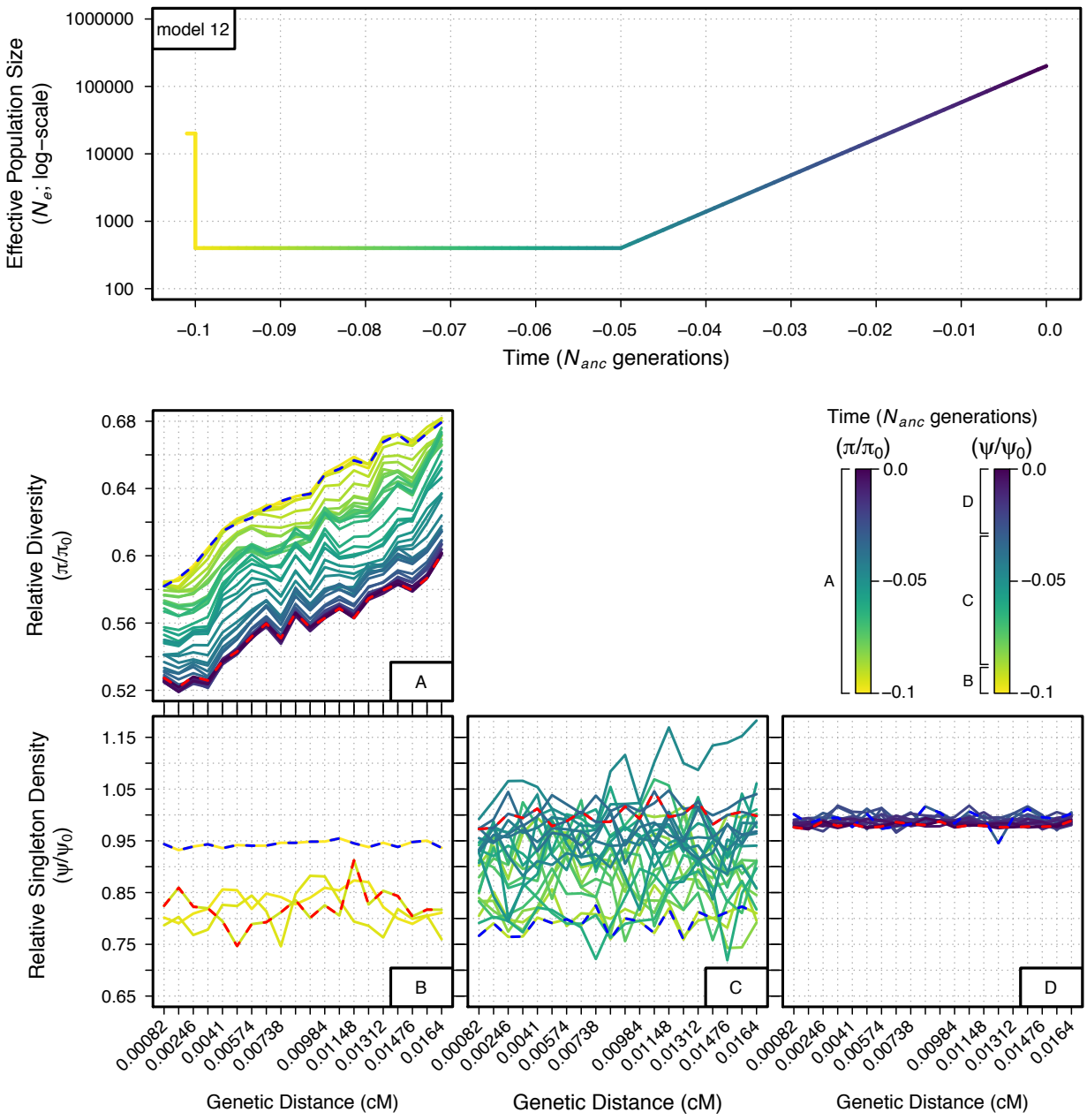

model 12
Effective Population Size
(N_e; log−scale)
1000000
100000
10000
1000
100
−0.1
−0.09
−0.08
−0.07
−0.06
−0.05
−0.04
−0.03
−0.02
−0.01
0.0
Time (N_{anc} generations)
Relative Diversity
(π/π_0)
0.68
0.64
0.6
0.56
0.52
A
Time (N_{anc} generations)
(π/π_0)
(ψ/ψ_0)
0.0
−0.05
−0.1
A
D
C
B
Relative Singleton Density
(ψ/ψ_0)
1.15
1.05
0.95
0.85
0.75
0.65
B
C
D
0.00082
0.00246
0.0041
0.00574
0.00738
0.00984
0.01148
0.01312
0.01476
0.0164
Genetic Distance (cM)
Genetic Distance (cM)
Genetic Distance (cM)

Figure B.20

Figures B.10-B.20. Relative diversity (π/π_0) and singleton density (ψ/ψ_0) through time for demographic models 2-12 measured across a neutral 200 kb region under the effects of BGS.

The genetic distance of each 10 kb bin from the selected locus is indicated on the x-axes, with genetic distance increasing from left to right. Each line measuring π/π_0 and ψ/ψ_0 across the 200 kb neutral region represents a specific generation of the demographic model (401 discrete generations for demographic models 2-8, 41 discrete generations for demographic models 9-12). Specific generations are indicated by the color of the demographic model at the top of each figure (time proceeds forward from left to right; time is scaled by the N_e of the population at the initial generation (N_{anc})) and in the figure legend. When necessary, multiple plots are given for π/π_0 and ψ/ψ_0 in order to prevent overlap of the measurements between generations (see legend for specific generations covered in each plot). Blue dashed lines and red dashed lines indicate the first generation and last generation measured, respectively, for each specific plot.

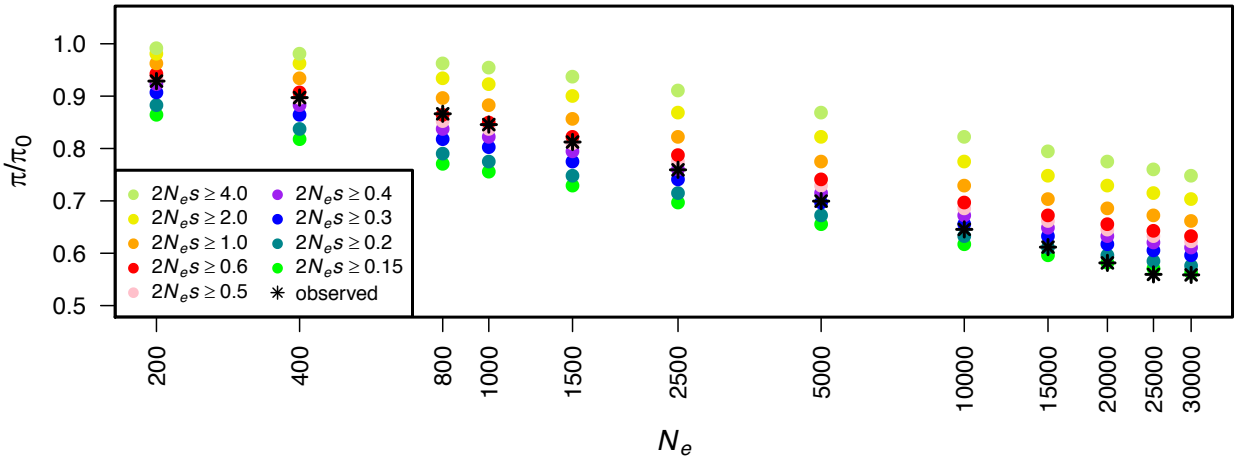


Figure B.21. Estimate of π/π_0 from the Nordborg model across different population sizes and different truncation thresholds on selection.
The resulting π/π_0 calculated from the Nordborg model is shown on the y-axis across various populations sizes. Different γ values were used to truncate s for the Nordborg model and are shown in the legend ($2N_e s \geq \gamma$). The black stars represent the observed π/π_0 from running simulations of BGS.

parameter	model 1	model 2	model 3	model 4	model 5	model 6	model 7	model 8	model 9	model 10	model 11	model 12
ancestral population size (N_{anc})	20000	20000	20000	20000	20000	20000	20000	20000	20000	20000	20000	20000
bottleneck/ expansion population size (N_e)	NA	2000	400	40000	2000	400	2000	400	2000	400	2000	400
bottleneck time length (N_{anc} generations)	NA	1	1	NA	0	0	0.05	0.05	0	0	0.05	0.05
expansion time length (N_{anc} generations)*	NA	NA	NA	1	1	1	0.95	0.95	0.1	0.1	0.05	0.05
final population size (N_e)	20000	2000	400	40000	200000	200000	200000	200000	200000	200000	200000	200000

Table B.1. Demographic parameters for demographic models 1-12

*note that the population expansions of models 5-12 are exponential growth models and that model 4 is an instantaneous growth model

model	difference under neutrality (relative difference)	difference under BGS (relative difference)
model 2	0.0011901 (0.0979034)	0.00111566 (0.101607)
model 3	0.0013119 (0.0160799)	0.00122008 (0.0169049)
model 4	0.00129788 (1.97708)	0.00115996 (1.94102)

Table B.2. Absolute and relative difference between final and initial generations for ψ under neutrality and BGS.

Absolute difference is taken as the difference for ψ under BGS or neutrality between the final generation and initial generation of each demographic model. Relative difference is taken as the quotient from dividing the final generation by the initial generation (note: relative difference is the same as the measurements of the final generation in Figure B.4).

model	difference under neutrality (relative difference)	difference under BGS (relative difference)
model 2	0.00119038 (0.1052)	0.000671873 (0.132625)
model 3	0.00130409 (0.0206869)	0.000751267 (0.0311911)

Table B.3. Absolute and relative difference between final and initial generations for π under neutrality and BGS.
Absolute difference is taken as the difference for π under BGS or neutrality between the final generation and initial generation of each demographic model. Relative difference is taken as the quotient from dividing the final generation by the initial generation (note: relative difference is the same as the measurements of the final generation in Figure B.4). We do not include differences for π for model 4 because equilibrium does not appear to have been reached for that model.

model	difference at initial generation (relative difference)	difference at minimum generation (relative difference)	difference at final generation (relative difference)
model 1	0.100 (1.171)	NA	0.094 (1.163)
model 2	0.098 (1.168)	0.087 (1.183)	0.025 (1.034)
model 3	0.098 (1.169)	0.074 (1.149)	-0.005 (0.995)
model 4	0.095 (1.163)	NA	0.099 (1.159)
model 5	0.098 (1.168)	0.075 (1.146)	0.069 (1.110)
model 6	0.098 (1.168)	0.059 (1.123)	0.021 (1.024)
model 7	0.099 (1.170)	0.087 (1.174)	0.072 (1.111)
model 8	0.099 (1.171)	0.073 (1.150)	0.022 (1.025)
model 9	0.099 (1.170)	NA	0.098 (1.172)
model 10	0.099 (1.170)	NA	0.098 (1.181)
model 11	0.099 (1.170)	NA	0.106 (1.195)
model 12	0.098 (1.168)	NA	0.074 (1.140)

Table B.4. Absolute and relative difference of π/π_0 between the last bin and first bin of the 200 kb neutral region for the initial and final generations and the generation with minimum π/π_0.
Absolute difference is taken as the difference of π/π_0 between the last 10 kb bin and the first 10 kb bin of the 200 kb neutral region for the initial generation and final generation of each model's demographic history. Relative difference is taken as the quotient from dividing π/π_0 in the last 10 kb bin by the first 10 kb bin. For models that experienced a decrease in π/π_0 followed by an increase in π/π_0 (models 2-3, 5-8), we also calculated the absolute and relative difference at the generation where minimum π/π_0 was observed in the first 10 kb bin. The first and last 10 kb bins have a genetic distance of 0.00082 cM and 0.0164 cM from the 2 Mb selected locus, respectively.

APPENDIX C: Supplemental Material to Chapter 4

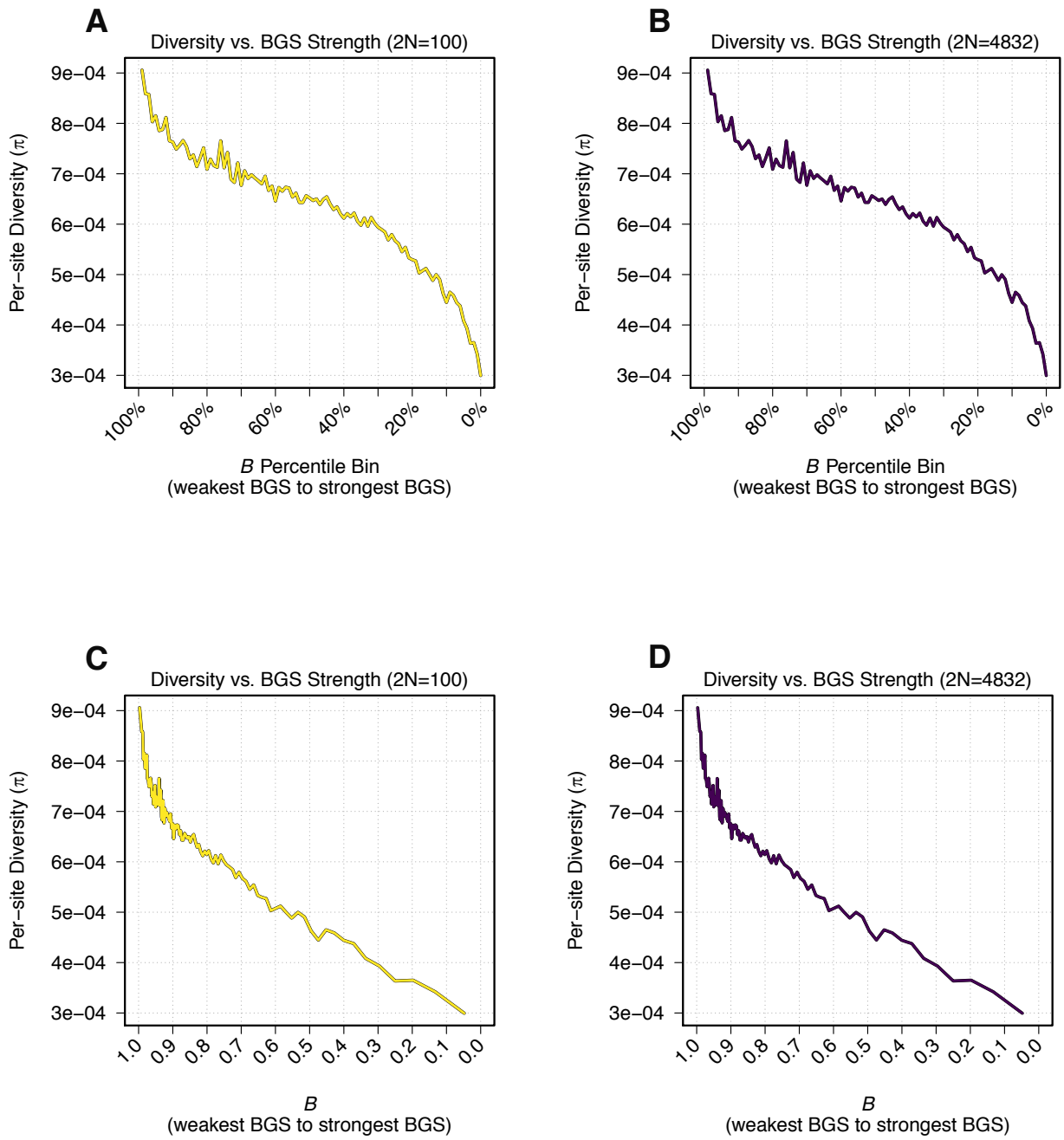


Figure C.1. Per-site diversity as a function of *B* for 2N=100 and 2N=4832.
A) and B) Average pairwise genetic diversity (π) is shown across percentile bins of *B* (lower percentile bins indicate stronger background selection) for the smallest (2N=100) and largest (2N=4832) sample sizes. C) and D) Average pairwise genetic diversity (π) is shown across *B* (lower *B* indicates stronger background selection) for the smallest (2N=100) and largest (2N=4832) sample sizes.

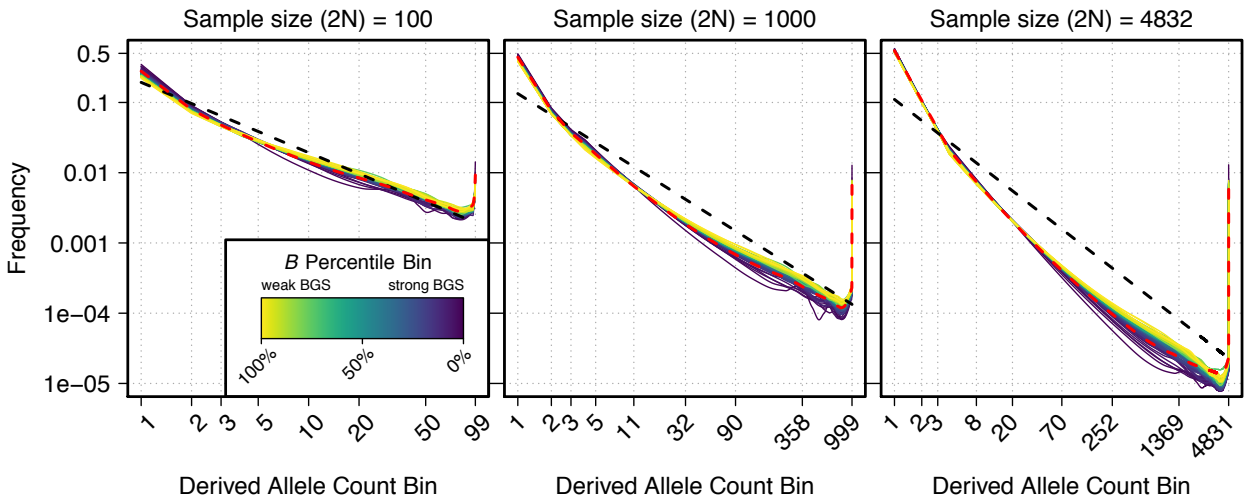


Figure C.2. Site-frequency spectrum (SFS) for different sample sizes and *B*.
SFS data is shown for each of 100 percentile bins of *B* (higher percentiles indicate weaker BGS). Each separate plot shows a different sample size from which the SFS was made. Dashed red lines show the SFS from fourfold degenerate sites. Dashed black lines show the SFS from a standard neutral model for the given sample size. Loess smoothing was conducted for derived allele count bins 4 to 2N-4 with a span of 0.4 to decrease the high variance associated with the middle bins of the SFS. The y-axis is on a $\log_{10}$ scale to show better detail across lower frequencies. The x-axis is on a $\log_e$ scale.

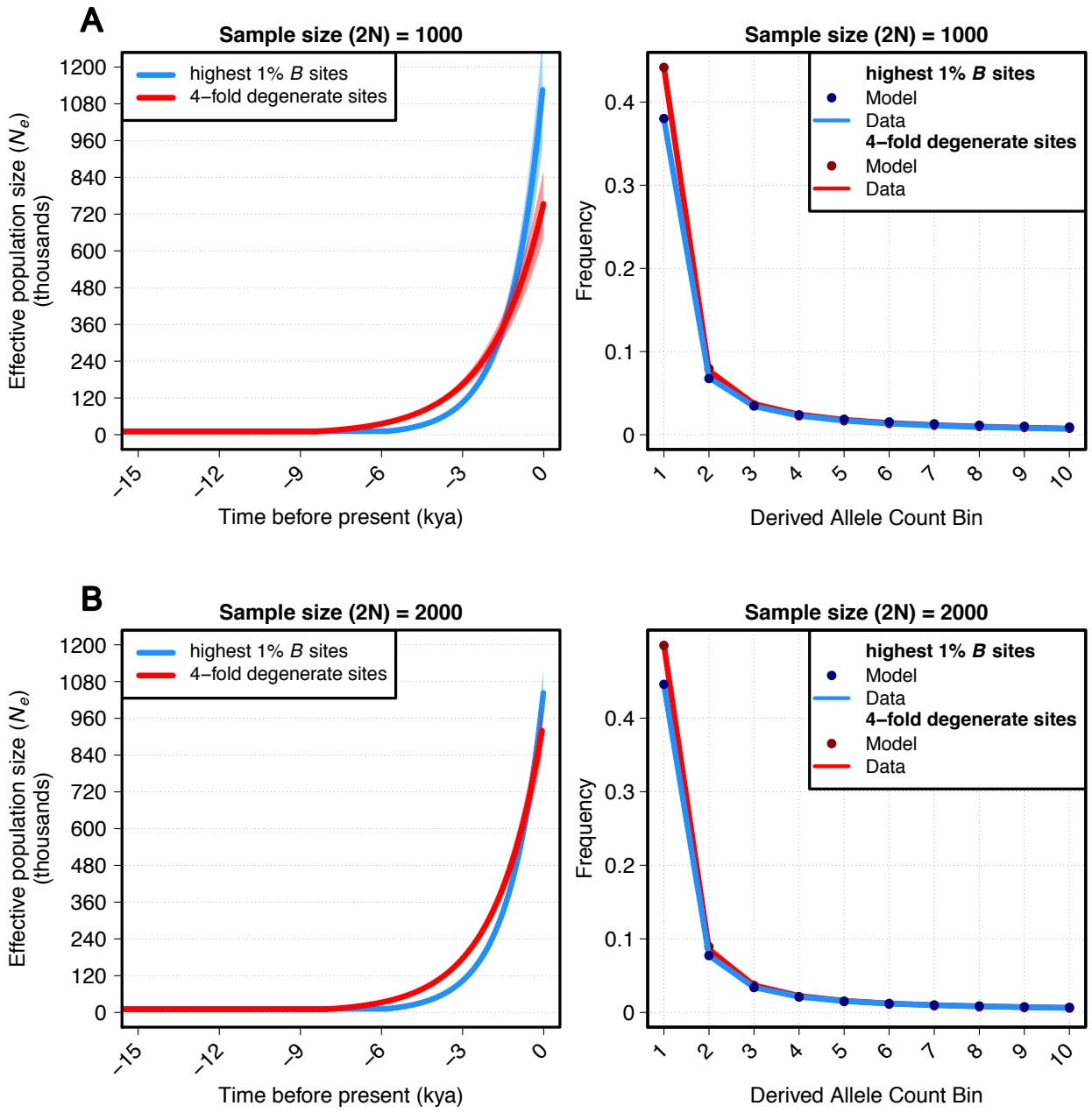


Figure C.3. Results from performing demographic inference using 2N=1000 and 2N=2000 samples.
Results from performing demographic inference on an exponential growth model (see Table C.1 for demographic parameter values) using data from sites in the highest 1% *B* bin (weak BGS) or fourfold degenerate sites for 2N=1000 samples (A) and 2N=2000 samples (B). Shaded envelopes represent 95% confidence intervals. In the plots on the right side of A) and B), the observed site-frequency spectrums from the two types of sites used for inference are shown as solid lines. The resulting fits to the site-frequency spectrum from the fitted demographic models are shown as points.

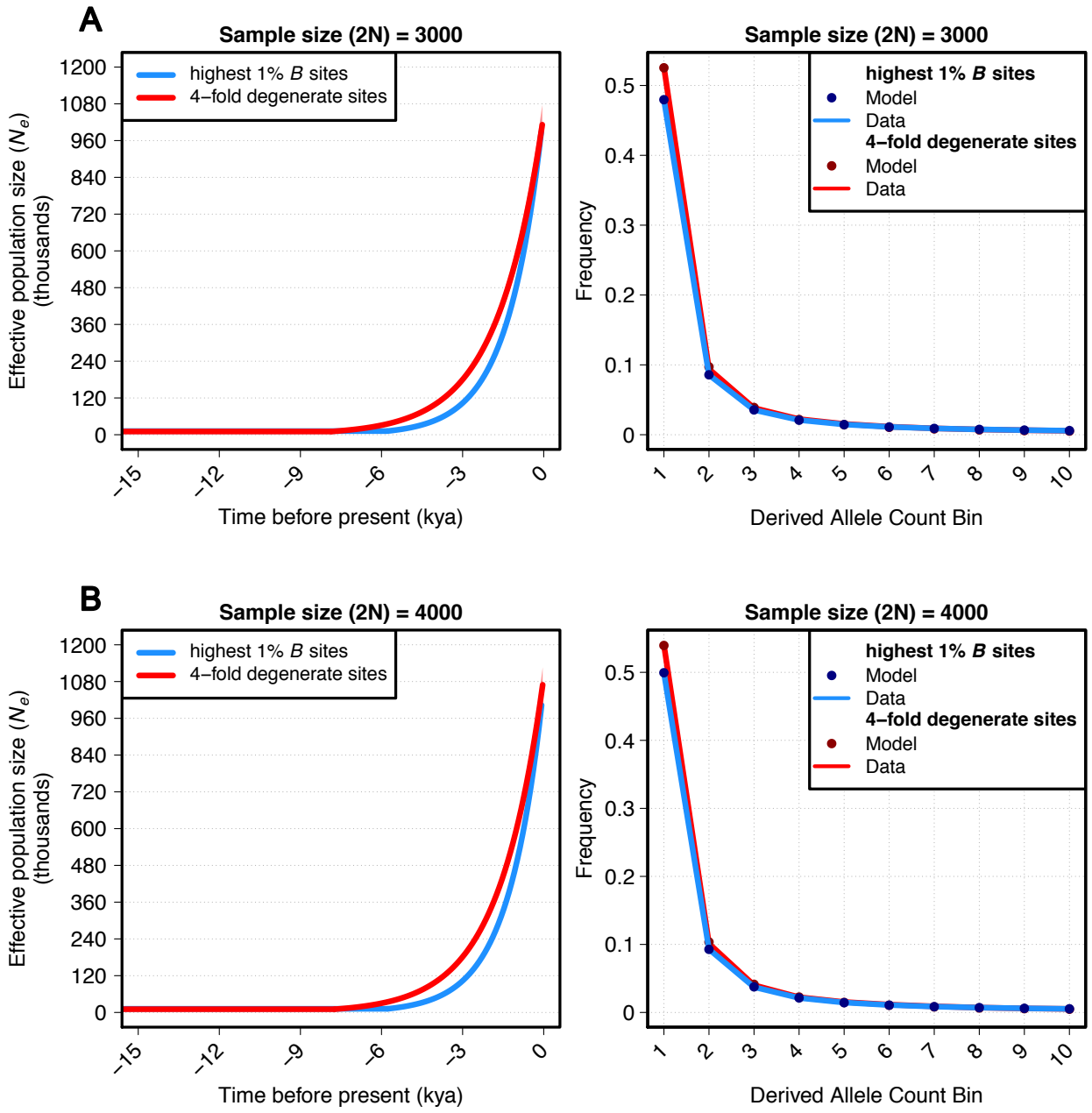


Figure C.4. Results from performing demographic inference using 2N=3000 and 2N=4000 samples.
Results from performing demographic inference on an exponential growth model (see Table C.1 for demographic parameter values) using data from sites in the highest 1% *B* bin (weak BGS) or fourfold degenerate sites for 2N=3000 samples (A) and 2N=4000 samples (B). Shaded envelopes represent 95% confidence intervals. In the plots on the right side of A) and B), the observed site-frequency spectrums from the two types of sites used for inference are shown as solid lines. The resulting fits to the site-frequency spectrum from the fitted demographic models are shown as points.

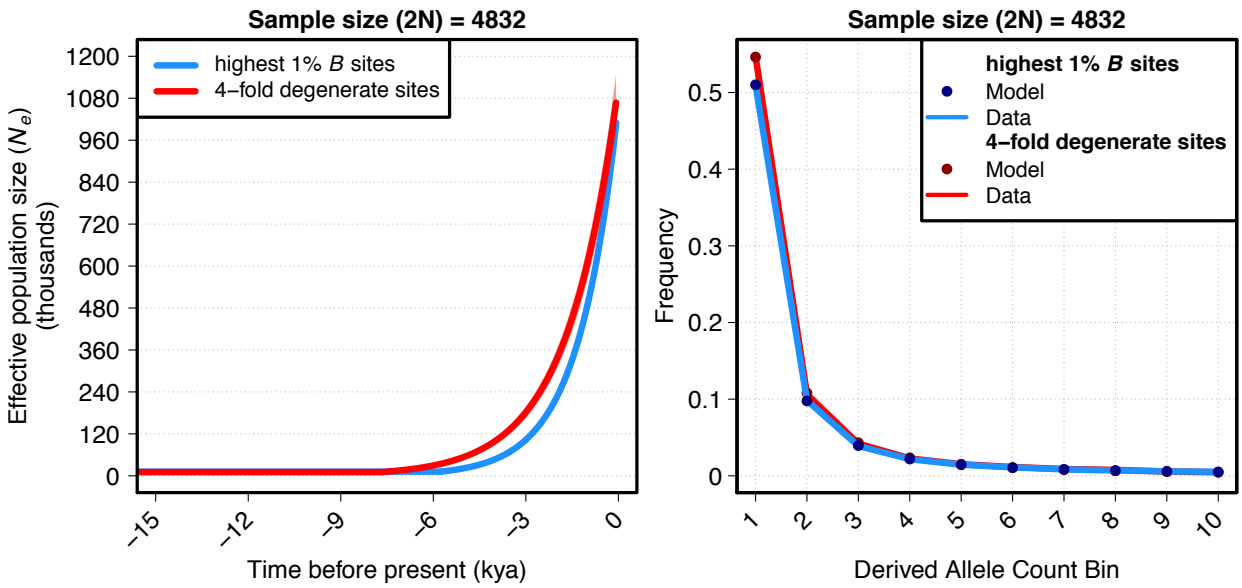


Figure C.5. Results from performing demographic inference using 2N=4832 samples. Results from performing demographic inference on an exponential growth model (see Table C.1 for demographic parameter values) using data from sites in the highest 1% *B* bin (weak BGS) or fourfold degenerate sites for 2N=4832 samples. Shaded envelopes represent 95% confidence intervals. In the plots on the right side of the figure, the observed site-frequency spectrums from the two types of sites used for inference are shown as solid lines. The resulting fits to the site-frequency spectrum from the fitted demographic models are shown as points.

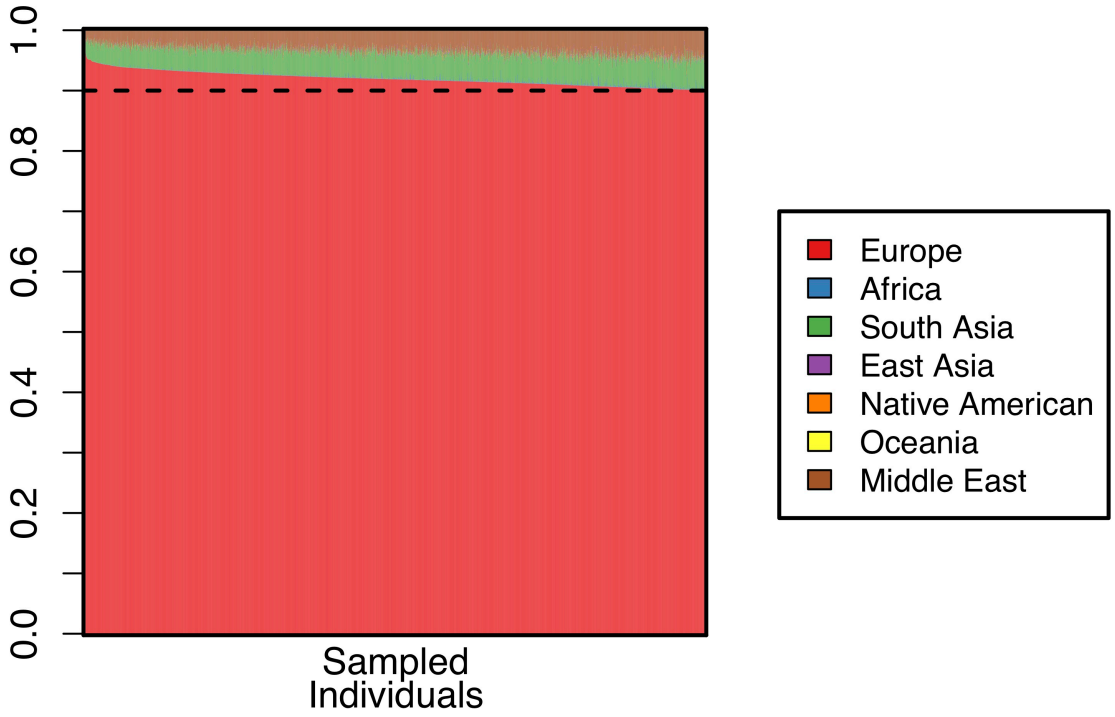


Figure C.6. Global ancestries from RFMix for 2,416 TOPMed samples.
The 7 HGDP super-populations used as references for RFMix are shown in the legend.

Table C.1

sample size (2N)	data	*NEur*$_0$ (95% CI)	*NEur* (95% CI)	*TEur* (95% CI)	rEur*100 (95% CI)
1000	highest 1% *B*	11,689 (11,586-11,792)	1,163,554 (971,534-1,355,573)	5,741 (5,566-5,916)	2.02 (1.94-2.09)
	4-fold degenerate	10,059 (9,902-10,217)	755,334 (644,947-865,722)	8,505 (8,131-8,880)	1.28 (1.23-1.32)
2000	highest 1% *B*	11,762 (11,658-11,866)	1,053,407 (974,047-1,132,768)	5,813 (5,682-5,943)	1.95 (1.92-1.98)
	4-fold degenerate	10,313 (10,159-10,466)	949,761 (867,911-1,031,612)	8,058 (7,802-8,314)	1.41 (1.38-1.44)
3000	highest 1% *B*	11,835 (11,735-11,936)	1,060,540 (1,003,916-1,117,164)	5,793 (5,687-5,899)	1.96 (1.93-1.98)
	4-fold degenerate	10,487 (10,334-10,639)	1,043,842 (976,228-1,111,456)	7,853 (7,634-8,072)	1.48 (1.45-1.5)
4000	highest 1% *B*	11,895 (11,790-12,000)	1,065,695 (1,022,542-1,108,848)	5,778 (5,676-5,879)	1.96 (1.95-1.98)
	4-fold degenerate	10,619 (10,462-10,775)	1,094,072 (1,035,292-1,152,852)	7,738 (7,524-7,952)	1.51 (1.49-1.53)
4832	highest 1% *B*	11,937 (11,716-12,158)	1,066,503 (1,029,592-1,103,414)	5,770 (5,600-5,940)	1.97 (1.95-1.98)
	4-fold degenerate	10,709 (9,469-11,948)	1,118,240 (1,033,993-1,202,487)	7,677 (6,416-8,938)	1.53 (1.5-1.55)

Table C.1. Resulting fitted parameters from performing demographic inference with 4-fold degenerate sites or highest 1% *B* sites (physical units).

Parameters include the starting population size before growth ($NEur_0$), the ending population size after growth (*NEur*), and the time span over which exponential growth occurred (*TEur*). The rate of growth is shown in the last column (rEur). Population size is in given in units of N_e and time is given in years assuming a generation time of 25 years. See Materials and methods for how genetic units (Table C.2) are converted to physical units.

Table C.2

sample size (2N)	data	*NEur*$_0$	*NEur* (95% CI)	*TEur* (95% CI)	theta (95% CI)
1000	highest 1% *B*	1	99.54 (83.12-115.97)	0.0098 (0.0095-0.0101)	8519.94 (8444.87-8595.02)
	4-fold degenerate	1	75.09 (64.11-86.06)	0.0169 (0.0162-0.0177)	3151.8 (3102.35-3201.25)
2000	highest 1% *B*	1	89.56 (82.81-96.31)	0.0099 (0.0097-0.0101)	8573.51 (8497.61-8649.41)
	4-fold degenerate	1	92.1 (84.16-100.03)	0.0156 (0.0151-0.0161)	3231.12 (3183.0-3279.24)
3000	highest 1% *B*	1	89.61 (84.82-94.39)	0.0098 (0.0096-0.01)	8626.78 (8553.38-8700.18)
	4-fold degenerate	1	99.54 (93.09-105.99)	0.015 (0.0146-0.0154)	3285.62 (3237.95-3333.29)
4000	highest 1% *B*	1	89.59 (85.96-93.22)	0.0097 (0.0095-0.0099)	8670.27 (8593.92-8746.63)
	4-fold degenerate	1	103.03 (97.5-108.57)	0.0146 (0.0142-0.015)	3327.06 (3278.0-3376.11)
4832	highest 1% *B*	1	89.34 86.25-92.44)	0.0097 (0.0094-0.01)	8700.85 (8540.01-8861.68)
	4-fold degenerate	1	104.42 (96.55-112.29)	0.0143 (0.012-0.0167)	3355.31 (2966.94-3743.67)

Table C.2. Resulting fitted parameters from performing demographic inference with 4-fold degenerate sites or highest 1% *B* sites prior to scaling (genetic units).
Parameters include the starting population size before growth ($NEur_0$), the ending population size after growth (*NEur*), and the time span over which exponential growth occurred (*TEur*). Since these parameters are not scaled using theta to generate units of N_e, parameter values are given relative to $NEur_0$ (which is always 1). The last column gives the inferred theta from performing inference.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Raul Jones
Author Signature

12-21-18
Date