Population Genomic Insights Into Recent Human Evolutionary History

Leslie S. Emery

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2014

Reading Committee:
Joshua M. Akey, Chair
Joseph Felsenstein
Jay Shendure

Program Authorized to Offer Degree:
Genome Sciences

© Copyright 2014

Leslie S. Emery

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.
http://creativecommons.org/licenses/by-nc-sa/4.0/

University of Washington

Abstract

Population Genomic Insights into Recent Human Evolutionary History

Leslie S. Emery

Chair of the Supervisory Committee:
Associate Professor Joshua M. Akey
Department of Genome Sciences

The advent of large-scale population genomic datasets has enabled detailed inferences regarding human evolutionary history. Demographic changes and positive selection have left their marks on the genome and we can now begin to decipher them. In this dissertation, I present the work I have completed on the topic of human population genomic inference. In chapter 1, I begin by reviewing the importance of human genetic variation and the factors that influence it, focusing on the effects of demographic changes and positive selection. Chapter 2 describes an analysis of genetic ancestry in a worldwide sample of human populations. I show that mitochondrial lineage tests overlook large amounts of variation in genetic ancestry. In chapter 3, I focus on inferences regarding the effective sex ratio in the recent evolutionary past. I present a reanalysis of SNP and resequencing data that resolves a set of conflicting results from previous studies. Using coalescent simulations, I present a model of a recent male bias in effective population size, coupled with an earlier

female bias, which is consistent with existing genetic variation on the X chromosome and the autosomes. In chapter 4, I present a comprehensive study of the performance of a battery of neutrality test statistics under a wide range of realistic models of positive selection in recent human evolution. I demonstrate that existing tests perform better than expected for detecting the signatures of a soft sweep from standing variation. Then, I develop a genome-wide approach, the Cumulative Selection Score (CSS), for combining the signals from multiple neutrality test statistics to detect the signatures of positive selection with greater accuracy. By implementing this approach in genomic variation data for chromosome 2, I show that the CSS can be applied to whole-genome datasets. I conclude in chapter 5 by discussing the potential of population genomic inferences and the future of the field.

Acknowledgements

I would like to begin by saying thank you to the many people who have contributed the encouragement and support I needed to be able to complete my doctorate.

First, I would like to thank my professors from Alfred University, Jean Cardinale, Gordon Godshalk, Cheryld Emmons, and Robert Myers, for their enthusiasm and encouragement. Their support was important in my decision to pursue a doctorate, and their training has been instrumental to my success. Dr. Godshalk in particular provided words of encouragement that have stuck with me and been a comfort in the most trying times.

I am grateful also to my classmates and fellow graduate students in Genome Sciences, who have helped to make my experience in graduate school enjoyable as well as productive. They have been valuable colleagues and collaborators, and also dear friends.

I have also benefited from the expertise and enthusiasm of my thesis committee: Joe Felsenstein, Jay Shendure, Celeste Berg, and Toby Bradshaw. It has been a privilege to share my work with them and my research has benefited greatly from their feedback and suggestions. In particular, Joe's enthusiasm for one of my projects resulted in a valuable collaboration and a published paper.

I would like to thank Josh Akey for providing a supportive lab environment and fostering the camaraderie that has made his lab what it is. Thank you also to the members of the Akey lab, past and present. Tom Nicholas, Dan Skelly, Caitlin

Connelly, Marnie Johansson, Ben Vernot, Jenny Andrie, Jacob Tennessen, and Tim O'Connor have especially been influential in sharing their ideas, their friendship, and their laughter.

To Brian McNally, thank you for your unwavering support and confidence, and especially for your patience and understanding during the most difficult periods of my graduate work.

I would finally like to thank my parents for being a constant source of encouragement and support. Without them I would not have accomplished any of the work I am about to present.

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 Factors that influence human genetic variation

The amount and distribution of genetic variation in human populations is a direct result of our evolutionary past. Each individual carries with them a genetic record of this evolutionary past that rivals archaeological and paleontological evidence for its objectivity and its detail. The two strongest evolutionary factors that influence extant human genetic variation are demography and selection. These two factors have interacted to leave us with a genetic legacy that affects the distribution of genetic ancestry, the prevalence of genetic diseases, and the wide variety of human phenotypes across the earth.

1.1.1 Effects of demographic history on human variation

Human genetic variation can only be properly understood in the context of human demographic history. Anatomically modern humans arose in Africa and experienced at least one major population expansion before a subset of individuals dispersed to new regions between 50,000 - 100,000 years ago (Goldstein and Chikhi 2002; Cavalli-Sforza 2007). This group was responsible for the colonization of the Middle East, Europe, Asia, and the Americas in a series of successive colonization events that fits either a serial founder effect (Ramachandran *et al.* 2005; Deshpande *et al.* 2009; Barbujani and Colonna 2010) or an isolation-by-distance model (Handley *et al.* 2007). Regardless of the appropriate model, each subsequent colonization was associated with a reduction in population size (a bottleneck) (Cavalli-Sforza 2007). As humans encountered new environments, geographically restricted adaptations

occurred. Population sizes expanded drastically with the development of agriculture at different times in different parts of the world, accompanied by adaptations to new diets and lifestyles (Cavalli-Sforza 2007). During this colonization phase, migration occurred at low levels between geographically close populations. As means of transportation have improved over the past centuries, migration has increased significantly, resulting in high levels of admixture between formerly separated populations and far-flung genetically related groups.

This demographic history has significant implications for interpreting human genetic variation. The small human ancestral effective population size means that the human species contains relatively little genetic diversity compared to the other great apes (Kaessmann *et al.* 2001). The African origin of humans results in higher levels of genetic variation found within Africans (Vigilant *et al.* 1991; Jorde *et al.* 2000; Tishkoff *et al.* 2009), as well as a negative correlation between measures of diversity and distance from hypothesized points of origin in Africa (Ramachandran *et al.* 2005; Tishkoff *et al.* 2009). There is some evidence that recent population growth and the peopling of the world has led to very rapid adaptation and an accelerated rate of evolution (Hawks *et al.* 2007). Importantly, human demographic history also determines the amount of variation and specific variants that are available for selection to act upon.

The relatively recent divergence of human populations, and high migration rates between them, mean that most of the variation seen in humans is found within all populations, and not between them (Lewontin 1995; Barbujani *et al.* 1997; Jorde *et al.* 2000). Because of the nature of human dispersal, genetic variation is distributed in a

clinal manner, rather than in distinct pools of alleles for each population (Handley *et al.* 2007).

Some cultural and biological factors result in demographic forces that affect males and females differently and this can leave unique patterns of variation in the human genome, particularly on the sex chromosomes and mtDNA (Wood *et al.* 2005). Sex biases in dispersal affect sex-specific rates of gene flow between populations (Hedrick 2007). Differences in sex-specific effective population sizes affect rates of genetic drift and leave corresponding patterns on the sex chromosomes. Notably, different mating systems (polygamy, polygyny, polyandry, *etc.*) affect sex-specific effective population sizes in an interesting example of human cultural practices affecting human variation. These effects on the variation of the sex chromosomes can be detected and used to determine sex-specific population sizes in the recent human past.

1.1.2 Effects of selection on genetic variation

Some forms of selection will reduce the amount of nearby neutral variation, while other forms of selection will increase it. The variation that remains after the action of selection is determined by the interaction of selection with recombination. Recombination links the fate of nearby drifting neutral variants to that of a selected allele (Maynard Smith and Haigh 1974). The effect of selection on these linked neutral variants can be described in terms of two related forces: hitch-hiking near advantageous alleles and background selection near deleterious alleles.

1.1.2.1 Hitch-hiking effects of selection

When a mutation creates an advantageous allele *A*, the surrounding genomic region contains linked neutral variants which comprise *A*'s haplotype. As natural selection acts to increase the frequency of *A*, the haplotype on which it is found will rise in frequency as well. In this sense, the linked neutral variants are "hitch-hiking" along with the selected allele in a phenomenon that is also called a selective sweep (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989; Barton 1998). When *A* and its haplotype become fixed in the population, the sweep is complete and the only variation left in the region surrounding *A* is the set of linked neutral variants in the hitch-hiking haplotype. The end result is a decrease in overall genetic variation (Barton 1998) and an excess of rare variants, which skews the region's site frequency spectrum (SFS) towards lower frequency variants (Kaplan *et al.* 1989; Stephan 2010). The extent of these effects is proportional to the amount of recombination in the region, the strength of selection on *A*, and the temporal stage of the selective sweep (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989; Barton 1998; Stephan 2010).

1.1.2.2 Background selection

If instead we have a newly created deleterious allele, *a,* the haplotype on which *a* is found will decrease in frequency as natural selection acts to remove it from the population. As a consequence, the neutral variants linked to *a* are also removed in a phenomenon known as background selection (Charlesworth *et al.* 1993). Both hitch-hiking and background selection reduce the amount of variation in the human genome and it is remarkably difficult to distinguish the two forces from sequence evidence alone

(McVicker *et al.* 2009; Stephan 2010). Weakly deleterious alleles can have a noticeable effect on variation, especially as the effects of continuous background selection near a functional element accumulate.

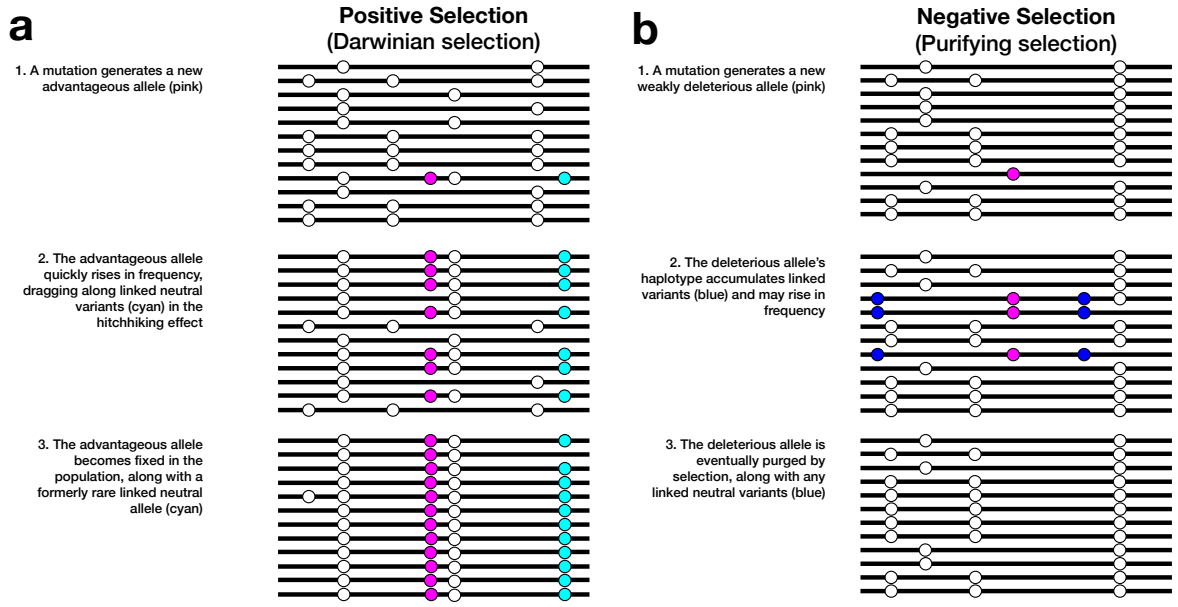


Figure 1.1 Types of selection
Each line is a chromosome in the population; each circle is a variant segregating in the population, colored according to selective role

1.1.2.3 Negative selection

Negative selection is the force responsible for decreasing the frequency of deleterious alleles (Fig. 1b). These deleterious alleles are affected by background selection, and in many cases a functional element is subject to repeated episodes of negative selection (Charlesworth *et al.* 1993). Because negative selection is responsible for purging deleterious alleles, it is also known as purifying selection (Charlesworth *et al.* 1993; McVicker *et al.* 2009; Lohmueller *et al.* 2011). The effects of negative selection are difficult to estimate for a single locus, but a pattern of reduced variation near

evolutionarily conserved elements suggests its genome-wide prevalence (Reed *et al.* 2005; McVicker *et al.* 2009; Lohmueller *et al.* 2011).

1.1.2.4 Positive selection

Positive selection is responsible for increasing the frequency of advantageous alleles in a population (Fig. 1a). The hitch-hiking effect is the major force acting on variation near advantageous alleles, and in fact it was first conceived as a way to describe the effects of positive selection (Maynard Smith and Haigh 1974). Much debate has centered on the relative contribution of purifying and positive selection (Stephan 2010). Current evidence suggests that weak negative selection is widespread, whereas positive selection has affected a smaller portion of the genome more strongly. Thus, while negative selection has affected a larger proportion of the genome, positive selection has had a stronger effect at the modest number of regions it has influenced (Lohmueller *et al.* 2011).

Because advantageous alleles provide information about specific human adaptations, positive selection has been widely studied. Adaptation of specific human populations to their respective environments and subsistence strategies is responsible for most of these ongoing, or recently completed, selective sweeps.

1.2 The importance of human genetic variation

1.2.1 Human variation and population differentiation

Past understanding of population differentiation was largely driven by the cultural construction of race, which is of questionable biological meaning (Lewontin 1995; Barbujani *et al.* 1997; Bamshad *et al.* 2004; Vitti *et al.* 2012). A thorough

understanding of the true genetic differentiation between human populations will provide insight into human history and evolution, as well as enabling ancestry-aware medical treatment and diagnosis. The amount of genetic differentiation between two populations is a product of shared ancestry, migration, genetic drift, geographic separation, and population-specific selective pressures. Recent widespread admixture between human populations also complicates the picture. Most targets of selection that have been identified to date are under selection in a subset of populations and this contributes to population differentiation. It is clear that selection has played a role in human population differentiation (Barreiro *et al.* 2008), but the extent of that role is thus far uncertain.

1.2.2 Human variation and disease

The present day distribution of disease burden is a direct function of our evolutionary past; therefore learning about the evolutionary history of human alleles can help us to understand why some populations are more susceptible to a particular disease and why common disease alleles persist. One goal of evolutionary medicine is to understand both proximate (mechanistic) and ultimate (evolutionary) causes of genetic disease (Gluckman *et al.* 2011). Knowledge of selection's effects on variation is also important for the selection of markers in disease gene mapping and association studies (Tishkoff and Verrelli 2003). Furthermore, positive selection can have important effects on disease susceptibility.

In particular, positive selection can result in adaptions to specific diseases, for example the Δ32 allele of the *CCR5* gene confers resistance to HIV-1 infection (Libert *et al.* 1998). *CCR5* encodes the CC chemokine receptor 5, a G-protein-coupled

chemokine receptor that is necessary for entry of HIV into the cell. Other alleles at *CCR5* have also been shown to affect HIV susceptibility (Bamshad *et al.* 2002). Interestingly, *CCR5* shows evidence for positive selection older than the HIV/AIDS epidemic and proposed historical explanations have included bubonic plague and smallpox (Galvani and Slatkin 2003). These historical hypotheses would explain the present-day frequency of HIV resistance alleles, which are most prevalent within Europe, although more recent studies have suggested that CCR5- Δ32 was subject to even more ancient selection (Sabeti *et al.* 2005; Hedrick and Verrelli 2006). Positive selection can also have negative side effects, which is why it is important to understand their selective advantage as well. Interestingly, current evidence suggests that positively selected alleles are more likely to increase risk for complex diseases than to decrease that risk (Barreiro and Quintana-Murci 2010) and are also found in disease-related genes more often than expected (Barreiro *et al.* 2008). These connections to human disease are among the reasons that it is important to identify the signatures that positive selection has left in the human genome.

1.3 Objectives

While I have taken on a variety of projects during my thesis work, they are united by an interest in characterizing the marks that human evolutionary history has left in the genome. Specifically, I sought to explore the evolutionary insights that can be gained by analyzing genomic variation data in human populations. My research objectives were as follows:

1. To characterize the extent to which mitochondrial lineage tests capture genetic ancestry information. I compared ancestry inferences from

mtDNA lineage tests with those obtained from SNP genotypes and found that the two are often decoupled.

2. To examine the genetic signatures of sex biased demographic forces in order to draw conclusions about past human mating systems. I found that genetic signatures of sex-biased demography are consistent with higher female effective population sizes in early human evolution, followed by higher male effective population sizes later on.

3. To comprehensively and accurately identify adaptive alleles across the entire human genome. To begin to achieve this goal, I examined the performance of neutrality test statistics under realistic models of selection in human populations. I also developed and tested a method for combining the signals from multiple test statistics in a genome-wide scan.

2. Continental ancestry varies widely within most human mitochondrial haplogroups

This chapter is based on the following manuscript, currently in preparation for submission at *The American Journal of Human Genetics*:

Leslie S. Emery*, Kevin M. Magnaye*, Abigail W. Bigham, Joshua M. Akey, and Michael J. Bamshad. Continental ancestry varies widely within most human mitochondrial haplogroups in a worldwide population sample.

* denotes co-first authorship.

Kevin Magnaye genotyped mitochondrial SNPs, classified samples into mtDNA haplogroups, and contributed to data analysis and interpretation. Abigail Bigham sequenced mtDNA samples and oversaw mtDNA SNP genotyping and haplogroup classification. Kevin Magnaye and I wrote the manuscript, with revisions from Abigail Bigham, Joshua Akey, and Michael Bamshad.

2.1 Summary

Human mitochondrial DNA can be organized into unique haplotypes and further classified into haplogroups according to a phylogeny or network. A standardized haplogroup classification is used for forensics, mitochondrial disease studies, characterizing migrations and recent human demography, and direct-to-consumer (DTC) ancestry testing. To determine the extent to which mtDNA haplogroups capture ancestry information, we undertook a study of variation in continental ancestry within mtDNA haplogroups in ~940 worldwide samples from the Human Genome Diversity Panel. We compared a sample's mtDNA haplogroup classification to its continental ancestry proportions (determined from ~650,000 autosomal SNPs) and found that continental ancestry varied widely from person to person within each haplogroup. MtDNA haplogroups were often associated with their primary continental ancestry components; however, most secondary continental ancestry components were not. The association between continental ancestry and mtDNA haplogroup was even weaker within a set of recently admixed populations from the 1,000 Genomes Project. Furthermore, continental ancestry proportions could not be reliably used to predict a sample's mtDNA haplogroup classification. These results have important implications for the continued usefulness of mtDNA haplogroups as indicators of ancestry and demographic history.

2.2 Background

The high level of polymorphism and overall abundance of mitochondrial DNA (mtDNA) has made it a useful tool for studying human demographic history (Underhill and Kivisild 2007). Early studies classified branches of the human mtDNA tree into

groups of phylogenetically related haplotypes (Cavalli-Sforza and Feldman 2003; Underhill and Kivisild 2007), defined by lineage-specific polymorphisms in continental-scale populations such as Native Americans (Torroni, Schurr, *et al.* 1993; Torroni, Sukernik, *et al.* 1993), Africans (Chen *et al.* 1995), and Europeans (Torroni *et al.* 1996; Herrnstadt *et al.* 2002). Populations with shared ancestry and/or living in geographic proximity displayed similar haplogroup frequencies (Cavalli-Sforza and Feldman 2003; Underhill and Kivisild 2007). These standardized haplogroups and subhaplogroups (van Oven and Kayser 2009) enabled detailed studies of migration history (*e. g.* Kivisild *et al.* 2003; Kong *et al.* 2003) and detection of sex-biased demography (*e. g.* Wen *et al.* 2004; Wood *et al.* 2005; Boattini *et al.* 2013).

For the past decade or so, genealogical testing companies have capitalized on the results of genetic ancestry research by adapting mtDNA and non-recombining Y chromosome (NRY) haplogroup analyses to provide direct-to-consumer (DTC) ancestry tests. The association of geographical region to mtDNA and NRY haplogroups provides the basis for using these haplogroups to infer an individual's genetic ancestry, but such lineage-based analyses overlook the contribution of the vast majority of an individual's ancestors to their genome (Shriver and Kittles 2004). Moreover, DTC ancestry tests have proven controversial because they use proprietary methods that lack transparency, present conflicts between cultural and scientific conceptions of ancestry, and lack federal regulation (Lee *et al.* 1993; Shriver and Kittles 2004; Bolnick *et al.* 2007; Frudakis 2008; Sarata 2008; The American Society of Human Genetics 2008; Via *et al.* 2009; Royal *et al.* 2010; Wagner *et al.* 2012).

The major alternative to lineage-based ancestry tests is model-based ancestry inference using genome-wide SNP genotype data (Pritchard *et al.* 2000; Tang *et al.* 2005; Alexander *et al.* 2009). Another alternative is to use Ancestry Informative Markers (AIMs), which are specific autosomal SNPs with documented allele frequency differences between continental groups (Shriver *et al.* 1997). Both of these approaches result in per-individual estimates of the proportion of ancestry from one or more reference populations that are assumed to correspond to specific ancestral populations. Several recent population-specific studies have assessed the relationship between ancestry inferences from mtDNA haplogroups and autosomal ancestry, and found frequent discrepancies, particularly in recently admixed populations (Salas *et al.* 2008; Watkins *et al.* 2012; Cardena *et al.* 2013; Poetsch *et al.* 2013). These results highlighted the limitations of lineage-based ancestry tests in general, and mtDNA ancestry tests in particular.

In 2008 and again in 2012, The American Society of Human Genetics (ASHG) acknowledged the prominence of commercial ancestry testing and provided a series of recommendations for academic scientists and DTC ancestry testing companies (The American Society of Human Genetics 2008; Royal *et al.* 2010). These recommendations expressed critical concern for the lack of information about the accuracy of lineage-based ancestry estimation compared to multi-locus estimation from autosomal markers (The American Society of Human Genetics 2008). In particular, the extent to which ancestry information is captured by mtDNA haplogroups is unknown. In addition, the ASHG ancestry testing statements called for further investigations into how the use of reference panel populations affects ancestry

inference. To begin to address some of these concerns, we quantified the variation in continental ancestry proportions among individuals with the same mtDNA haplogroup in a panel of reference populations. We then compared these results to a large sample of individuals from recently admixed populations. We will conclude by discussing the implications of our findings for the continued use of mtDNA haplogroup-based ancestry inference.

2.3 Materials and Methods

2.3.1 HGDP dataset

We downloaded Illumina 650Y SNP array genotype data for the Human Genome Diversity-CEPH Panel (HGDP) (Cann *et al.* 2002) consisting of 1,043 individuals from 52 worldwide populations (Figure 1A) as previously reported (Li *et al.* 2008). After removing previously-identified relatives and duplicate samples (Rosenberg 2006), and samples with low-quality SNP genotype data, 938 samples remained in our HGDP dataset. Next, we obtained Hypervariable Region 1 (HVR1) sequence data for 891 of these 938 samples from the National Center for Biotechnology Information (NCBI; PopSet accession #189174470). For the 47 individuals without publicly available sequence data, we Sanger sequenced HVR1 using DNA obtained from the Centre d'Etude Polymorphisme Humain (CEPH).

2.3.2 1,000 Genomes dataset

We downloaded 1,000 Genomes Project (1000 Genomes Project Consortium *et al.* 2012) variant call format (.vcf) files for phase 1 low coverage whole genome sequence data (release 20101123). We selected five populations from world regions with high levels of recent admixture: ASW (Americans of African ancestry in SW USA),

CLM (Colombians from Medellin, Colombia), GBR (British in England and Scotland), MXL (Mexican ancestry from Los Angeles USA), and PUR (Puerto Ricans from Puerto Rico). Our total sample consisted of data from 327 people. Using `Tabix` (Li 2011) (v. 0.2.6) and `VCFtools` (Danecek *et al.* 2011) (v. 0.1.10), we removed indels, extracted variant sites with dbSNP reference SNP ID numbers (rs numbers) matching SNPs from the HGDP 650Y SNP data, and converted the data to `PLINK`'s map/ped file format. In `PLINK` (Purcell *et al.* 2007) (v. 1.07), we merged the 1,000 Genomes data with the HGDP data, removing 77 SNPs with unresolvable strand mismatches and 141 SNPs that could not be converted from hg18 to hg19 coordinates. Our final dataset consisted of 646,356 SNPs. We also downloaded .vcf files for all mitochondrial variants in each 1,000 Genomes population.

2.3.3 Mitochondrial haplogroup typing

We obtained DNA for 965 of the HGDP individuals from the Centre d'Etude Polymorphisme Humain for use in mtDNA haplogroup typing. To begin, we first reviewed the literature for SNPs that uniquely identify each of the 23 major mtDNA haplogroups (diagnostic SNPs). Initially, we selected 24 candidate SNPs from Mitomap (Ruiz-Pesini *et al.* 2007) and the Genographic Project (Behar *et al.* 2007) (one haplogroup required two diagnostic SNPs). We checked each of these candidate diagnostic SNPs in the PhyloTree (van Oven and Kayser 2009) comprehensive mtDNA phylogeny to confirm that the SNP was diagnostic. If a candidate diagnostic SNP was not supported by information from PhyloTree, we selected an additional diagnostic SNP based on the PhyloTree phylogeny. Using a total of 28 diagnostic SNPs, we could classify samples into the 23 mitochondrial haplogroups (Figure A.1).

We next used the Genographic Project's nearest-neighbor haplogroup prediction tool to assign a predicted haplogroup to each sample based on its HVR1 sequence. The prediction tool uses a sample's haplotype affinity to HVR1 sequences in the Genographic Project's extensive reference database to predict the sample's haplogroup (Behar *et al.* 2007). Next, we experimentally confirmed these haplogroup predictions by genotyping each HGDP sample (either by Sanger sequencing or restriction digest) for the diagnostic SNP(s) of its predicted haplogroup. Finally, if our initial prediction was incorrect, we genotyped the sample using our 28 diagnostic SNPs, beginning with the SNP diagnostic of ancestral haplogroup L0/L1, and working our way towards the tips of the mtDNA phylogeny.

Using the mitochondrial variant calls from the 1,000 Genomes phase 1 low-coverage genome data, we extracted all of the variants corresponding to our 28 diagnostic SNPs. We used these diagnostic SNP variant calls to assign a mitochondrial haplogroup to samples from all 327 people.

2.3.4 Continental ancestry estimation from autosomal SNPs

We used `ADMIXTURE` (Alexander *et al.* 2009) (v. 1.22) to estimate ancestry proportions from each of seven continental regions in each sample from the HGDP Illumina 650Y genotype data. Because `ADMIXTURE` does not account for linkage disequilibrium (LD), we pruned the genotype markers according to observed correlation coefficients in the data using a threshold of $R^2 \geq 0.1$ and a 50-SNP window advancing by 10 SNPs in `PLINK`. We used this dataset in `ADMIXTURE` with seven ancestral groups ($k = 7$), according to previously established population structure parameters in the HGDP (Li *et al.* 2008). The seven inferred populations correspond to continental

regions: Africa, the Americas, Central/South Asia, East Asia, Europe, the Middle East, and Oceania. Continental ancestry proportions for each sample in each HGDP population are shown in Figure A.6.

To estimate continental ancestry in samples from the 1,000 Genomes dataset, we first selected HGDP pseudo-ancestors, to serve as proxies for the ancestral populations we sought to identify (Tang *et al.* 2005). For each of the seven continental groups, we selected the twenty people with the highest fraction of ancestry from the respective continent, resulting in 140 pseudo-ancestors. By including these proxies for ancestral populations, we ensured that the seven continental ancestry components identified in the 1,000 Genomes data would match those identified in the HGDP dataset. Combining SNP data from the pseudo-ancestors and the 1,000 Genomes populations, we used `PLINK` to prune the SNPs according to the same pruning settings used above and estimated ancestry from this pruned dataset. We estimated ancestry proportions for the pseudo-ancestors twice (once with the HGDP data, and once with the 1,000 Genomes data). Each sample's two sets of estimated ancestry proportions were highly correlated (Pearson's $R^2 > 0.99$, $p < 0.0001$) (Figure A.2). Continental ancestry proportions for each sample in each 1,000 Genomes population are shown in Figure A.7.

2.3.5 Analysis of continental ancestry estimates within each mtDNA haplogroup

We first examined the average continental ancestry proportions within each mtDNA haplogroup (Table A.1). This produced an *h* x 7 matrix of means, where *h* is the number of haplogroups and 7 is the number of continental regions. To determine

whether each average continental ancestry component was significantly higher within a haplogroup than we would expect by chance, we performed a permutation test. For each replicate of the permutation test, we shuffled the haplogroup labels for the sample and recalculated the *h* x 7 matrix of means. Then we used 999 replicates, plus the original data, to calculate *p* values for each of the means in the original matrix.

To examine inter-individual variation in the composition of each individual's continental ancestry proportions within mtDNA haplogroups, we calculated the standard deviation of the ancestry estimates for each continental region within each haplogroup. To measure this variability in more detail, we calculated the mean pairwise Euclidean distance, *d*, within each haplogroup. Specifically, we considered each person $P_i = (p_1, p_2, \ldots p_7)$ to be a point in 7-dimensional space (defined by their ancestry proportions). Then we calculated the 7-D Euclidean distance between each pair of individuals, P_i, Q_j, within the haplogroup. Finally, we averaged these distances for all unique pairs within haplogroups, according to the equation below:

$$d = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sqrt{\sum_{i=1}^{k} (q_i - p_i)^2}}{\binom{n}{2}} \qquad \textbf{Eq. 2.1}$$

To determine how well an mtDNA haplogroup can be linked to a particular continental region, we estimated consistency, the proportion of individuals within a haplogroup whose highest proportion of continental ancestry was the same as the haplogroup's highest average continental ancestry (Table A.2). For example, the highest average continental ancestry within haplogroup L2 in the HGDP dataset is

Africa. The consistency of haplogroup L2 in the HGDP dataset is 0.67 as the highest continental ancestry proportion is Africa in 67% of L2 members.

2.3.6 Multinomial logistic regression model (logit model)

To explore whether there was a significant predictive relationship between a sample's mtDNA haplogroup and its inferred continental ancestry composition, we used a training set of combined HGDP and 1,000 Genomes samples to fit a multinomial logistic regression model. First, we excluded all haplogroups with fewer than ten total samples, or fewer than six samples in either dataset. Ten haplogroups remained, including L0/L1, L2, L3, C, D, A, H, B, T, and U. The training set consisted of a randomly selected third of the samples from each dataset. The remaining two thirds of samples were reserved for the test set.

We first used the `multinom` function from the R package `nnet` (Venables and Ripley 2002) to fit a preliminary logit model, with mtDNA haplogroup as the dependent variable and continental ancestry proportions as independent variables. To prevent collinearity, we excluded one continental region as a variable, and chose to exclude Oceanian ancestry because it was least common within our datasets. Each of the six continental ancestry variables improved the fit of the model to the data (likelihood ratio test (LRT), $p < 0.001$). We observed possible nonlinear patterns in the data, so we tested logarithmic and exponential relationships for each of the continental ancestry variables and included any nonlinear relationships that improved the model fit by a likelihood ratio test ($p < 0.001$). With 6 independent variables there are 57 possible interaction terms that can be included in the model. We used likelihood ratio tests to determine which of these interaction terms contribute significantly to a better fit of the

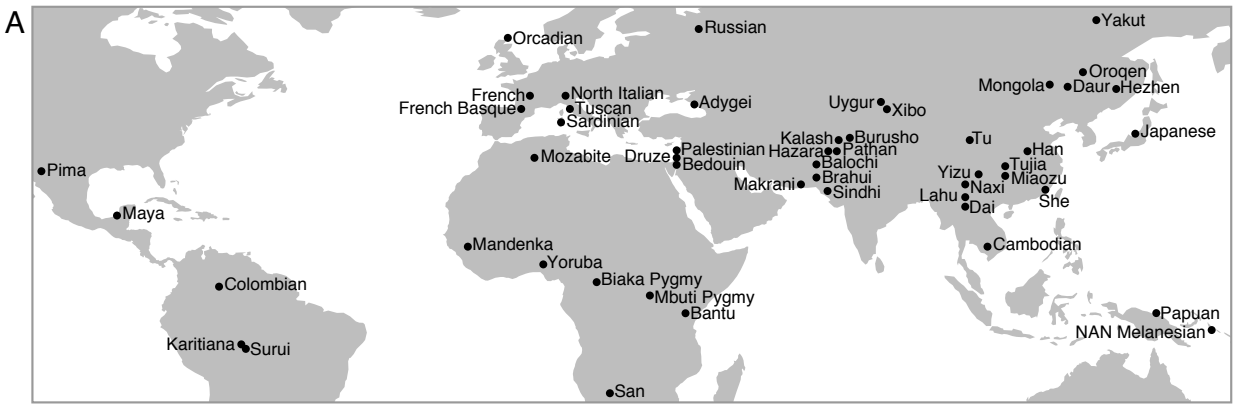
model and identified 39 significant interaction terms to include; the inclusion of these interaction terms significantly improved the fit of the model (LRT, $p < 0.001$).

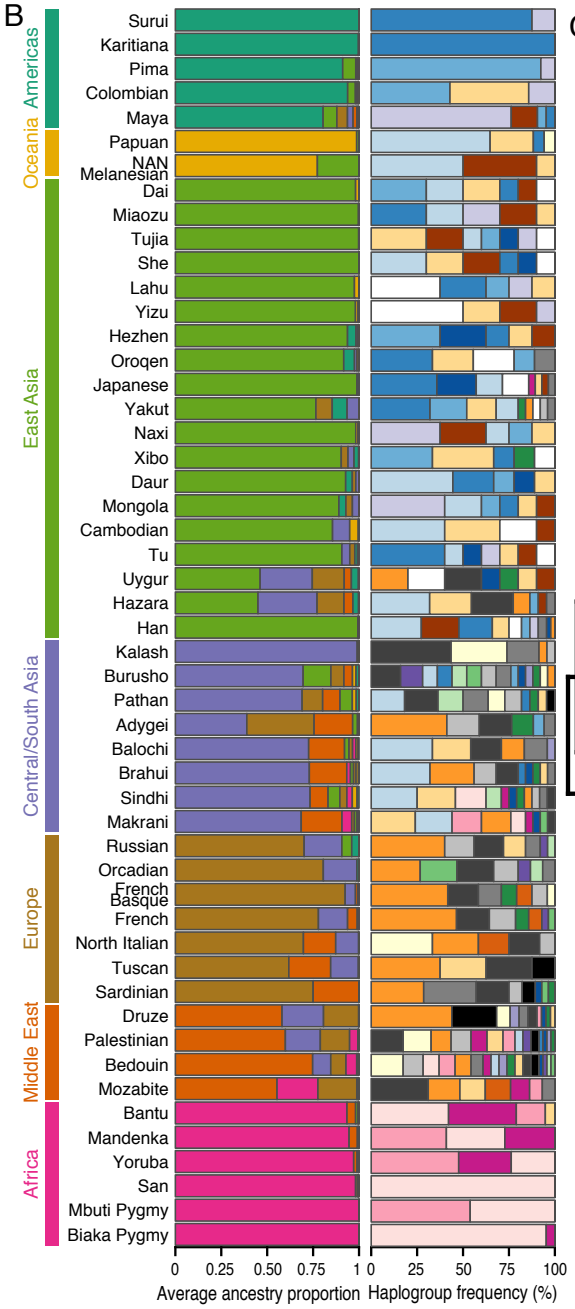
Our final model produced a set of nine logit equations describing the relative odds of belonging to each haplogroup compared to the reference haplogroup, L0/L1. We used these relative odds to determine each sample's classification probability for each of our ten haplogroup categories. Additionally, we used the logit equations to calculate the fitted classification probabilities of each sample in our test set for each haplogroup. For each individual, the haplogroup with the highest classification probability was the haplogroup predicted by the model. We repeated the model fitting procedure with three different randomly selected training subsets from our data. While particular details did change depending on the training set used, the general performance of the model was consistent for multiple training sets.

2.4 Results

2.4.1 Variation in haplogroup frequencies between populations

We explored the relationship between mtDNA haplogroups and continental ancestry by first examining populations of the HGDP, which is often used as a reference panel in human genetic studies. For each of the samples in our HGDP dataset, we assigned an mtDNA haplogroup and the estimated proportion of ancestry from each of the seven continental regions (sub-Saharan Africa, the Middle East, Europe, Central/South Asia, East Asia, Oceania, and the Americas). We then averaged continental ancestry proportions among samples in each population (Figure 2.1B, left column). Additionally, we tabulated the frequency of each mtDNA haplogroup in each population (Figure 2.1B, right column).


A
Orcadian
Russian
Yakut
Oroqen
Mongola
Daur
Hezhen
French
French Basque
North Italian
Tuscan
Sardinian
Adygei
Uygur
Xibo
Japanese
Kalash
Burusho
Tu
Han
Tujia
Miaozu
Yizu
Naxi
She
Lahu
Dai
Pathan
Hazara
Palestinian
Bedouin
Balochi
Brahui
Sindhi
Makrani
Druze
Mozabite
Pima
Maya
Mandenka
Yoruba
Biaka Pygmy
Mbuti Pygmy
Bantu
Cambodian
Colombian
Karitiana
Surui
Papuan
NAN Melanesian
San


B
Americas
Oceania
East Asia
Central/South Asia
Europe
Middle East
Africa
Surui
Karitiana
Pima
Colombian
Maya
Papuan
NAN Melanesian
Dai
Miaozu
Tujia
She
Lahu
Yizu
Hezhen
Oroqen
Japanese
Yakut
Naxi
Xibo
Daur
Mongola
Cambodian
Tu
Uygur
Hazara
Han
Kalash
Burusho
Pathan
Adygei
Balochi
Brahui
Sindhi
Makrani
Russian
Orcadian
French Basque
French
North Italian
Tuscan
Sardinian
Druze
Palestinian
Bedouin
Mozabite
Bantu
Mandenka
Yoruba
San
Mbuti Pygmy
Biaka Pygmy
0
0.25
0.50
0.75
1
Average ancestry proportion
0
25
50
75
100
Haplogroup frequency (%)

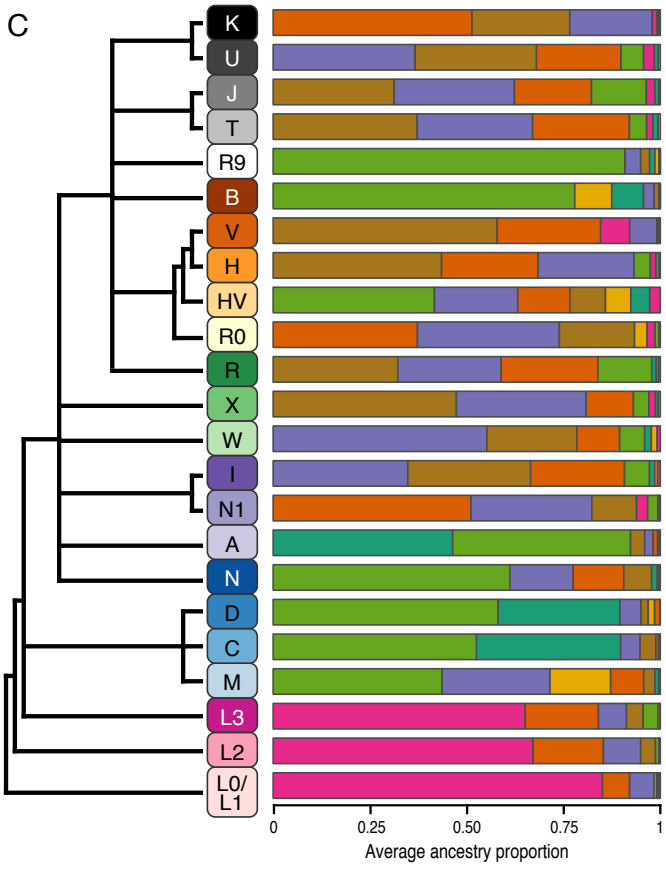

C
K
U
J
T
R9
B
V
H
HV
R0
R
X
W
I
N1
A
N
D
C
M
L3
L2
L0/
L1
0
0.25
0.50
0.75
1
Average ancestry proportion

Figure 2.1 Population locations, mtDNA haplogroup frequencies, and average ancestry proportions in the HGDP data.
A) World map showing sample locations (points) for each of the populations included in the HGDP dataset (labels). B) Left column: Barplots of continental ancestry proportions averaged within each HGDP population. Barplots are colored by continental region (labeled by colored bars on left) and sorted by continental ancestry. Right column: Barplots of haplogroup frequencies within each population, colored by haplogroup (labeled by the haplogroup tree in panel C). C) Barplots of continental ancestry proportions averaged within each mtDNA haplogroup within the HGDP dataset. Barplots are colored by continental region (labeled by colored bars on the left of panel A). The unscaled phylogeny at left shows the relationships between the mtDNA haplogroups (from Behar *et al.* 2007).

We observed each mtDNA haplogroup in several populations, and most populations contained many different haplogroups, with a median of 6 haplogroups per population (Figure 2.1B). Populations from sub-Saharan Africa (*e. g.* Biaka Pygmy, Bantu, Yoruba) consisted almost exclusively of samples assigned to mtDNA haplogroup L0/L1, L2, or L3. Haplogroup H appeared at high frequencies in populations from Europe and the Middle East. Haplogroups M, C, D, N, and A were most frequent in populations from East Asia and the Americas. Haplogroup M was frequent in populations of mostly Central/South Asian, East Asian, or Oceanian ancestry. Despite these observations, there was marked heterogeneity in the overall relationship between a population's geographical origin and its haplogroup composition (Figure 2.1B).

Each of the five 1,000 Genomes populations (ASW, CLM, GBR, MXL, and PUR) contained individuals from 8 different haplogroups, which is significantly higher than the HGDP's median value of 6 haplogroups per population (1-tailed t test, $p < 0.001$). This included GBR, which consisted almost entirely of European continental ancestry. The high frequency of haplogroup H in GBR and of haplogroup L3 in ASW was

consistent with the coincidence of these haplogroups with European ancestry and African ancestry, respectively, in the HGDP populations. Although haplogroup A was not observed in any European HGDP populations, and in fact was most frequent in the Maya (mainly ancestry from the Americas), A was the most common haplogroup in the PUR, MXL, and CLM populations. These observations suggest that the relationship between mtDNA haplogroups and geographic origin has broken down in these recently admixed populations.

2.4.2 Continental ancestry within haplogroups

To examine the continental ancestry of each mtDNA haplogroup, we averaged the individual ancestry proportions among samples within each haplogroup in populations from the HGDP (Figure 2.1C) and the 1,000 Genomes Project (Figure 2.3B). The maximum proportion of ancestry for any single continental region for each haplogroup was, on average, higher in the 1,000 Genomes (1-tailed t test, $p < 0.01$). This may be attributed to higher diversity of continental ancestry proportions in the HGDP populations than in the 1,000 Genomes population, and to the high proportion of continental ancestry from Europe within the 1,000 Genomes populations.

Next, we performed a permutation test to determine whether, in any mtDNA haplogroup, one or more of the seven continental ancestry proportions was higher than expected by chance. One or two continental ancestry components per haplogroup were significantly higher than expected by chance (Figure A.3). These results suggest that some haplogroups are associated with a higher average ancestry proportion from a continental region than expected by chance, and affirm the *ad hoc* visual relationships suggested by the co-occurrence of high average continental ancestry and

high haplogroup frequency within the HGDP populations (Figure 2.1B). The 1,000 Genomes populations had very high proportions of European ancestry due to recent admixture, and this decreased our ability to detect associations between continental ancestry regions and mtDNA haplogroups using a permutation test (Figure A.3). Many of the associations we observed within the HGDP populations were also present in the other dataset, but fewer associations were significant in the 1,000 Genomes populations.

2.4.3 Heterogeneity of individual continental ancestry proportions within mtDNA haplogroups

How applicable are the haplogroup-level results presented above to any one individual? To answer this question we examined the inter-individual variation in inferred continental ancestry composition within mtDNA haplogroups. Individual continental ancestry proportions in the HGDP dataset varied considerably among individuals within mtDNA haplogroups, as measured by the standard deviation of ancestry proportions within each haplogroup. For example, the s.d. of East Asian ancestry was > 0.40 in haplogroups M, C, D, N, A, and HV, which limits the conclusions we could make regarding East Asian ancestry in individuals from these haplogroups. Standard deviations of continental ancestry proportions within each haplogroup were generally lower in the 1,000 Genomes populations, perhaps because these populations have lower diversity of geographic origin than those from the HGDP.

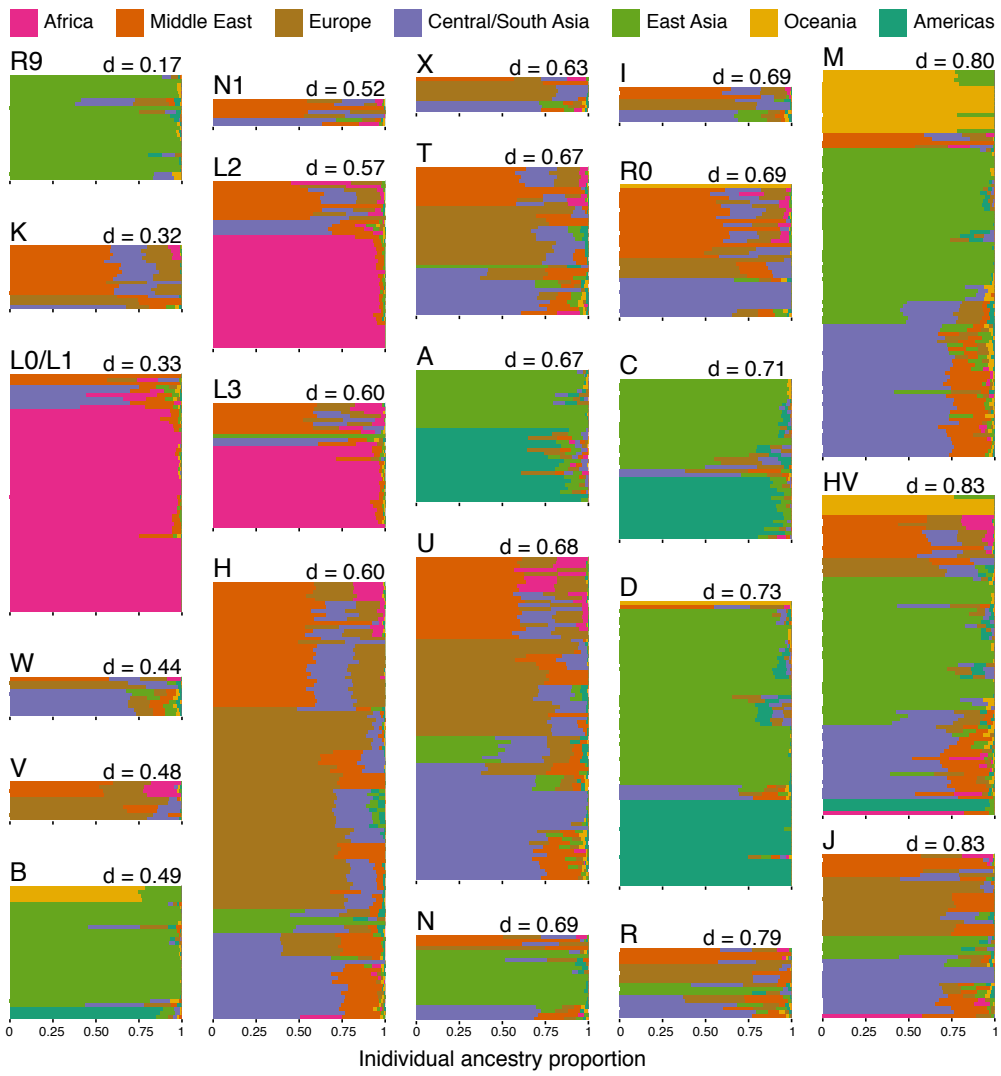


Figure 2.2 Individual continental ancestry proportions within each haplogroup in the HGDP
Each horizontal line is a barplot for one of the HGDP samples, indicating individual continental ancestry proportions with continental regions colored by the key at top. Individual barplots are grouped into the 23 mtDNA haplogroups. Each haplogroup is labeled with the mean pairwise Euclidean distance (see methods), and haplogroups are sorted by increasing mean pairwise Euclidean distance from top to bottom and left to right.

To further understand the variability of individual continental ancestry within a haplogroup, we calculated the mean pairwise Euclidean distance between continental ancestry proportions among individuals within each haplogroup. This distance is a

quantitative measure of the inter-individual variability in continental ancestry proportions within a haplogroup (Figure 2.2, Figure 2.3C). The mean pairwise Euclidean distance was relatively low in some haplogroups (*e. g.* R9 in HGDP, T and H in 1,000 Genomes), indicative of a stronger association between individual continental ancestry proportions and an mtDNA haplogroup. However, the mean pairwise Euclidean distance of other haplogroups was high (*e. g.* J, HV, M in HGDP, L3 in 1,000 Genomes), suggesting that these haplogroups are less informative for individual ancestry determination. The mean pairwise Euclidean distance within each haplogroup was much lower in the 1,000 Genomes populations (Figure A.4) than in the HGDP. Additionally, mean pairwise Euclidean distances were not necessarily similar between the same haplogroup in each dataset (Figure 2.2, Figure 2.3C). For example, haplogroup H had a much lower mean pairwise Euclidean distance in the 1,000 Genomes (d = 0.15) than in the HGDP (d = 0.60). This indicates that inferences drawn from the HGDP dataset are not necessarily applicable to the 1,000 Genomes populations and vice versa.

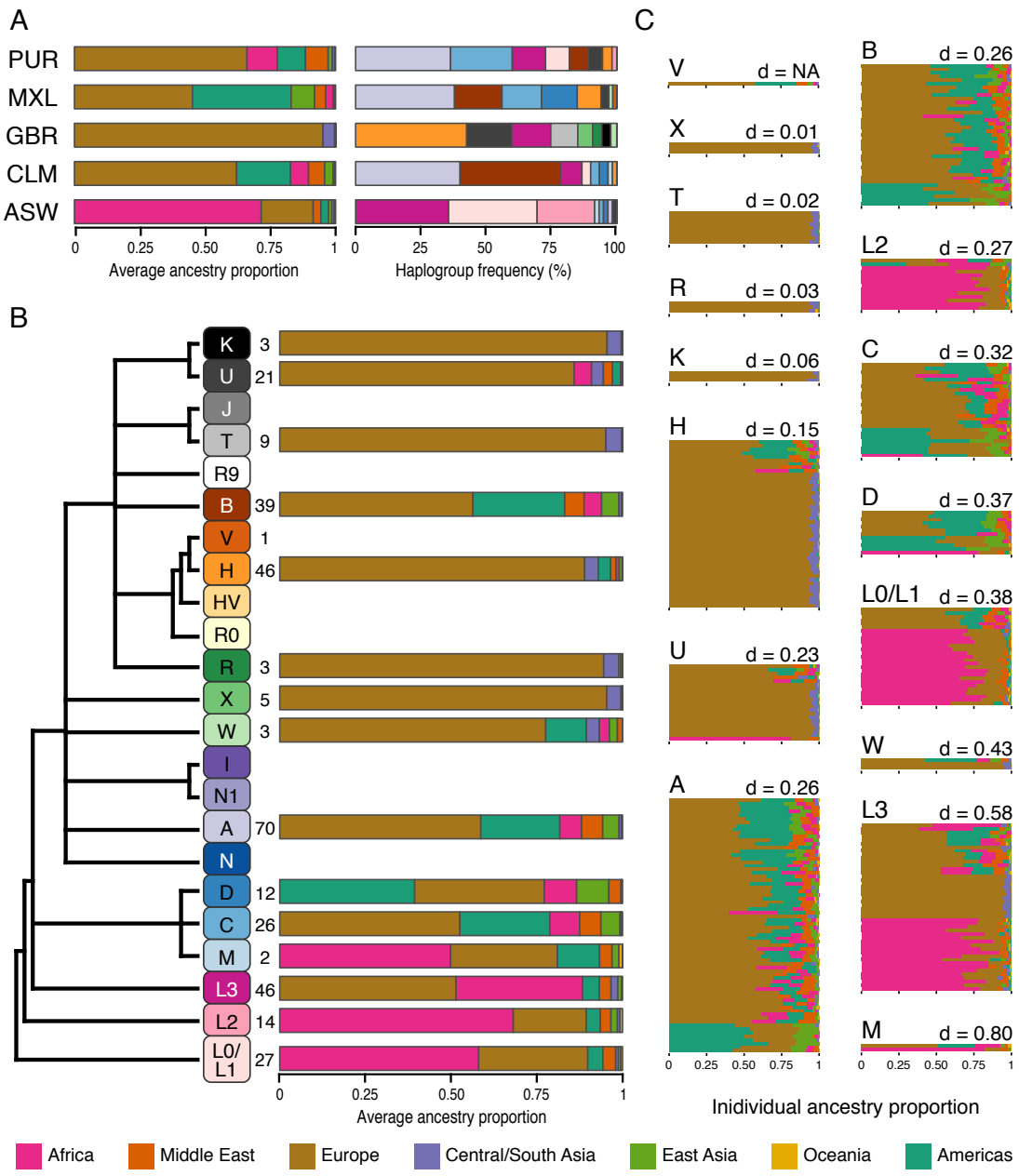


Figure 2.3 mtDNA haplogroup frequencies, population-averaged and haplogroup-averaged continental ancestry proportions, and individual continental ancestry proportions in the 1,000 Genomes dataset.
A) Left column: Barplots of continental ancestry proportions averaged within each 1,000 Genomes population. Barplots are colored by continental region (labeled by colored bars in Figure 1A) and sorted by continental ancestry. Right column: Barplots of haplogroup frequencies within each population, colored by haplogroup (labeled by the haplogroup boxes in panel B). B) Barplots of continental ancestry proportions averaged within each mtDNA haplogroup within the 1,000 Genomes dataset. Barplots are colored by continental region (see key at bottom). The unscaled phylogeny at left is

the same as in Figure 1C. Numbers to the right of haplogroup labels are the sample size for each haplogroup. C) Individual continental ancestry proportions within each haplogroup. Each horizontal line is a barplot for one of the 1,000 Genomes samples, indicating individual continental ancestry proportions with continental regions colored by the key at bottom. Individual barplots are grouped into mtDNA haplogroups. Each haplogroup is labeled with the mean pairwise Euclidean distance (see methods), and haplogroups are sorted by increasing mean pairwise Euclidean distance from top to bottom and left to right.

To quantitatively assess how informative each haplogroup in the HGDP populations is for predicting individual ancestry, we calculated consistency within each haplogroup in the HGDP (Table A.4). Consistency measures the percentage of the time that an individual's highest ancestry proportion matches the highest average ancestry proportion in that individual's haplogroup. Haplogroups in the HGDP dataset had a mean consistency of 0.56, meaning that slightly more than half of the haplogroups were more consistent than not (Figure A.5). Haplogroups with high consistency scores had low mean pairwise Euclidean distance values (Pearson's product-moment correlation; HGDP: $R^2 = -0.86$, $p < 0.001$). Haplogroup R9, found only in East Asia, had the lowest mean pairwise Euclidean distance value ($d = 0.17$) and was also the most consistent (consistency = 0.92), suggesting that it could reasonably be described as an East Asian haplogroup. We measured consistency within the 1,000 Genomes populations, using the HGDP as a reference panel to determine the maximum continental ancestry component for each haplogroup. Consistency measures what proportion of 1,000 Genomes individuals from a given haplogroup have 50% or more continental ancestry from the maximum ancestry region of the haplogroup in the HGDP. About half of the haplogroups had high consistency scores (*i. e.* > 50%) in the 1,000 Genomes data – in many cases with higher consistency scores than in the HGDP

itself (Figure A.5, Table A.4). The correlation between consistency and mean pairwise Euclidean distances was both weaker and not statistically significant in the 1,000 Genomes data (R^2 = -0.49, p = 0.07). As with mean pairwise Euclidean distance estimates, the consistency score of some haplogroups varied between the HGDP and 1,000 Genomes datasets (*e. g.* haplogroup A: d = 0.67 in HGDP, d = 0.26 in 1,000 Genomes), and several even had consistency scores of 0 in the 1,000 Genomes (M, C, D, W, B, U, & K).

2.4.4 Predicting mtDNA haplogroup from continental ancestry

To determine the predictive relationship between individual continental ancestry and mtDNA haplogroups, we fit a multinomial logistic regression (logit) model to predict a sample's mtDNA haplogroup from its continental ancestry proportions. Our logit model was a significantly better fit to the data than a null model (LRT, $p < 0.001$). McFadden's pseudo-R^2 for this final model is 0.84, indicating a very good fit to the data. The effect sizes of the estimated coefficients revealed interesting relationships between continental ancestry proportions and mtDNA haplogroups. For example, the combination of Middle Eastern and Central/South Asian ancestry drastically decreased a person's probability of belonging to haplogroup C – the highest coefficient effect size (β = -1,166.12). The second-highest coefficient effect size (β = -1,073.69) indicated that the interaction term between European and African ancestry drastically decreased a person's probability of belonging to haplogroup B.

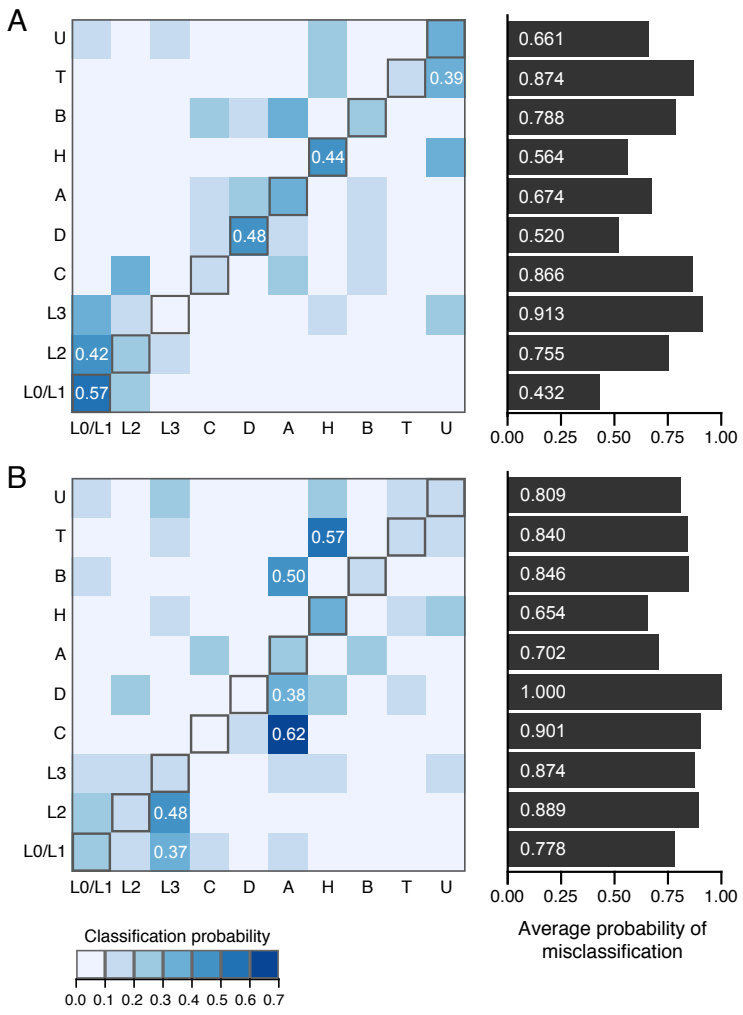


Figure 2.4 Misclassification probabilities in the HGDP and 1,000 Genomes datasets. Each cell in row *i*, column *j*, denotes the probability of a sample experimentally determined as haplogroup *i* being classified as haplogroup *j*, based on a fitted logit model. Cells are colored by increasing classification probability values from white to blue (see key at bottom). Diagonal entries (gray outlines) are the probability of being classified correctly. Barplots on the right show the average probability of misclassification for each haplogroup, which is the total of all non-diagonal values in each row. The top 10% of non-zero classification probabilities are labeled (white text). A) misclassification in the HGDP dataset B) misclassification in the 1,000 Genomes dataset.

Using the logit model to predict the haplogroup for each of the individuals in our test set, 24% of the predictions were correct in the HGDP populations; for comparison, a random guess would be correct 10% of the time. Furthermore, the classification

probabilities for the correct haplogroup of each sample were significantly higher than the classification probabilities for the incorrect haplogroups (t test, $p < 0.001$). In the 1,000 Genomes populations, correct haplogroup classification probabilities were also significantly higher than incorrect haplogroup classification probabilities (t test, $p = 0.02$) (Figure A.8). However, only 7% of the predictions were correct in the 1,000 Genomes populations. Prediction accuracy was significantly higher in the HGDP (χ^2, $p = 1 \times 10^{-14}$), even though both datasets were used to build the model. For both datasets, there were a considerable number of samples for which the prediction's classification probability was very high, but the prediction was nonetheless incorrect (Figure A.8).

We examined the classification probabilities within each haplogroup in more detail to identify which haplogroups were most likely to be classified incorrectly according to our logit model (Figure 2.4). In the HGDP samples, classification accuracy ranged from 9% (L3) to 57% (L0/L1) (Figure 2.4A). Except for haplogroup L0/L1, all mtDNA haplogroups were more likely to be classified incorrectly than correctly, based on continental ancestry proportions. For many haplogroups, the highest classification probability did not match the correct haplogroup (*e. g.* L2 classified as L0/L1: 42%, classified correctly 9%; T classified as U: 39%, classified correctly: 13%; L2 classified as L0/L1:%, classified correctly: 24%). Samples from haplogroups A, B, C, and D were often incorrectly classified as one other (cluster of higher values in Figure 2.4A). Classification probability showed similarly poor performance in the 1,000 Genomes dataset. All haplogroups had higher probabilities of misclassification than observed for the HGDP data, and none were more likely to be classified correctly than incorrectly

(Figure 2.4B). One haplogroup (D) was never classified correctly. Classification probabilities for a given haplogroup in the 1,000 Genomes dataset usually differed from their counterparts in the HGDP populations.

2.5 Discussion

2.5.1 Conclusions

To address the American Society of Human Genetics' concerns regarding lineage-based ancestry testing (The American Society of Human Genetics 2008; Royal *et al.* 2010), we have presented a detailed portrait of the variation in continental ancestry observed within major mtDNA haplogroups. We used a novel combination of genome-wide SNP data and mtDNA haplogroup classifications in a worldwide sample of geographically diverse populations. We found that there is striking heterogeneity of intra-haplogroup continental ancestry proportions. While many haplogroups were associated with their primary ancestry components, secondary ancestry components rarely showed such associations. Furthermore, predicting an individual's mtDNA haplogroup from their continental ancestry proportions was incorrect more often than not. Finally, we found that the relationship between an individual's mtDNA haplogroup and their continental ancestry was less informative in recently admixed populations.

We found that all but a few of the major mtDNA haplogroups display a high amount of variation in continental ancestry proportions. MtDNA haplogroup distributions in the HGDP agree with previously observed descriptions of geographic distributions (*e. g.* L0/L1, L2, and L3 frequent in Africa, H frequent in Europe, *etc.*). However, there are some significant departures from these generalizations; for example, 24% of the Brahui population belonged to haplogroup H, though average

European ancestry in the Brahui was only 0.01. Moreover, high inter-individual variation in continental ancestry proportions, as measured by mean pairwise Euclidean distance (Figure 2.4), indicates that several mtDNA haplogroups are comprised of individuals from many diverse ancestry backgrounds.

Our results show that most mtDNA haplogroups are associated with high ancestry in a particular continental region. Furthermore, the highest continental ancestry proportion for most individuals is consistent with that of the haplogroup to which they belong. While this information appears to support the usefulness of mtDNA haplogroups for determining ancestry, an appreciable number of haplogroups do not exhibit this pattern. Some mtDNA haplogroups also show significant associations with higher-than average continental ancestry from a secondary region, but most do not. The second-highest and third-highest continental ancestry proportions within an individual's overall ancestry profile can vary significantly within a haplogroup, despite comprising significant portions of an individual's ancestry. For example, haplogroup K is associated with high Middle Eastern ancestry, but not with European or Central/South Asian ancestry, although these regions contribute an average of 25% and 21% ancestry, respectively. Thus, while an mtDNA haplogroup classification might provide accurate information about a person's highest continental ancestry component, information regarding an individual's other ancestry components is limited.

So far, we have discussed our approaches for inferring a person's continental ancestry proportions from their mtDNA haplogroup, which is unsurprisingly difficult. If the relationship between mtDNA haplogroups and continental ancestry is consistent and considerable, we should be able to reverse this approach by using continental

ancestry proportions to infer a sample's mtDNA haplogroup with a logit model. However, our logit-based mtDNA haplogroup predictions were often incorrect (76% incorrect in the HGDP), despite the good fit of our model. Some haplogroups were difficult to distinguish from one another and many samples were misclassified as a phylogenetically distant mtDNA haplogroup (*e. g.* C misclassified as L2, Figure 2.4A). Our failed predictions were not necessarily "close" to being correct. Often the failed predictions had very high classification probabilities, so the strength of the prediction did not necessarily indicate its confidence level (Figure A.8). The inaccuracy of these predictions provides another line of evidence regarding the amount of ancestry information that is not captured by mtDNA haplogroup classifications.

We also repeated all of our analyses in five recently admixed populations from the 1,000 Genomes Project: ASW, CLM, GBR, MXL, and PUR. We used the HGDP populations as a reference panel for our analyses of the 1,000 Genomes dataset, to assess the effects of reference panels on lineage-based ancestry inferences. We first observed that the geographic association of mtDNA haplogroups could break down in recently admixed populations. Additionally, we found fewer significant associations between a haplogroup and its primary ancestry component, different amounts of intra-haplogroup ancestry variation, and less accurate logit-based haplogroup predictions in the 1,000 Genomes compared to the HGDP. We conclude that, for recently admixed populations, the predicted continent of origin based on mtDNA lineage testing may not accurately reflect a high proportion of ancestry from the same continent. For example, in the admixed populations that we examined, "Native American", "East Asian", and "African" mtDNA haplogroups (A, B + C + D, and L3, respectively) show mostly

European continental ancestry for the autosomes. Indeed, similar discrepancies between autosomal ancestry and ancestry based on mtDNA haplogroup have previously been observed in Colombian (Salas *et al.* 2008), Bolivian & Totonac (Watkins *et al.* 2012), Brazilian (Cardena *et al.* 2013), and Malagasy (Poetsch *et al.* 2013) populations. This suggests that lineage-based ancestry inferences may be more misleading in recently admixed populations, and are highly dependent upon which populations are included in the reference panel.

2.5.2 Limitations of lineage-based ancestry testing

Our results demonstrate the impact that reference panel populations have on lineage-based ancestry inferences. In particular, reference panels drawn from historically isolated populations may produce misleading results when interpreting mtDNA-based ancestry within recently admixed populations. The HGDP dataset notably under-samples populations from Oceania and the Americas, and lacks populations from India. This decreases the geographic scope of our conclusions, and DTC testing companies may have similarly underrepresented world regions whose ancestry they cannot assess. If mtDNA-based ancestry is often decoupled from continental ancestry, as in the admixed populations we included, the results reported to consumers may be misleading or difficult to interpret.

We have also characterized the extent of ancestry information captured by mtDNA haplogroup classifications. A highly informative mtDNA haplogroup would have a low mean pairwise Euclidean distance, an association with ancestry proportion(s) from specific continental region(s), low standard deviation of ancestry proportions, and high consistency. Only three haplogroups satisfy these criteria: R9, K, and L0/L1.

Additionally, an informative haplogroup would have a high consistency score in both the reference panel and in admixed populations, which is the case for L0/L1, but not for K (the 1,000 Genomes populations lacked R9). Our results demonstrate that the majority of mtDNA haplogroups convey information about one, or possibly two, top ancestry components while other ancestry information is lost. MtDNA haplogroups are frequently assigned to a continental group (*e. g.* an "African haplogroup" or a "European haplogroup"), but these descriptions do not present the whole picture of continental ancestry variation within each haplogroup. For any particular consumer, the mtDNA haplotype could have been inherited from a population in which it is less common, despite its high frequency elsewhere (Bolnick *et al.* 2007). Effectively communicating this complexity to a DTC test consumer poses a significant challenge (Wagner *et al.* 2012), which may be alleviated by the transition to new continental ancestry testing products, based on panels of autosomal markers. These SNP-based tests report results as continental ancestry proportions, which will make results easier for consumers to interpret.

It is particularly problematic to reduce mtDNA sequence information to haplogroup labels. Using mtDNA haplogroup classifications to look for mitochondrial disease associations can lead to spurious associations that suggest haplogroup-defining polymorphisms are responsible for disease (Achilli *et al.* 2008; Benn *et al.* 2008). Because mtDNA is a single locus without recombination, selection anywhere on the mito-genome would dramatically affect the coalescent history of human mtDNAs (Underhill and Kivisild 2007; Barbujani and Colonna 2010). Others have suggested that the haplogroup-defining polymorphisms might themselves be adaptive variants, which

would make mtDNA haplogroups poor tools for making inferences regarding neutral demographic processes (Ruiz-Pesini *et al.* 2004). There are also technical issues related to the mtDNA haplogroup classification methods, such as the difficulty in classifying samples based only on HVR sequences (Behar *et al.* 2007; Wong *et al.* 2011). New evidence also demonstrates that there is a very high rate of back-mutation and heteroplasmy in at least one subhaplogroup's diagnostic SNP (Duggan and Stoneking 2013). This suggests that other understudied diagnostic SNPs may be subject to back-mutation and heteroplasmy as well. While mtDNA lineage tests have been a powerful tool in the past, their usefulness has diminished with the advent of genome-wide SNP and sequence data.

2.5.3 The continued utility of lineage-based studies

Mitochondrial haplogroup analyses are particularly suited to analysis of ancient DNA. The abundance of mtDNA makes it easier to obtain from degraded ancient samples than nuclear DNA. Recent examples have used ancient mtDNA to trace ancient migrations (Bollongino *et al.* 2013; Brandt *et al.* 2013) and to calibrate a molecular clock for dating population separations (Fu *et al.* 2013). Ancient mtDNA demographic analysis is particularly effective when it includes whole mito-genomes and does not reduce mtDNA sequences to haplogroup classifications (Bollongino *et al.* 2013).

For modern DNA samples, there are many alternatives and complementary approaches to using mtDNA haplogroups for making inferences about genetic ancestry, demography, migrations, and sex-biased gene flow. Several recent studies have included analysis of nuclear DNA sequences alongside mtDNA, to striking effect

(Simonson *et al.* 2011; Jinam *et al.* 2012; Watkins *et al.* 2012). Additionally, the demographic history of an admixed population can be inferred in much more detail using nuclear genome SNP data. For example, Moreno-Estrado *et al.* used ancestry tract length information to make very detailed conclusions about the demographic history of Caribbean populations (Moreno-Estrada *et al.* 2013). For investigating sex-biased demographic processes, comparing variation on the X chromosome to that on the autosomes provides a much more detailed description than comparing haplogroup frequencies from NRY and mtDNA (*e. g.* Hammer *et al.* 2008; Keinan *et al.* 2009). These studies, among others available in the recent literature, provide good examples for alternatives to mtDNA demographic studies, and for integrating information from mtDNA with that from the nuclear genome.

2.6 Web Resources

The URLs for data presented herein are as follows:

1,000 Genomes Project: http://www.1000genomes.org

Genographic Project Haplogroup Prediction Tool: http://nnhgtool.nationalgeographic.com/

HGDP SNP data: http://www.hagsc.org/hgdp/files.html

`PLINK` v1.07: http://pngu.mgh.harvard.edu/purcell/plink/

3. Inferring the effective sex ratio in recent human evolution

This chapter is based on the following published paper:

Leslie S. Emery, Joseph Felsenstein, and Joshua M. Akey. (2010) Estimators of the human effective sex ratio detect sex biases on different timescales. *The American Journal of Human Genetics.* 87 (6), 848-856.

Joseph Felsenstein developed the coalescent model (sections B.1 & B.2) for the expected sex ratio, and I performed all other analyses. Joshua Akey and I wrote and revised the manuscript, with additional revisions by Joseph Felsenstein.

3.1 Summary

Determining historical sex ratios throughout human evolution can provide insight into patterns of genomic variation, the structure and composition of ancient populations, and the cultural factors that influence the sex ratio (*e. g.* sex-specific migration rates). Although numerous studies have suggested that unequal sex ratios have existed in human evolutionary history, a coherent picture of sex-biased processes has yet to emerge. For example, two recent studies compared human X chromosome to autosomal variation to make inferences about historical sex ratios, but reached seemingly contradictory conclusions, with one study finding evidence for a male bias and the other study describing a female bias. Here we show that a large part of this discrepancy can be explained by methodological differences. Specifically, through reanalysis of empirical data, derivation of explicit analytical formulae, and extensive simulations we demonstrate that two estimators of the effective sex ratio based on population structure and nucleotide diversity preferentially detect biases that have occurred on different timescales. Our results clarify apparently contradictory evidence on the role of sex-biased processes in human evolutionary history and show that extant patterns of human genomic variation are consistent with both a recent male bias and an earlier, persistent female bias.

3.2 Main Text

Although studies of DNA variation have revealed important insights into human demographic history, comparatively little is known about mating patterns and sex ratio during human evolution (Quinlan 2008). Sex-biased processes, such as matrilocality—when females remain in their natal territory—and polygyny—when males have multiple

female mates—are widespread in mammals and can have profound effects on genomic patterns of variation (Greenwood 1980; Hedrick 2007). One measure of the sex bias within a population is the effective sex ratio (ESR)—defined here as the female proportion of the effective population ($ESR = \frac{N_e^{female}}{(N_e^{female} + N_e^{male})}$). Diversity measures from the mtDNA and the non-recombining portion of the Y chromosome (NRY) provide relative estimates of N_e^{female} and N_e^{male}. Previous studies comparing mtDNA and NRY have shown evidence for local-scale sex biases in migration rates of humans (Seielstad *et al.* 1998; Oota *et al.* 2001; Wilder, Kingan, *et al.* 2004; Wilder, Mobasher, *et al.* 2004; Hamilton and Stoneking 2005) and other species (Melnick and Hoelzer 1992; Eriksson *et al.* 2006; Douadi *et al.* 2007; Langergraber *et al.* 2007). Because these uniparentally inherited markers experience no recombination, however, selection on any part of the mtDNA or NRY will affect the entire locus and make ESR estimates difficult to interpret (Wilkins and Marlowe 2006).

Recently, the availability of sequence data has enabled comparisons of X chromosome and autosomal variation levels (Ramachandran *et al.* 2004; Hammer *et al.* 2008; Ramachandran *et al.* 2008; Keinan *et al.* 2009; Labuda, J.-F. Lefebvre, *et al.* 2010), which have higher power to make global-scale inferences about human sex biases than inferences based on mtDNA or NRY (Goudet *et al.* 2002; Wilkins and Marlowe 2006). These comparisons rely on a consequence of male hemizygosity: the effective number of X chromosomes in a population (N_e^X) depends on the ESR. If males and females are present in equal numbers (ESR = 0.5), then the effective population

size of the X is three quarters that of the autosomes (Hedrick 2007). This relationship is described by the ratio *Q* (Keinan *et al.* 2009):

$$Q = \frac{N_e^X}{N_e^A} \cong 0.75 \qquad \textbf{Eq. 3.1}$$

However, sex biases can lead to deviations from $Q = 0.75$: in cases of a male bias ($N_e^{female} < N_e^{male}$) there is a relative reduction in the number of X chromosomes, decreasing *Q* ($Q < 0.75$); in cases of a female bias ($N_e^{female} > N_e^{male}$) there is a relative increase in the number of X chromosomes, which increases *Q* ($Q > 0.75$). Because the effective population size of the X chromosome determines the rate of genetic drift on the X, *Q* can be estimated by comparing levels of genetic diversity between the X chromosome and the autosomes. In population data, *Q* can be estimated from statistics such as the fixation index (F_{ST}) and nucleotide diversity (π) (Hammer *et al.* 2008; Keinan *et al.* 2009), and serves as a proxy for the ESR in detecting sex biases. Several recent studies have compared X and autosomal variation to make inferences regarding sex biases in *Drosophila* (Begun and Whitley 2000; Musters *et al.* 2006; Singh *et al.* 2007) and in humans (Hammer *et al.* 2004; Ramachandran *et al.* 2004; Hammer *et al.* 2008; Ségurel *et al.* 2008; Keinan *et al.* 2009).

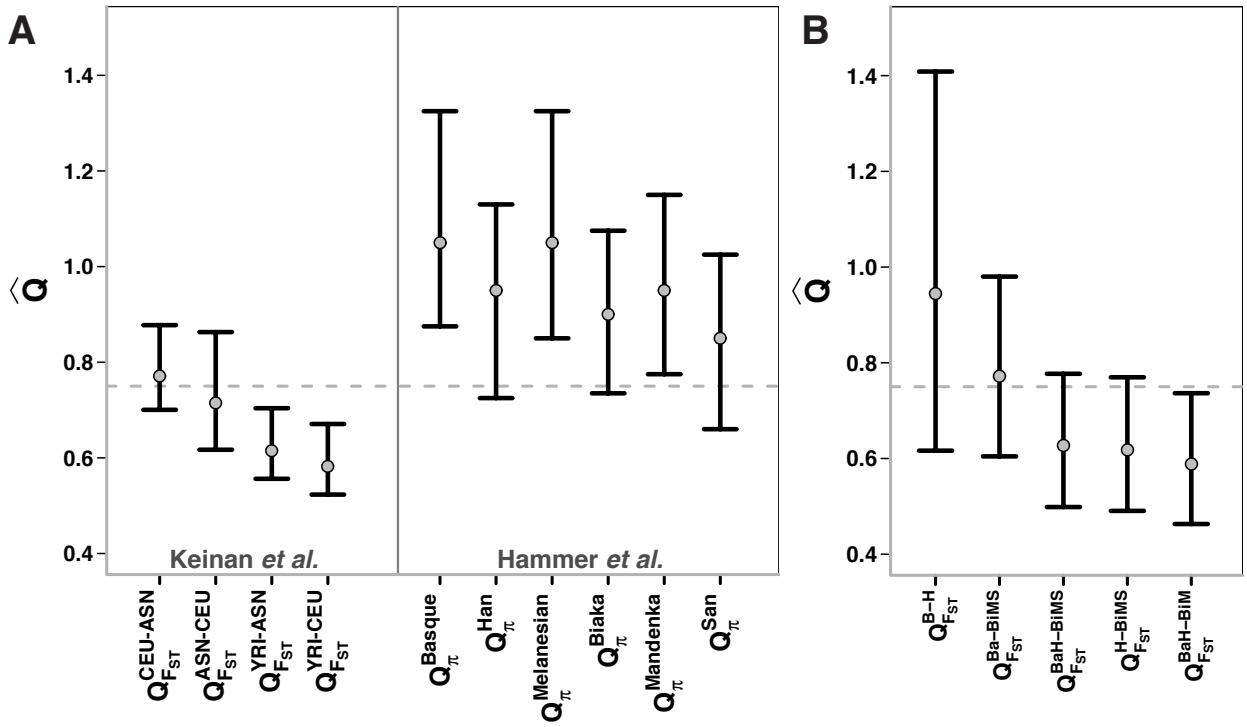


Figure 3.1 Estimates of Q_{FST} in Hammer's resequencing data are consistent with a male bias during the Out of Africa disperal.
Point estimates of *Q* are denoted by gray dots and vertical black bars represent 95% confidence intervals. The dashed gray line indicates the expected value of $Q = 0.75$ with no sex bias. **(A)** Summary of two previous studies of human sex bias (Hammer *et al.* 2008; Keinan *et al.* 2009). The x-axis shows the populations (superscript) and the variation measure used (subscript). A comparison between two populations is denoted by the population names connected by a hyphen. HapMap population abbreviations: ASN = Japanese in Tokyo and Han Chinese in Beijing; CEU = Utah residents with ancestry from northern and western Europe; YRI = Yoruba in Ibadan, Nigeria. **(B)** Reanalysis of Hammer's data with the F_{ST} method. Several *Q* estimates are below 0.75, in contrast to estimates of Q_π in the same data set. Ba: Basque; H: Han; BaH: Basque+Han; BiM: Biaka+Mandenka; BiMS: Biaka+Mandenka+San.

Recently, two studies estimated *Q* in order to detect sex biases in similar human populations (Hammer *et al.* 2008; Keinan *et al.* 2009) and found seemingly contradictory conclusions (Bustamante and Ramachandran 2009). Using SNP data from the International HapMap Project (International HapMap Consortium *et al.* 2007), Keinan *et al.* found evidence for a male bias during the dispersal of modern humans out of Africa (Figure 3.1A) (Keinan *et al.* 2009). Hammer and colleagues, however,

found evidence for a female bias throughout human history in six populations from the Human Genome Diversity Panel (HGDP) (Figure 3.1A) (Hammer *et al.* 2008). Although these two analyses differ in several respects, such as the specific populations and markers analyzed, we were especially interested in whether methodological differences could account for the disparate results. In particular, the primary analysis of Keinan *et al.* used F_{ST} to estimate Q while Hammer *et al.* estimated Q using π (we will denote these estimators as Q_{FST} and Q_{π}):

$$Q_{\pi} = \frac{\pi^{X}}{\pi^{Autosomes}} \qquad \textbf{Eq. 3.2}$$

$$Q_{F_{ST}} = \frac{\ln(1 - 2F_{ST}^{A})}{\ln(1 - 2F_{ST}^{X})} \qquad \textbf{Eq. 3.3}$$

We investigated the properties of these estimators of Q using three independent methods, including reanalysis of empirical data, a coalescent theoretical model, and detailed coalescent simulations.

The most direct way to explore the methodological differences in Q obtained by Keinan *et al.* and Hammer *et al.* is to calculate both Q_{π} and Q_{FST} on the same empirical data set. To this end, we obtained the resequencing data from Hammer *et al.*, which consists of 20 regions (~5 kb) each for the X chromosome and autosomes, and calculated Q_{FST} for all possible population pairs. The populations included in the data set are French Basque, Biaka, Han Chinese, Mandenka, Melanesian, and San (Wall *et al.* 2008). To mitigate the effects of recent sex biases unique to one population on estimates of Q_{FST}, we also performed analyses with combinations of populations (Biaka+Mandenka, Basque+Han, & Biaka+Mandenka+San). To calculate F_{ST} from this

data, we tabulated allele frequencies in each population, excluding SNPs with a minor allele frequency < 0.05 ($SNP_A = 276$, $SNP_X = 252$). We calculated all pairwise F_{ST} estimates using Weir and Cockerham's estimator (Weir 1996), and performed non-parametric bootstrapping over the SNPs to estimate 95% confidence intervals from 1,000 bootstrap replicates. Combining the 1,000 F_{ST}^X estimates and 1,000 F_{ST}^A estimates into all 1,000,000 possible combinations, we obtained 1,000,000 estimates of Q_{FST} for each pair of populations using Eq. 3.3.

In the populations most closely related to the HapMap populations, most estimates of Q_{FST} are below 0.75 (Figure 3.1B), which is more consistent with the observations of Keinan *et al.* (using the same method) than with those of Hammer *et al.* (using the same data). Of the twenty-five possible comparisons of a non-African to an African population, only four exhibit a female bias (Figure B.1). In fact, over half of these non-African vs. African comparisons display a male bias (four significantly so). Given the exact same data set, we still see marked differences between Q_π and Q_{FST} estimates in resequencing data, which is a compelling reason to investigate the methodological differences further.

To better understand the differences between Q_{FST} and Q_π, we first derived analytical expressions for both estimators under a coalescent model. Eq. 3.4 can be used to calculate the expected value for π on either the X chromosome or autosomes, using the appropriate mutation rate μ (see Appendix B for a complete derivation):

$$\pi = 2\mu \sum_{i=1}^{n} \left(e^{-\left(\sum_{k=i+1}^{n} \frac{T_k}{2N_k} \right)} \right) \left(1 - e^{-\frac{T_i}{2N_i}} \right) \left[2N_i + \sum_{k=i+1}^{n} T_k - T_i \left(\frac{e^{\frac{-T_i}{N_i}}}{1 - e^{\frac{-T_i}{N_i}}} \right) \right] \quad \textbf{Eq. 3.4}$$

Eq. 3.4 is derived from a model based on a single lineage that is partitioned into nonoverlapping intervals described in terms of a series of population sizes N_n, N_{n-1}, N_{n-2}, … N_2, N_1 proceeding from the present backwards in time. Each interval has an associated duration describing how long the population remained at that size, giving a series of durations T_n, T_{n-1}, T_{n-2}, … T_2, T_1 measured in generations.

We can extend the above model to two subpopulations that diverged from the same ancestral population t generations ago. The ancestral population has population sizes and associated durations N'_n, N'_{n-1}, N'_{n-2}, … N'_2, N'_1 and T'_n, T'_{n-1}, T'_{n-2}, … T'_2, T'_1. We can also derive an expression (Eq. 3.5) for the expectation of π between subpopulations 1 and 2, which we denote as π_{12}:

$$\pi_{12} = 2\mu \sum_{i=1}^{n} \left(e^{-\left(\sum_{k=i+1}^{n} \frac{T'_k}{2N'_k} \right)} \right) \left(1 - e^{-\frac{T'_i}{2N'_i}} \right) \left[2N'_i + 2t + \sum_{k=i+1}^{n} T'_k - T'_i \left(\frac{e^{\frac{-T'_i}{N'_i}}}{1 - e^{\frac{-T'_i}{N'_i}}} \right) \right] \quad \textbf{Eq. 3.5}$$

By using a formula for F_{ST} in terms of the three measures π_1, π_2, and π_{12} (Hudson *et al.* 1992), we can calculate approximate expected values for F_{ST} on the X chromosome or on the autosomes (see Appendix B):

$$F_{ST} = \frac{2\pi_{12} - \pi_1 - \pi_2}{2\pi_{12}} \quad \textbf{Eq. 3.6}$$

A model of the African (Af), Asian (As), and European (Eu) populations in terms of N and T parameter pairs is given in Figure B.2. We used equations Eq. 3.4, Eq. 3.5, and

Eq. 3.6 to calculate π^X, π^A, F_{ST}^X, and F_{ST}^A under this model, and then equations Eq. 3.2 and Eq. 3.3 to obtain Q_{FST} and Q_π. The expected values of both Q_π and Q_{FST} from the theoretical model without sex biases are slightly below 0.75, with the notable exception of Q_π^{Af} ($Q_\pi^{Af} = 0.778$; $Q_\pi^{Eu} = 0.740$; $Q_\pi^{As} = 0.736$; $Q_{FST}^{Eu\text{-}Af} = 0.740$; $Q_{FST}^{As\text{-}Af} = 0.738$; $Q_{FST}^{As\text{-}Eu} = 0.735$). Population size dynamics alone can have a significant impact on the null expectation of Q_π in the absence of sex bias (Pool and Nielsen 2007), and it is interesting to note that this phenomenon also affects Q_{FST}.

To investigate the effects of a sex bias, we calculated Q_{FST} and Q_π in each population for sex biases of varying severity at 295 different time points for each of the three populations (Figure 3.2, Figure B.3, Figure B.4, & Figure B.5). Specifically, we introduced a 1,400 generation-long sex bias into a single population 225,000 generations ago, and moved this bias forward in time in 250-generation increments. At each increment, Q_{FST} and Q_π were calculated as described above. The African lineage most clearly demonstrates the different effects of the same sex bias on the two estimators of Q (Figure 3.2 A & D). Q_π in Africans is virtually unaffected by the time of the bias, while the magnitude of Q_{FST} in Africans shifts further away from $Q = 0.75$ for recent biases. The same general patterns are observed for biases introduced into the European and Asian lineages (Figure 3.2B & C, E & F).

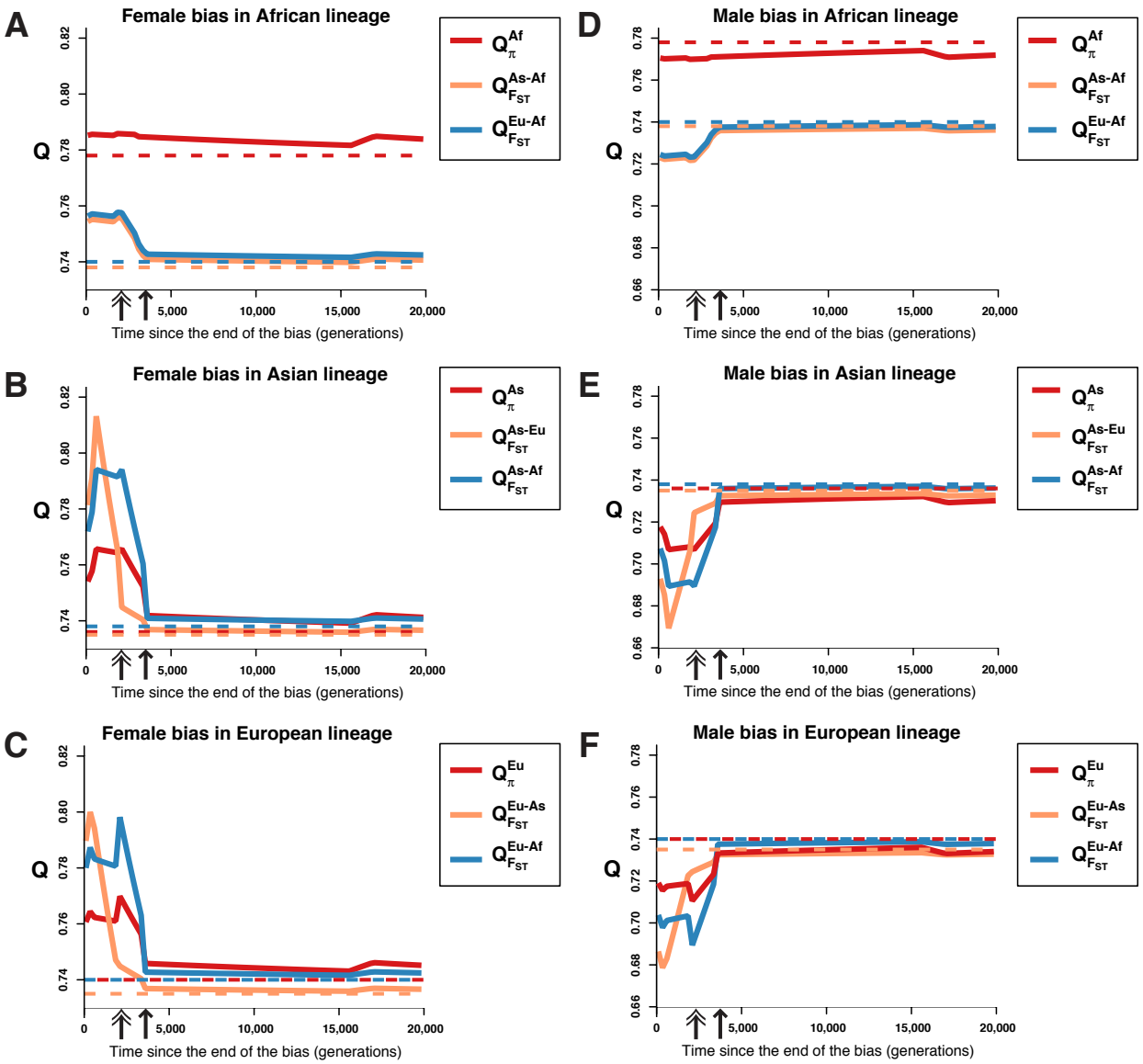


Figure 3.2 Expected values of Q_{FST} and Q_{π} after a bias in a population's history, using a theoretical coalescent model.
Each plot displays the expected values of relevant *Q* estimators after a single female bias (ESR = 0.9) (**A** - **C**) or male bias (ESR = 0.1) (**D** - **F**) lasting 1,400 generations introduced into a single lineage's history. The x-axis indicates the number of generations elapsed since the bias ended. Double-headed arrows indicate the time of the split between Asian and European populations and single-headed arrows indicate the time of the split between Africans and non-Africans. Solid lines denote the results after introducing a bias, while dashed lines in a corresponding color indicate the null theoretical expectation of each *Q* estimator in the absence of a sex bias.

For non-African populations, Q_π does change as the bias becomes more recent, but the magnitude of the changes in Q_{FST} is much larger. These changes in Q_π for non-Africans are attributable to the introduction of a bias during or near a population bottleneck, growth, or expansion event that amplifies the signal of the bias. The sensitivity of Q_{FST} to a recent bias is explained by closely examining the expected values of the estimator in a non-African population, for instance Europeans. There are two Q_{FST} comparisons for the European population: $Q_{FST}^{Eu\text{-}As}$ and $Q_{FST}^{Eu\text{-}Af}$. There is a region of time—between double- and single-headed arrows—when $Q_{FST}^{Eu\text{-}As}$ is still largely unchanged while $Q_{FST}^{Eu\text{-}Af}$ has already decreased (male bias—Figure 3.2B) or increased (female bias—Figure 3.2E). When the bias in question ends before the divergence of the populations being compared, it has less of an effect on $Q_{FST}^{As\text{-}Eu}$ because the bias occurred during their shared history. When the bias starts after the two populations diverge, however, only one population experiences the bias so the differences between the two populations are greater. These results suggest a timescale hypothesis, which posits that Q_{FST} and Q_π are influenced by biases on different timescales: Q_{FST} is mainly influenced by sex biases occurring in the portion of time after two populations diverge while Q_π is influenced by biases along the whole lineage.

To complement and extend the theoretical analyses described above, we also performed extensive coalescent simulations to explore the behavior of Q_{FST} and Q_π under more complex demographic situations with different magnitudes and durations of male and female sex biases. We used the program `ms` (Hudson 2002) to simulate samples from Africans, Europeans, and Asians using a best-fit model of evolution

derived from the HapMap data (Schaffner *et al.* 2005), including bottlenecks, population expansions, and population splits with no migration (see Table B.1 and Table B.2 for model parameters and `ms` command lines). Using this basic coalescent model, we first simulated autosomal and X-chromosomal regions with recombination, similar to those sampled by Hammer *et al.* To simulate regions on the X chromosome (with no sex bias) we scaled θ ($\theta = 4N_e\mu$) from the autosomal simulation by 0.75 and also scaled the population recombination rate ρ ($\rho = 4N_e r$) to be half that of the autosomes. For this simulated sequence data we calculated both π^A and π^X for Africans, Europeans, Asians, and non-Africans (nA) and used Eq. 3.2 to estimate Q_π in each population. To model the sampling method of Keinan *et al.* we simulated unlinked SNPs on the autosomes and on the X chromosome—again scaling the X chromosome parameters appropriately—and simulated the ascertainment process by matching the global minor allele frequency spectrum to that of the HapMap SNPs. From these SNPs, we obtained F_{ST}^A and F_{ST}^X for the four comparisons: 1) non-Africans vs. Africans (nA-Af), 2) Europeans vs. Africans (Eu-Af), 3) Asians vs. Africans (As-Af), and 4) Asians vs. Europeans (As-Eu). Then we used Eq. 3.3 to estimate Q_{FST} for each comparison.

To incorporate sex biases into the coalescent model of the X chromosome, we scaled the population size during the Out of Africa bottleneck event (50 generations long) using Eq. 3.7 (Hedrick 2007) to determine the effective size on the X chromosome given the autosomal effective size during the bottleneck and an arbitrary ESR.

$$N_e^X = \frac{9(N_e^A)^2}{16N_e^A - 9N_e^A(ESR)} \qquad \textbf{Eq. 3.7}$$

We used this scaling procedure to simulate sex biases on the X chromosome using the N_e^A value corresponding to the Out of Africa bottleneck, with ESR values ranging from 0.1 (extreme male bias) to 0.9 (extreme female bias). Because our simulations directly manipulate the number of X chromosomes found in human populations, they are agnostic to the specific mechanism causing the bias. According to Eq. 3.7, the X chromosome experiences a more severe bottleneck than the autosomes in the case of a male bias (ESR < 0.5) and a less severe bottleneck in the case of a female bias (ESR > 0.5). We also simulated an extended bias beyond the duration of the bottleneck, with 150 to 1,350 additional generations of sex bias, resulting in biases lasting from 50 to 1,400 generations in total.

Consistent with the observations of Pool & Nielsen (Pool and Nielsen 2007), some of our null estimates of Q_π are above 0.75 ($Q_\pi^{nA} = 0.769$; $Q_\pi^{Af} = 0.791$; $Q_\pi^{Eu} = 0.753$; $Q_\pi^{As} = 0.751$) in the absence of a sex bias due to the bottlenecks and periods of expansion in our model of human evolution. In contrast to the theoretical results described above, simulations show that our model of human evolution leads to a lower Q_{FST} estimate than the 0.75 expectation in the absence of a sex bias ($Q_{FST}^{Eu\text{-}Af} = 0.708$; $Q_{FST}^{As\text{-}Af} = 0.708$; $Q_{FST}^{As\text{-}Eu} = 0.718$; $Q_{FST}^{nA\text{-}Af} = 0.719$). Simulated null estimates for Q_{FST} are lower than those obtained with the theoretical framework, which may be due to the differences between Eq. 3.6 and the Weir & Cockerham estimator of F_{ST}, or to the effects of recombination. For all of the analyses described below, we will use the simulated null estimates of Q for hypothesis testing purposes.

As shown in Figure B.6 and Figure B.7, both Q_π and Q_{FST} decrease in response to a simulated male bias and increase in response to a simulated female bias. Notably,

the relative change in the value of Q_π is much smaller than the change in the value of Q_{FST}. The simulated bias occurs after non-Africans diverge from Africans and has a stronger effect on Q_{FST} than on Q_π. From the data in Figure B.6, it is clear that Q_π is not well suited for detecting recent sex biases associated with the Out of Africa dispersal. These observations, along with the reanalysis of Hammer's data with Q_{FST} and some previous implications in the literature (Goudet *et al.* 2002; Prugnolle and Meeus 2002; Wilkins and Marlowe 2006; Keinan *et al.* 2009; Keinan and Reich 2010), support the hypothesis that Q_π and Q_{FST} detect biases on different timescales. π in each population is a function of polymorphism along the whole lineage, while F_{ST} is a function of polymorphism differences between two populations; therefore, Q_π is affected by sex biases both before and after two populations have split while Q_{FST} is primarily affected by sex biases occurring after the split.

To more explicitly evaluate the different timescales on which Q_π and Q_{FST} detect biases, we performed additional coalescent simulations using the best-fit model of human evolution and including two separate sex biases (Figure 3.3A). We first introduced a female bias lasting for 20,000 generations along the ancestral human lineage, corresponding to the time before the dispersal of modern humans out of Africa. We then introduced a male bias in the non-African lineage, lasting for the 1,400 generations before the split between European and Asian populations. Using this basic set of model parameters, we simulated both linked sequence regions and unlinked SNPs to repeat the Q_π and Q_{FST} estimation procedure described above. We simulated scenarios with an ESR of 0.9, 0.8, 0.7, or 0.6 for the female bias and an ESR of 0.1, 0.2, 0.3, or 0.4 for the male bias.

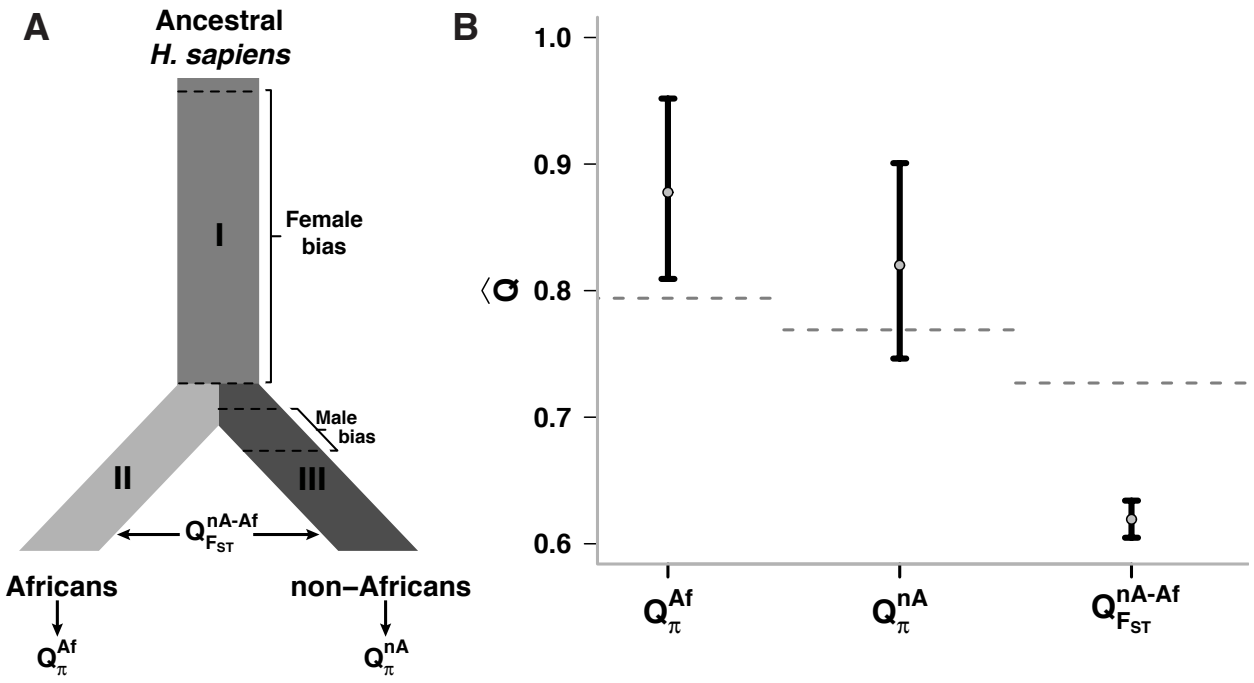


Figure 3.3 Coalescent simulations show that Q_{FST} and Q_{π} detect sex-biased events on different timescales.
(A) Graphical representation of the simulation model. For simplicity, population size changes are not shown and the branch lengths are not to-scale. The three possible *Q* estimates are shown. π^{Af} measures diversity along sections I and II of the lineage, π^{nA} measures diversity along sections I and III, and $F_{ST}^{nA\text{-}Af}$ measures divergence between sections II and III. (B) Results from the simulations of the scenario in A. Gray dashed lines are the null estimates determined by simulations (Figs. S6 & S7). As predicted by the timescale hypothesis, the two estimates of Q_{π} are above 0.75, detecting the early female bias. The estimate of Q_{FST}, however, is below 0.75, detecting the recent male bias. Female bias: ESR = 0.9, 20,000 generations; Male bias: ESR = 0.1, 1,400 generations.

If Q_{FST} and Q_{π} preferentially detect sex biases acting on different timescales, the model considered in Figure 3.3A leads to three testable predictions: 1) Q_{π}^{Af} should be much greater than 0.75, 2) Q_{π}^{nA} should be slightly greater than 0.75, and 3) $Q_{FST}^{nA\text{-}Af}$ should be less than 0.75. The simulation results are in complete agreement with these predictions (Figure 3.3A, Figure 3.4). Interestingly, at least one simulated scenario (Figure 3.3B) produces results for Q_{π}^{Af} and $Q_{FST}^{nA\text{-}Af}$ that are consistent with some of the observations by Hammer *et al.* and Keinan *et al.* Bilinear interpolation heat maps

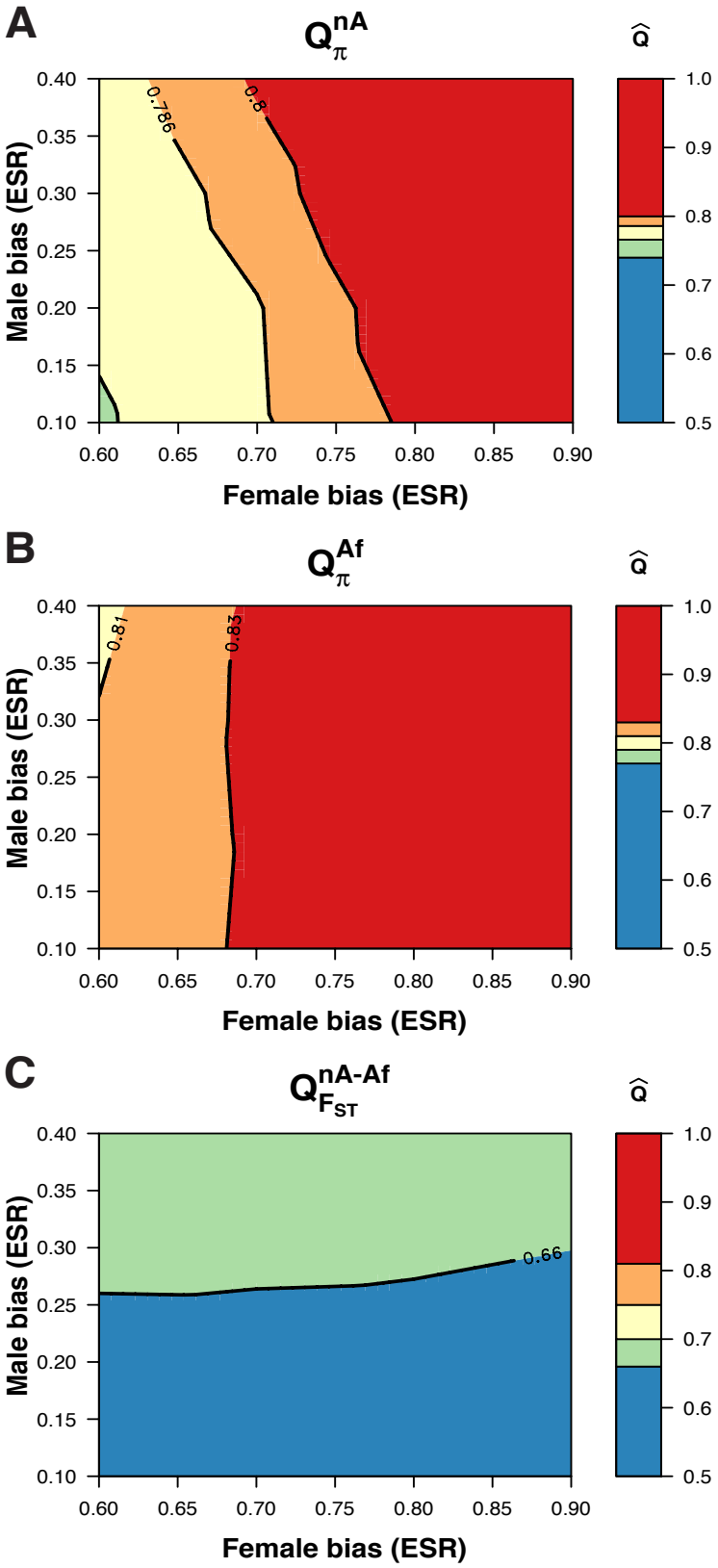


Figure 3.4 Coalescent simulations show that Q_{FST} is primarily influenced by the ESR of a recent bias, while Q_{π} is primarily influenced by the ESR of an ancestral bias. Each panel shows a set of Q estimates from simulated data under the model in figure 3A. The y-axis and x-axis indicate the ESR for the male and female bias, respectively, in the model. The color indicates the value of Q for a combination of male and female ESR's (note the Q scales are different in each plot). (A) Q_{π} in non-Africans is jointly influenced by both the recent and ancient biases, but the signal from the older bias dominates. (B) Africans experience only the female bias, shifting Q_{π} in Africans above 0.75. (C) While non-Africans experience both the male and female biases, Q_{FST} comparing the two shows no evidence of the older female-biased event.

(Figure 3.4) show the relationship between the magnitude (ESR) of male and female biases and the resulting *Q* ratio. Contours in Figure 3.4B and Figure 3.4C show that Q_{π}^{Af} depends primarily on the magnitude of the older female bias while $Q_{FST}^{nA\text{-}Af}$ primarily depends on the magnitude of the recent male bias. Figure 3.3A, however, shows that the more complex pattern of Q_{π}^{nA} is jointly influenced by the magnitudes of both biases.

In summary, our theoretical and simulation results demonstrate that the seemingly contradictory results of Hammer *et al.* and Keinan *et al.* are in fact reconcilable. Q_{π} is well-suited for detecting sex biases in the ancestral human population, so it is probable that the female biases detected by Hammer and colleagues represent a female bias that is a shared legacy of all human populations. Accordingly, long-term sex-biased processes, such as polygyny or higher female dispersal rates in ancestral human populations, likely caused the Q_{π} estimates found by Hammer *et al.* Furthermore, a recent study that compared relative recombination rates on the X chromosome and autosomes found evidence for an ESR greater than 0.5 (female bias) in all three HapMap populations (Labuda, J. Lefebvre, *et al.* 2010; Labuda, J.-F. Lefebvre, *et al.* 2010; Lohmueller *et al.* 2010). These results are consistent with the Q_{π} observations of Hammer *et al.* because, like π, recombination rates detect events along the whole lineage of the human population.

The male bias detected by Keinan *et al.* can be explained by a recent event associated with the Out of Africa dispersal, as initially proposed by the authors. The *Q* ratios detected by Keinan *et al.* suggest a very strong male bias for the entire portion of the non-African lineage before the split of Asians from Europeans. A subsequent study

has shown that a model of continuous male-biased migration from African into non-African populations before the split of Asians and Europeans can account for the magnitude of the previously observed male bias (Keinan and Reich 2010).

In their supplementary analyses, Keinan *et al.* also estimated a Q_π-like measure in shotgun genome sequences from a small number of individuals. Their results were consistent with the pattern they observed for Q_{FST} in that Q is much lower in non-African than in African populations (Keinan *et al.* 2009). Recently, a new study found that regions close to genes have a deficit of X-linked diversity ($Q_\pi < 0.75$) while regions further from genes have an excess of X-linked diversity ($Q_\pi > 0.75$), suggesting that positive selection has had a widespread effect on X-linked genes (Hammer *et al.* 2010). These results are consistent with previous studies that have detected signatures of selection across the X chromosome (Casto *et al.* 2010; Lambert *et al.* 2010). The correlation between Q_π and genetic distance from genes can potentially explain the contrasting results for Q_π, but not the discrepancy between Keinan's Q_{FST} and Hammer's Q_π. We have demonstrated methodological differences between Q_{FST} and Q_π that can account for the majority of this discrepancy, but it remains to be seen what other factors are contributing to the conflicting results such as small sample sizes, different sample populations, and different outgroup species.

More broadly, our results illustrate that complicated demographic models can influence different summary statistics of genetic variation in distinct ways. Thus, evaluating the operating characteristics and behavior of summary statistics under complex demographic models provides important insights into whether different summaries of genetic variation could have been generated by the same evolutionary

forces. These insights will be critical in interpreting the deluge of next-generation sequencing data sets (Li *et al.* 2010) and developing a more comprehensive understanding of human evolutionary history.

3.3 Acknowledgements

We thank Mike Hammer and Alon Keinan for providing data from their previous studies and Alon Keinan, Abigail Bigham, and members of the Akey lab for helpful discussions. We also acknowledge two anonymous reviewers for their thoughtful comments. This research was supported in part by National Institutes of Health grant RO1GM078105 and an NIH/NHGRI Genome Training Grant. J.F.'s participation was funded by National Science Foundation grant 0814322 (P.I. Mary Kuhner) and by "life support" funding from the Department of Genome Sciences.

4. Improving methods for detecting recent positive selection in human populations

4.1 Introduction

Recent advances in DNA sequencing technology and population genetic models have provided more information than ever before about the nature of human genetic variation. As a result, the role of natural selection in shaping patterns of human genetic variation is becoming increasingly clear. There is accumulating evidence that a significant portion of the genome has been the target of natural selection at some point in human evolution (Hahn 2008).

The ultimate goal of population genomic studies of selection is a comprehensive understanding of the alleles that have been the substrates of selection (Akey 2009). Among the questions we seek to answer are: What are the precise causal variants that have been the targets of natural selection? When were they subject to selection? In which human populations were they selected, and in which associated environments? And, perhaps most important and difficult of all, *Why were these alleles selected?* What advantage or disadvantage do they present compared to other alleles at the same locus? A clear understanding of the selective history of a locus satisfies an innate interest in our past, but can also provide information about the evolutionary relationships between human populations, mechanisms of evolutionary change, and the heritable basis of disease.

4.1.1 Candidate gene studies of positive selection

The first DNA studies presenting empirical evidence for the action of positive selection in the human genome examined single genes with a previous hypothesis for selection. The most notable example is that of *LCT*, the gene encoding the enzyme lactase, which digests the sugar lactose (Figure 4.1). In mammals, the expression of *LCT* is down-regulated after weaning; thus, lactase is only produced when a juvenile is dependent on a milk diet (Harris and Meyer 2006). This is also the ancestral state in humans, but adaptive regulatory mutations have arisen in pastoral populations. These adaptive alleles produce a lactase persistence phenotype, such that lactase production continues into adulthood and allows the digestion of dairy products such as milk, yogurt, and cheese.

Examinations of genetic variation around *LCT* show evidence for an incomplete (or in-progress) selective sweep, including long high-frequency haplotypes carrying the causative allele (*i. e.* these haplotypes show very low variation and very high LD for their young age, indicating they rose to a high frequency very quickly) (Figure 4.1b) (Bersaglieri *et al.* 2004; Tishkoff *et al.* 2007; Enattah *et al.* 2008) and high population differentiation (Figure 4.1c). Meanwhile, other haplotypes lacking an advantageous allele exhibit normal levels of variation (Tishkoff *et al.* 2007). Additionally, examination of various populations has identified several different causative lactase-persistence alleles (Enattah *et al.* 2002; Tishkoff *et al.* 2007; Enattah *et al.* 2008; Ingram *et al.* 2009; Gallego Romero *et al.* 2012; Jones *et al.* 2013) in an upstream intron of the gene *MCM6*, which shows enhancer activity (Olds and Sibley 2003).

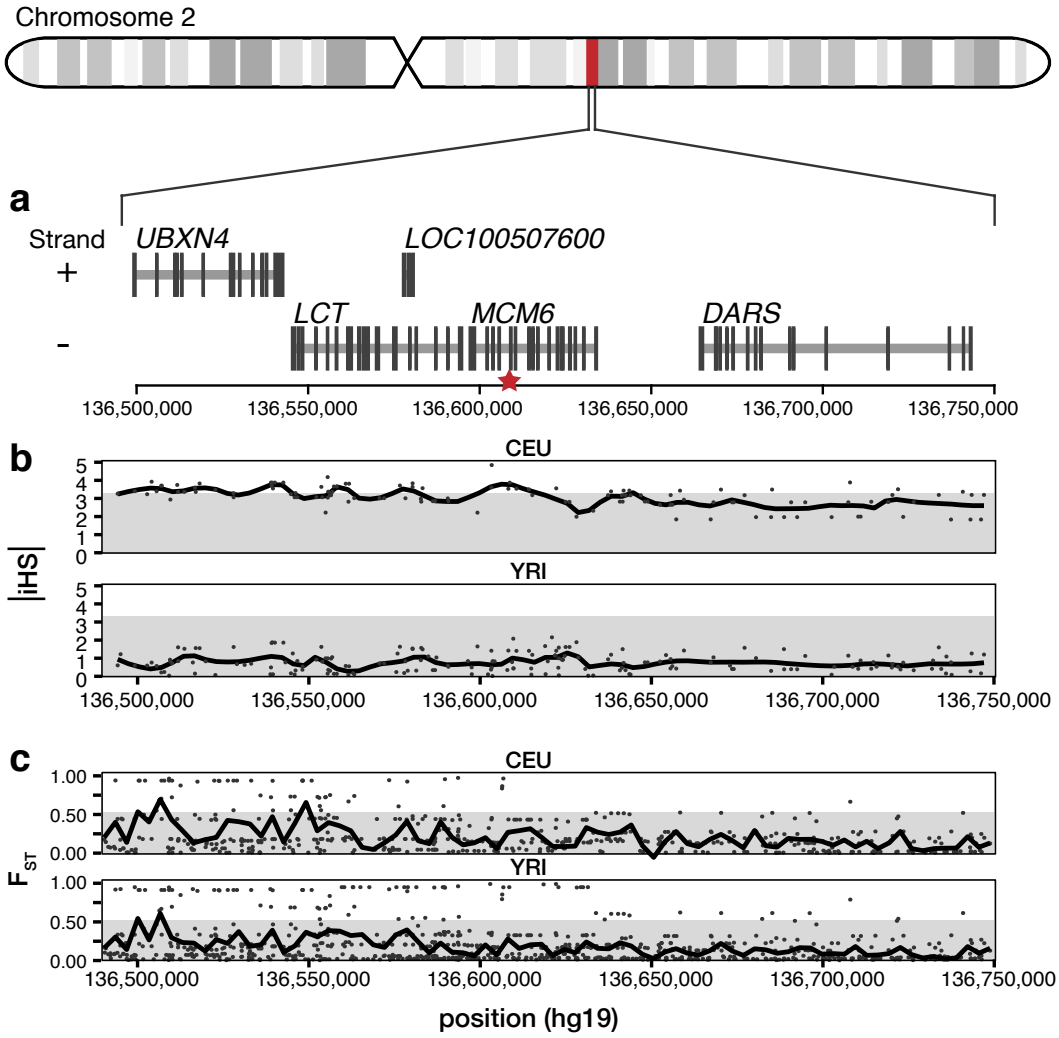


Figure 4.1 Signal of selection surrounding the LCT locus.
(a) LCT and nearby RefSeq genes on chromosome 2. Genes on the top line are on the + strand and genes on the bottom line are on the – strand. The red star marks the location of rs4988235, one of the polymorphisms in an enhancer region that is associated with lactase persistence (Bersaglieri *et al.* 2004). (b) The absolute value of the integrated haplotype score (iHS) for SNPs from the HapMap phase 2 data in two populations, from Voight *et al.* (Voight *et al.* 2006) Populations are (CEU): Utah residents with Northern and Western European ancestry and (YRI): Yoruba in Ibadan, Nigeria. The black line is a loess-smoothed curve of the plotted points. The gray box marks the 90% confidence interval of all iHS scores in the region, which is exceeded by variants in the enhancer region in the CEU. (c) F_{ST}, a measure of population differentiation, for variants from the 1,000 Genomes project phase 1 data. Populations are the same as above. The black line is a loess-smoothed curve of the plotted points. The gray box marks the 80% confidence interval of all F_{ST} measures in the region.

Although there are a handful of other well-supported examples of positive selection acting on identified advantageous alleles, the scope, throughput, and utility of such studies is limited. To be able to identify a candidate gene for positive selection requires detailed information on the genotype-phenotype relationship and some *a priori* hypothesis regarding why, when, and where selection acted on the gene (Akey 2009). This is an even more difficult prospect for advantageous regulatory alleles, which we still know relatively little about. Additionally, interpreting patterns of variation for a single gene at a time is difficult because of the complicating factors introduced by human demographic history (Akey *et al.* 2004; Stajich and Hahn 2005; Akey 2009; Li *et al.* 2012).

4.1.2 Insights from genome-wide scans for positive selection

In contrast, genome-wide scans provide an unbiased way to search for unsuspected targets of positive selection and also provide more rigorous approaches to distinguish the marks of selection from those of neutral demographic events. The advent of dense genome-wide SNP genotype data has made these scans possible within the past decade. Despite concerns that the signal of natural selection would be drowned out by genomic noise, early studies showed that selection's mark on human genome variation could be detected from sequence data alone (Akey *et al.* 2002). To date, scans for selection have utilized a wide variety of test statistics to detect unusual patterns of variation that are indicative of selection. These test statistics examine features such as population differentiation (Akey *et al.* 2002; Chen *et al.* 2010); decay of LD around a SNP (Sabeti *et al.* 2002; Voight *et al.* 2006); comparative LD (Sabeti *et al.* 2007), derived allele frequency (Grossman *et al.* 2010), or homozygosity between

populations (Zhong *et al.* 2010; Zhong *et al.* 2011); high frequency derived alleles (Fay and Wu 2000); an excess of rare variants (Tajima 1989); and differences between population- and pedigree-based recombination rate estimates (O'Reilly *et al.* 2008). Some of these statistics perform well for detecting completed selective sweeps, while others are appropriate for in-progress or incomplete sweeps. They also exhibit a wide range of sensitivities and specificities (Ronald and Akey 2005; Biswas and Akey 2006; Sabeti *et al.* 2007). While the effects of selection and demography are difficult to distinguish, a reasonable way to address this problem is to employ an outlier approach, which assumes that demographic events will affect variation throughout the genome, while selective events act at individual loci (Cavalli-Sforza 1966). Advantageous alleles can be detected by identifying loci that exhibit anomalous patterns of variation compared to the rest of the genome and are therefore outliers in the distribution of a statistic of interest (Akey *et al.* 2002).

4.1.3 Improving genome-wide scans

In this chapter, I will describe my work on improving current methods for genome-wide scans to identify the signatures of positive selection on a large scale. My objectives are 1) to characterize the performance of neutrality test statistics for detecting selection from standing variation and 2) to develop a genome-wide approach for combining signals across multiple neutrality test statistics.

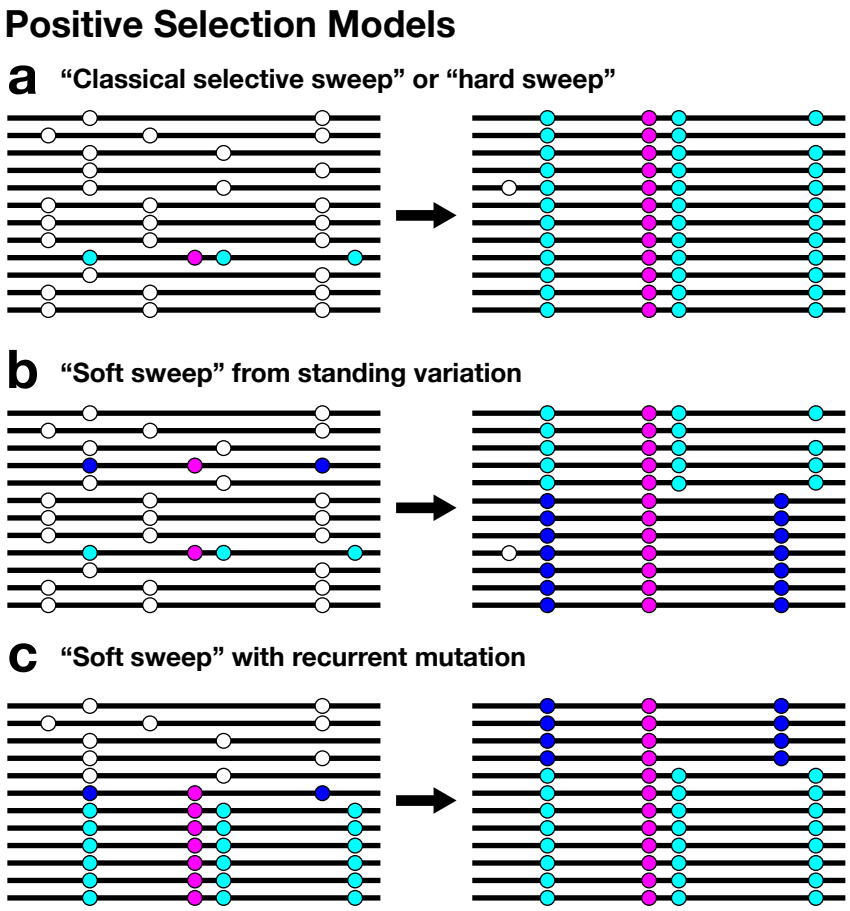


Figure 4.2 Classical and alternative models of positive selection.
Pink: causal variant under positive selection; Cyan: hitch-hiking variants; Blue: hitch-hiking variants on a second haplotype background; Red: epistatic variant that affects the fitness of the selected variant. Left and right panels denote patterns of variation before and after selection, respectively. (a) A new mutation arises and is immediately advantageous. (b) An existing variant with multiple haplotype backgrounds is newly advantageous due to a novel selective pressure. (c) During a selective sweep, the advantageous variant arises again due to a mutation on a second haplotype background.

The usefulness of previous genome-wide scans for recent selection is undermined by the simplicity of the classic selective sweep model, upon which most neutrality test statistics are based. In a classic selective sweep (“hard sweep”), a new mutation is immediately beneficial upon its introduction into the population and quickly reaches fixation, sweeping along linked neutral variants (Figure 4.2a). Some

approaches also exist to detect the signatures of incomplete or in-progress sweeps (Voight *et al.* 2006). Selection may also occur by "soft sweeps", in which a beneficial allele exists on multiple haplotype backgrounds before sweeping to fixation (Hermisson and Pennings 2005). A soft sweep may occur because the beneficial allele was a previously neutral variant segregating in the population (from standing variation) or because the beneficial allele arose multiple times (by recurrent mutation) (Figure 4.2). The model of a soft sweep from standing variation is intuitively appropriate for recent human evolution, which has been characterized by adaptation to novel environments (Hermisson and Pennings 2005; Pritchard *et al.* 2010; Hernandez *et al.* 2011). It is unclear how appropriate current neutrality test statistics are for detecting the signatures of a soft sweep (Przeworski *et al.* 2005; Pritchard *et al.* 2010; Hernandez *et al.* 2011). A newly-beneficial allele undergoing a soft sweep does reach fixation quickly and drag along linked neutral variants, but because the allele had time to accumulate different haplotype backgrounds, the signature of the hitch-hiking effect is more difficult to detect than in a classic hard sweep (Figure 4.2b) (Hermisson and Pennings 2005; Przeworski *et al.* 2005). To address this objective, I will investigate the ability of existing neutrality test statistics to detect the signatures of positive selection in both hard sweep and soft sweep models.

It is also imperative to develop methods for reconciling the conflicting signals produced by various neutrality test statistics. Previous work has developed a composite likelihood method to combine the information from five different neutrality test statistics, but applied this method to candidate regions identified by previous genome-wide scans with a single neutrality test statistic (Grossman *et al.* 2010;

Grossman *et al.* 2013). Methods that combine information from multiple test statistics also have the potential to increase the resolution of genome-wide scans for positive selection. The use of whole genome sequences for detecting selection has the potential to identify the causative variants underlying signatures of positive selection (Tennessen *et al.* 2010; Tennessen *et al.* 2012). To address this objective, I will develop an approach for combining information from multiple signals in a genome-wide scan for selection and investigate which combinations of neutrality test statistics provide the most information.

4.2 Detecting the signatures of selection from standing variation

4.2.1 Materials and Methods

4.2.1.1 Coalescent simulations

To assess the performance of existing neutrality test statistics under known conditions, I used the coalescent simulation programs `ms` (Hudson 2002) and `msms` (Ewing and Hermisson 2010) to generate simulated population genetic data. I used a model of human evolution in which a non-African population split off from an African population (see Figure 4.3) 3,500 generations ago. I used model parameters that were calibrated to match the allele frequency spectrum observed ASN, CEU, and YRI population in the HapMap phase 2 SNP data (Schaffner *et al.* 2005). The model included mutation, migration, recombination, bottlenecks, and rapid recent population growth. Using this model, I simulated 1 Mb of sequence data 100 times for each of the 29 parameter combinations in Table 4.1, for a total dataset consisting of 2,900 Mb of simulated sequence data in 200 sampled chromosomes. The full command line for each combination of parameters is presented in Table C.1.

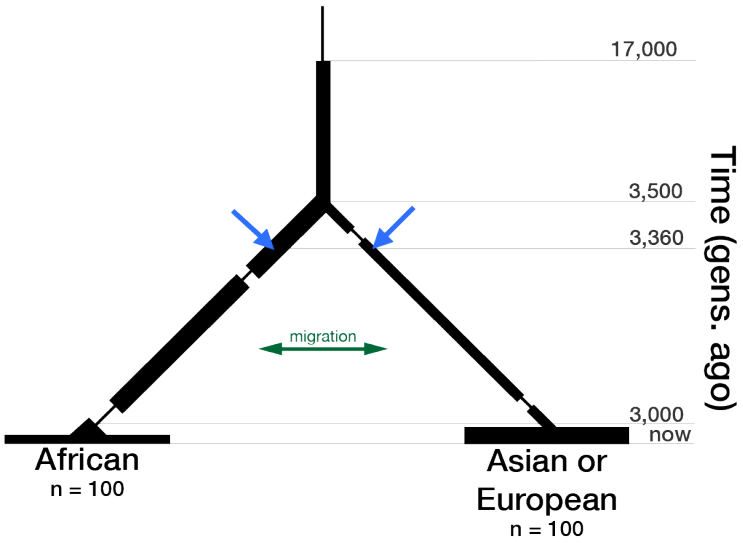


Figure 4.3 Coalescent model used in simulations
Selection was introduced into the model at one of the two blue arrows. Branches of the lineage indicate relative changes in population size, but are not to-scale. Sample sizes (n) are the total number of chromosomes simulated. The y-axis (time) is not to-scale.

For simulations including positive selection, I used the -Sc and -SI options in `msms` to specify positive selection in a single population beginning 3,360 generations ago and continuing forwards in time. The beneficial allele, located in the center of the simulated sequence, had a dominance coefficient of 0.50 and a rate of recurrent mutation that matched the overall mutation rate. The program `msms` uses forward simulations to generate the frequency trajectory of the beneficial allele in each population, and then runs coalescent simulations according to the ancestral recombination graph, conditioned on the beneficial allele frequency. I simulated beneficial alleles with selection coefficients (*s*) of 0.001, 0.01, and 0.1. To consider varying models of a soft sweep from standing variation, I specified the starting frequency (*f*) of the beneficial allele, including values of 0.05, 0.1, and 0.5. To account for the proportion of the time that the beneficial allele was lost due to drift, I kept only those simulations for which the beneficial allele had reached a frequency of at least 80%. To achieve this, I ran simulations for each parameter set until I had collected 100 datasets that satisfied this beneficial allele frequency cutoff. To simulate hard sweep

models, I set $f = 0$ and applied the same beneficial allele frequency filter. Some parameter combinations did not generate any datasets that met the beneficial allele frequency cutoff, and were therefore excluded from further analyses.

Table 4.1 Parameters used in coalescent simulations

Population experiencing selection	f b	s a	Population experiencing selection	f b	s a
African	0	0.1	Asian	0.5	0.1
Asian	0	0.1	European	0.5	0.1
European	0	0.1	African	0.5	0.01
Asian	0	0.01	Asian	0.5	0.01
African	0.05	0.1	European	0.5	0.01
Asian	0.05	0.1	African	0.5	0.001
European	0.05	0.1	Asian	0.5	0.001
African	0.05	0.01	European	0.5	0.001
Asian	0.05	0.01	neutral	-	0
European	0.05	0.01	neutral	-	0
African	0.1	0.1	Africanc	0	0.001
Asian	0.1	0.1	Asian	0	0.001
European	0.1	0.1	European	0	0.001
African	0.1	0.01	African	0.05	0.001
Asian	0.1	0.01	Asian	0.05	0.001
European	0.1	0.01	European	0.05	0.001
Asian	0.1	0.001	African	0.1	0.001
European	0.1	0.001	European	0.01	0.01
African	0.5	0.1	African	0.01	0.01

a: *s* = selection coefficient
b: *f* = frequency of the beneficial allele at the time that positive selection starts
c: cells shaded in red indicate parameter combinations that failed to meet the beneficial allele frequency cutoff

4.2.1.2 Calculating established neutrality test statistics

I selected 13 neutrality test statistics to investigate. These statistics, presented in Table 4.2, have been used widely in recent genome-wide scans for positive selection in humans.

Table 4.2 Neutrality test statistics examined in this study

Statistic	Other names	Full name	Category	Reference
H_{12}			Haplotype structure & linkage disequilibrium	(Garud *et al.* 2013)
H_2/H_1			Haplotype structure & linkage disequilibrium	(Garud *et al.* 2013)
$\Delta iEHH_D$	ΔiHH	Difference in integrated Extended Haplotype Homozygosity (iEHH) for derived haplotypes	Linkage disequilibrium	(Grossman *et al.* 2010)
$iEHH_S$	iES	Site-specific iEHH	Linkage disequilibrium	(Tang *et al.* 2007)
iHS		Integrated haplotype score; allele-specific iEHH	Linkage disequilibrium	(Voight *et al.* 2006)
ΔDAF		Difference in derived allele frequency	Population differentiation	(Grossman *et al.* 2010)
F_{ST}		Weir and Cockerham's F_{ST}	Population differentiation	(Wright 1950; Weir 1996)
$xp\text{-}iEHH_S$	$\ln(R_{sb})$	Cross-population site-specific iEHH	Population differentiation & linkage disequilibrium	(Tang *et al.* 2007)
$xp\text{-}iEHH_D$	XP-EHH	Cross-population allele-specific iEHH	Population differentiation & linkage disequilibrium	(Voight *et al.* 2006)
H_{FW}		Fay and Wu's H	Site frequency spectrum	(Fay and Wu 2000)
D*		Fu and Li's D*	Site frequency spectrum	(Fu and Li 1993)
F*		Fu and Li's F*	Site frequency spectrum	(Fu and Li 1993)
D_{Tajima}		Tajima's D	Site frequency spectrum	(Tajima 1989)

I developed a custom pipeline consisting of several `Python` modules for the analysis of simulated sequence data in order to calculate each of these test statistics on my simulated datasets. To implement functions for each statistic, I read the primary source and tested the functions for accuracy against examples in the literature whenever possible.

I defined common variant sites as having a derived allele frequency greater than 1% and calculated each statistic at every common variant site, within each population in a simulated dataset. For site frequency spectrum-based statistics (H_{FW}, D*, F*, & D_{Tajima}), I calculated an average value for the statistic within a 10 kb window around

each common site, including rare sites in the calculation. For linkage disequilibrium-based tests ($\Delta iEHH_D$, $iEHH_S$, iHS, xp-$iEHH_S$, & xp-$iEHH_D$), I calculated each statistic using each common site as a core site and excluded all rare variant sites from the calculations. I took the absolute value of xp-$iEHH_S$ and xp-$iEHH_D$ to account for differentiation of either population being considered. For haplotype structure statistics (H_{12} & H_2/H_1), I classified haplotypes in 10 kb windows around a site, using each common variant as the core site. I calculated F_{ST} and ΔDAF at every variant site.

LD-based tests (H_{12}, H_2/H_1, $\Delta iehh_D$, $iEHH_S$, iHS xp-$iEHH_S$ xp-$iEHH_D$) are highly dependent on the allele frequency of the variant at the core site. To account for this, I standardized each of these statistics based on the derived allele frequency of the variant at the core site, as is standard practice (Voight *et al.* 2006; Grossman *et al.* 2010). To accomplish this, I used data from the neutral simulation models. I separated the variant sites based on their frequency, in frequency bins of 2.5%. I then calculated the mean and the standard deviation of each LD-based statistic. For each LD-based statistic, I obtained the standardized value by subtracting the mean and dividing by the s.d. for variants of the same allele frequency, according to the following equation:

$$standardized(s_p) = \frac{s_p - mean(n_p)}{s.d.(n_p)} \qquad \textbf{Eq. 4.1}$$

where s_p is the estimated value of the statistic at a variant site with allele frequency p and n_p is the set of all variant sites in the neutral data with allele frequencies in the 2.5% bin including p.

4.2.1.3 Developing neutrality test statistics using haplogroup structure

A previous candidate gene study of *ALMS1* used the *ad hoc* statistic d_{xy} to identify the genetic signature of a soft sweep from standing variation (Scheinfeldt *et al.* 2009). This statistic measured the average pairwise nucleotide diversity between two deeply diverged haplogroups, each bearing the derived allele that was purported to be under selection. To assess the applicability of this statistic to genome-wide scans for positive selection, I developed methods for calculating d_{xy} at any variant site in an automated manner. To calculate d_{xy} at a variant site, I examined a window of 10 kb around a core variant site and selected those haplotypes carrying the derived allele at the core site. Using these derived haplotypes, I calculated a distance matrix using a simple measure of distance assuming two alleles (derived and ancestral) and no recurrent mutation. I then used this distance matrix as input in the program `neighbor` from the package `PHYLIP` (v. 3.6) (Felsenstein 2005). Using the tree generated by `neighbor`'s, UPGMA method, I identified the deepest split in the tree and used this split to classify the derived haplotypes into two derived haplogroups. Finally, I calculated the average pairwise nucleotide diversity between these two derived haplogroups. d_{xy} is described by the following equation, where d_{ij} is the number of nucleotide differences between haplotypes i and j and D_1 and D_2 are the set of derived haplotypes in the two respective haplogroups:

$$d_{xy} = \frac{1}{i \times j} \sum_{i \in D_1} \sum_{j \in D_1} d_{ij} \qquad \textbf{Eq. 4.2}$$

To determine if another similar test might be more informative, I used the same haplogroup classification procedure as described above, but instead calculated d_{12},

which is the absolute difference in Tajima's D between the two haplogroups, according to the following equation:

$$d_{12} = \left| D^1_{Tajima} - D^2_{Tajima} \right| \quad \textbf{Eq. 4.3}$$

where D^k_{Tajima} is the mean Tajima's D calculated in a 10 kb window around the core site in haplogroup *k*. I calculated d_{12} and d_{xy} for core sites at every common variant in the simulated datasets, and included rare variants in their calculation.

4.2.1.4 Calculating empirical p values

To identify the variants under selection in my simulated datasets, I calculated the empirical p value for each statistic at each site. I used the empirical cumulative distribution function of each statistic in the neutral simulations to calculate these p values. I transformed all of the p values to an upper-tailed p value (p_u) so that they are all comparable to one another. For the lower-tailed tests H_{12}, H_2/H_1, d_{xy}, d_{12}, H_{FW}, D*, F*, and D_{Tajima}, I subtracted the lower-tailed p value from 1 according to the equation $p_u = 1 - p_l$. For the two-tailed tests iHS and ΔDAF, I used the formula $p_u = |p_l - 0.5| \times 2$ to obtain an upper-tailed p value. Each empirical p value is a measure of how extreme the statistic is compared to the distribution of that statistic in the neutral simulations.

4.2.1.5 Calculating binned means

In order to plot the empirical p values from each statistic, it was necessary to reduce the dataset from the order of millions of points to thousands of points. I calculated binned means for empirical p values for each statistic in 10 kb bins across the length of the simulated sequences, within each set of simulation parameters and within each population. I used these binned mean values to identify peaks that indicate a signature of selection. For each bin, I also calculated the upper 95th percentile of the

statistic, as an estimate of the 95% confidence interval for the value of the statistic within that bin.

4.2.2 Results

The binned mean empirical p values for each neutrality test statistic are presented in Figure 4.4, Figure 4.5, Figure 4.6, & Figure 4.7 and supplemental Figure C.1, Figure C.2, Figure C.3, Figure C.4, Figure C.5, Figure C.6, Figure C.7, Figure C.8, Figure C.9, Figure C.10, and Figure C.11. Examining the peaks in these figures in detail, there were several interesting observations. First of all, many of these individual statistics (*e. g.* D_{Tajima}, D*, F*) were very noisy, exhibiting a wide peak around the selected site and very wide confidence intervals. Some statistics had very high peaks around the selected site (*e. g.* H_{FW}, xp-$iEHH_S$), which were very strong signals for the signatures of selection. Others had shorter peaks around the selected site, but those peaks were often more distinct from the background (*e. g.* ΔDAF, iHS, and Δ$iEHH_D$). To my knowledge, these results were the first direct comparison of so many of these neutrality test statistics to one another under known conditions. Using these results, I could start to pick out neutrality test statistics that performed well over a wide range of selection models. For example, H_{FW} had a very high peak around the selected site in all but one selection model (Figure 4.4). In contrast, F_{ST} had a very short peak around the selected site, in particular in the simulated Asian and European populations (Figure 4.5). There were also several models of selection for which F_{ST} exhibited no appreciable peak.

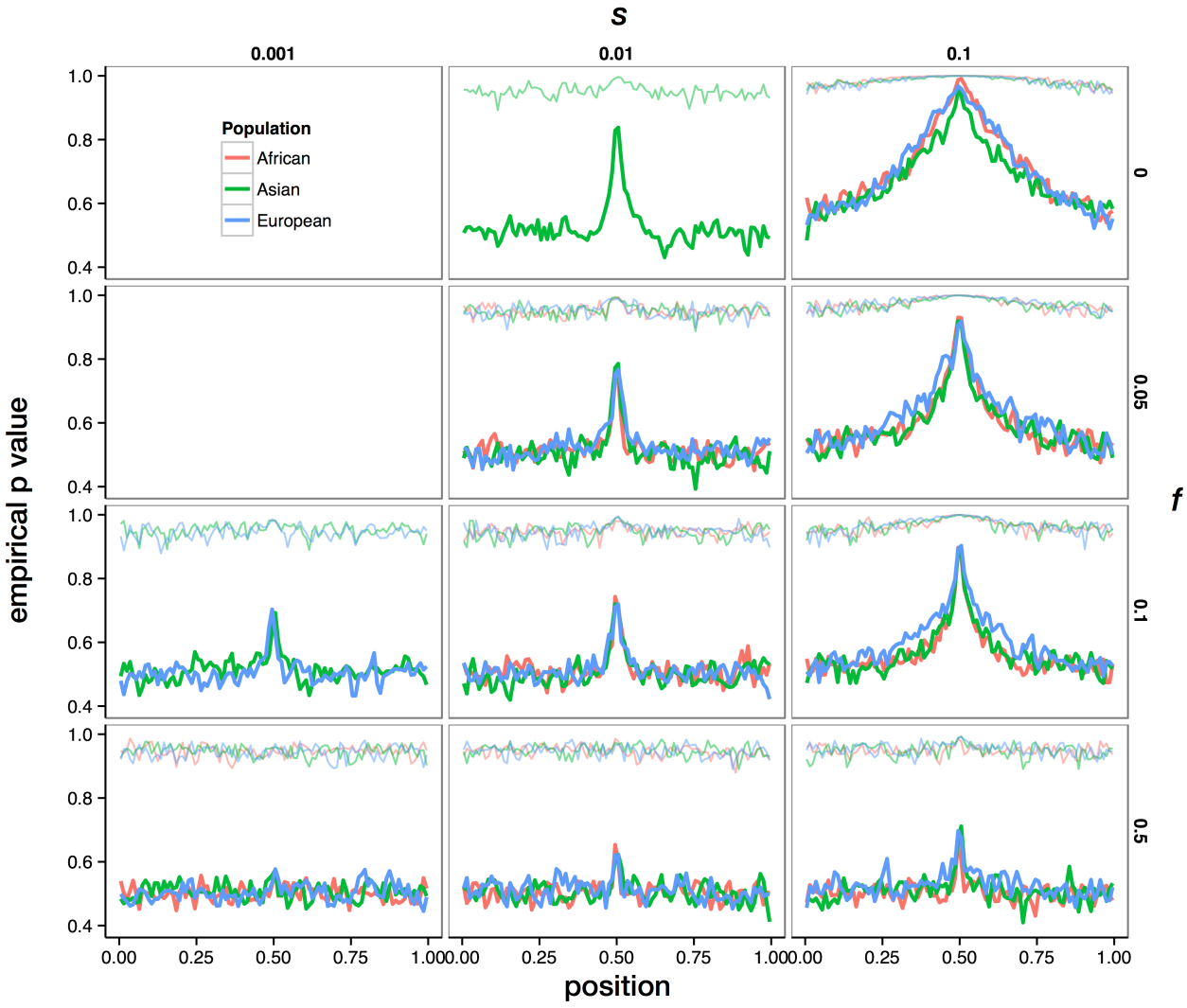


Figure 4.4 Binned mean empirical p value for H_{FW} depends on *s* and *f* in simulations including selection

For each combination of *s* and *f*, the average empirical p value binned by position is plotted in thick, darker lines. Thin, lighter lines indicate the binned 95% confidence interval. Simulations including selection in each of three populations are shown in each panel where available. Populations are colored according to the legend at top left.

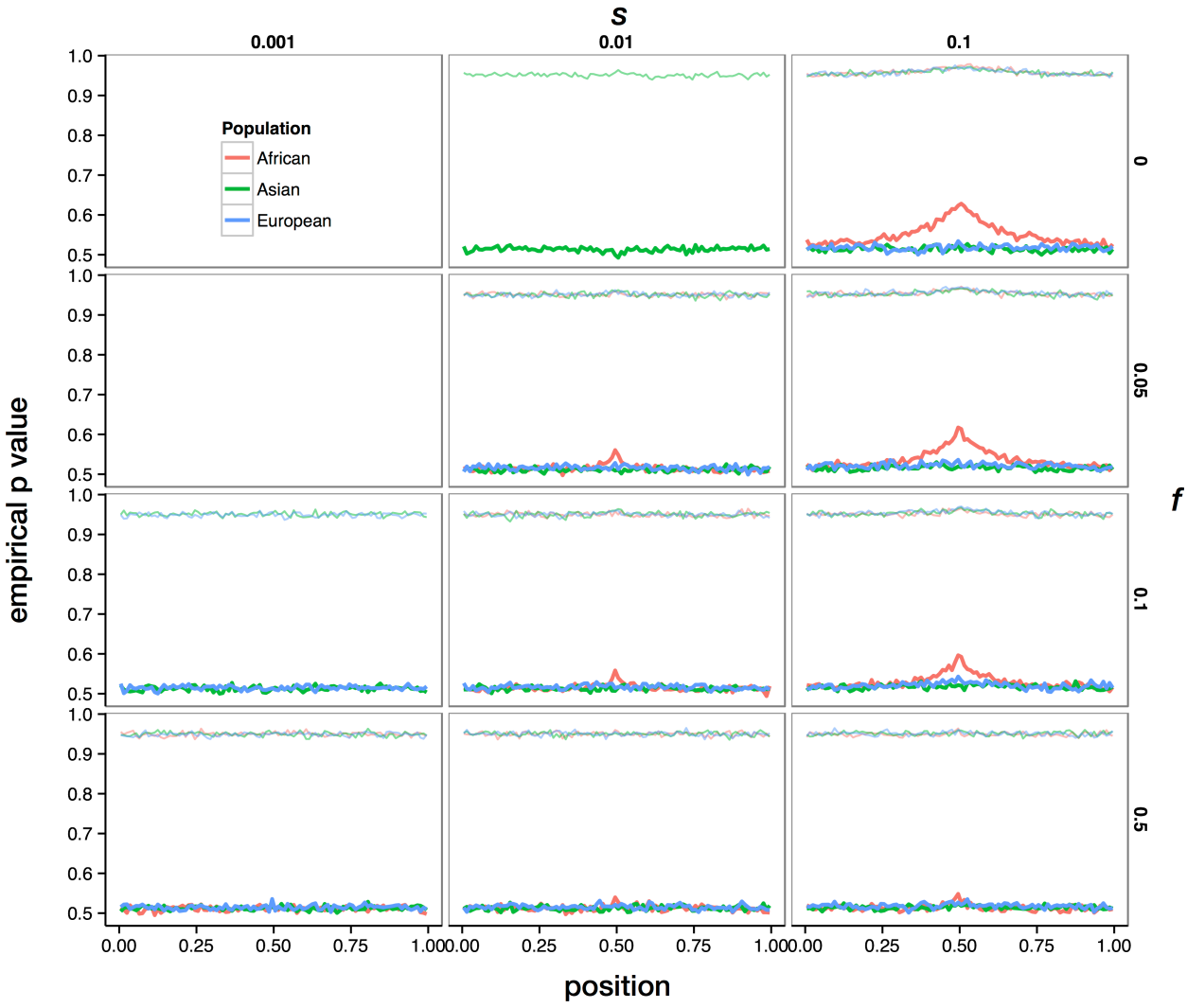


Figure 4.5 Binned mean empirical p value for F_{ST} depends on *s* and *f* in simulations including selection

For each combination of *s* and *f*, the average empirical p value binned by position is plotted in thick, darker lines. Thin, lighter lines indicate the binned 95% confidence interval. Simulations including selection in each of three populations are shown in each panel where available. Populations are colored according to the legend at top left.

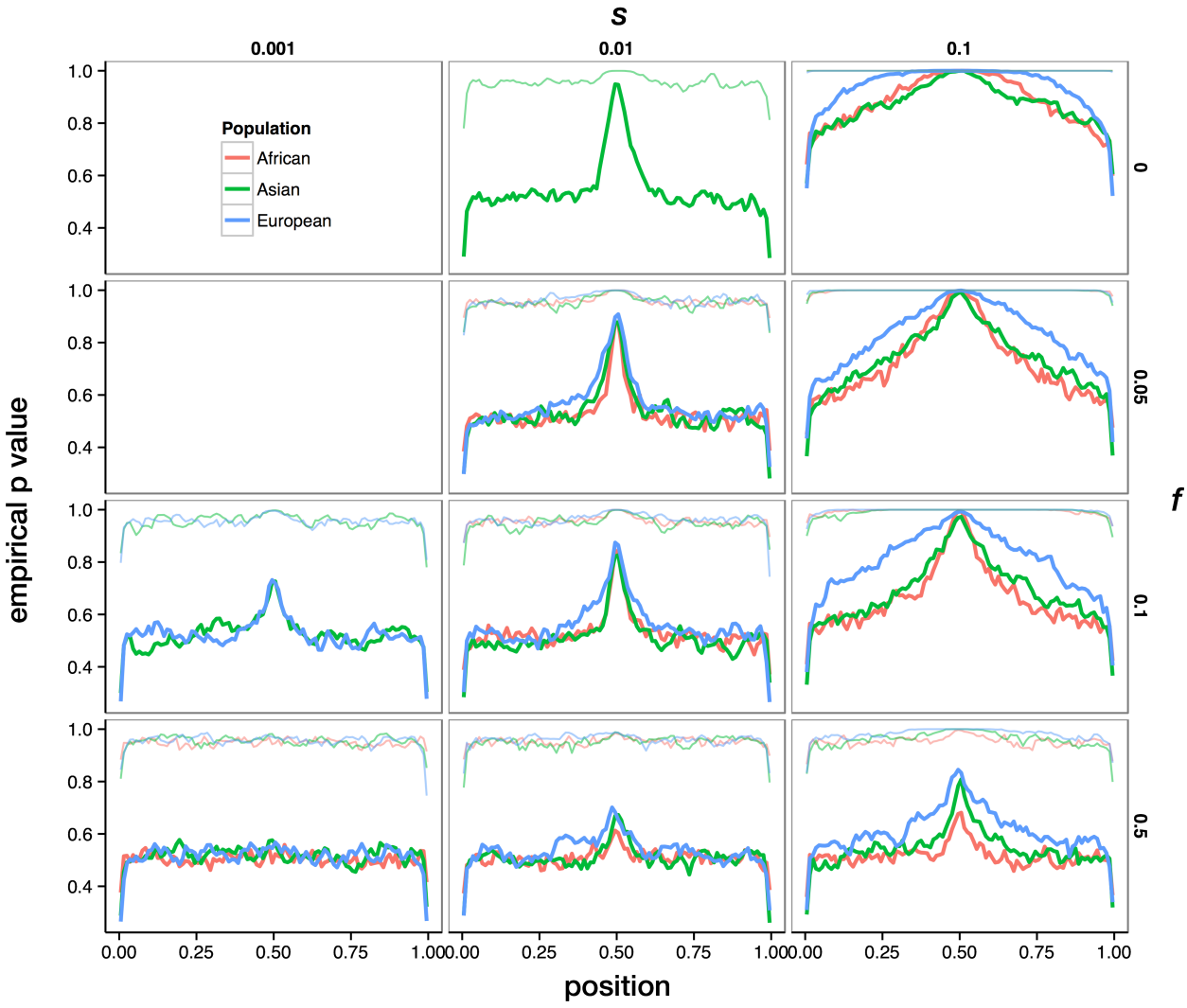


Figure 4.6 Binned mean empirical p value for iEHH$_S$ depends on *s* and *f* in simulations including selection

For each combination of *s* and *f*, the average empirical p value binned by position is plotted in thick, darker lines. Thin, lighter lines indicate the binned 95% confidence interval. Simulations including selection in each of three populations are shown in each panel where available. Populations are colored according to the legend at top left.

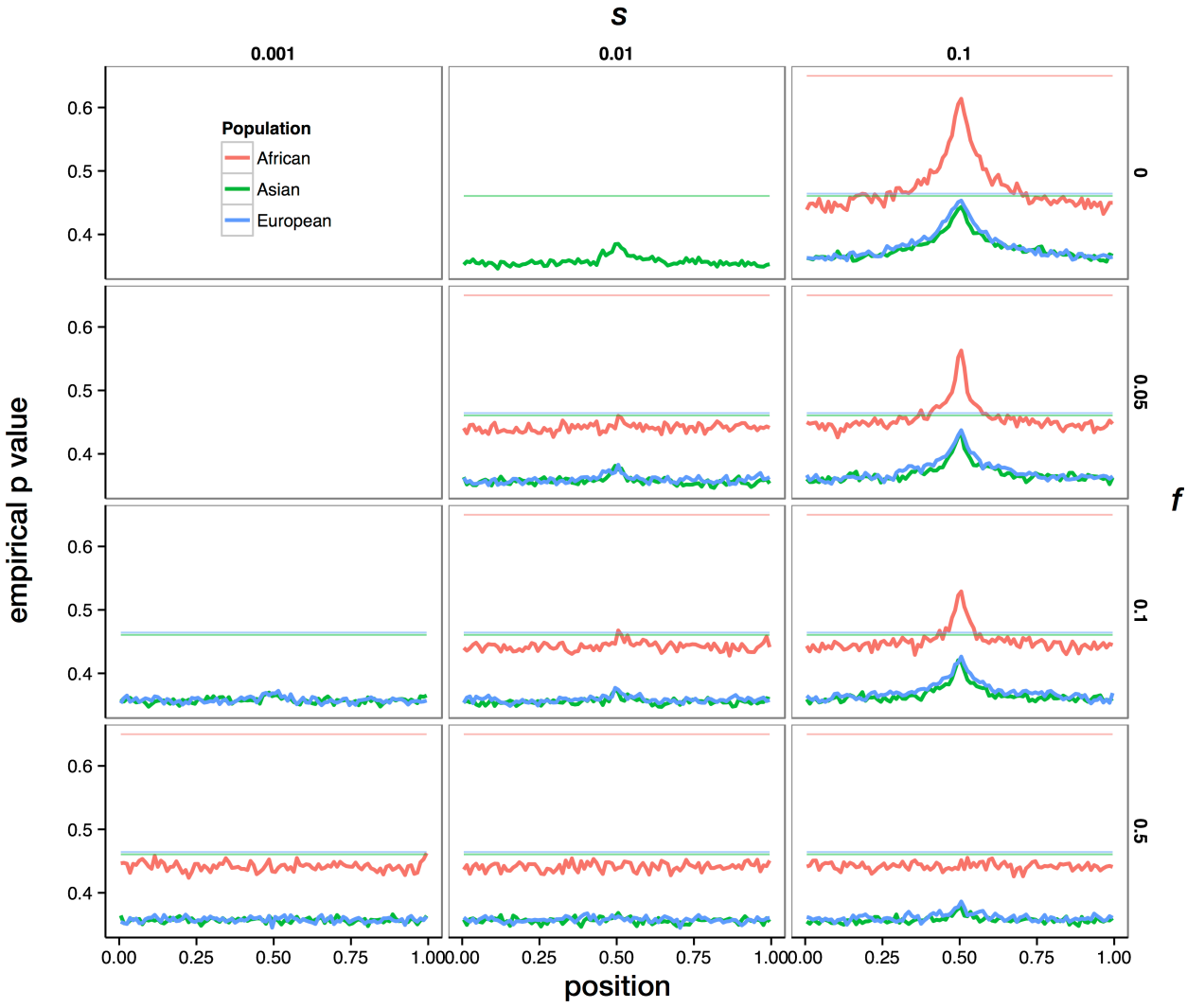


Figure 4.7 Binned mean empirical p value for d_{xy} varies little with *s* and *f* in simulations including selection

For each combination of *s* and *f*, the average empirical p value binned by position is plotted in thick, darker lines. Thin, lighter lines indicate the binned 95% confidence interval. Simulations including selection in each of three populations are shown in each panel where available. Populations are colored according to the legend at top left.

Additionally, several neutrality test statistics exhibited strong signals of positive selection in models of soft sweeps from standing variation, despite being designed for detecting classical hard sweeps. These statistics had high peaks in mean empirical p values around the selected site even for low selection coefficients and high starting allele frequencies. For example, xp-iEHH$_S$, D$_{Tajima}$, iEHH$_S$, and ΔiEHH$_D$ all had noticeable peaks at the selected site in every model of selection tested except for those with s = 0.001 and f = 0.5. The statistic H$_{12}$, which was recently developed specifically for the purpose of detecting the signatures of soft sweeps, similarly showed peaks at the selected site in all models of selection tested except for those with s = 0.001 and f = 0.5.

Another striking result was the difference in the strength of the signatures of selection between simulated African and non-African populations. In particular, the peak in the mean empirical p value of D$_{Tajima}$ was noticeably lower than the simulated non-African populations, while peaks in ΔDAF and F$_{ST}$ were higher in the simulated Africans. Furthermore, when the selection coefficient was high (s = 0.1), the simulated African population had lower peaks in the mean empirical p value of F*, iHS, and ΔiEHH$_D$.

As I've calculated it here, d_{xy} did not perform well in detecting the signatures of soft sweeps from standing variation. In fact, d_{xy} performed better for detecting the signatures of a hard sweep than those of a soft sweep (Figure 4.7). Furthermore, the overall performance of d_{xy} did not compare favorably to the other statistics that I investigated. The automated derived haplogroup classification procedure I developed to apply d_{xy} in a genome-wide scan could be altered and improved to determine if there

are other ways of calculating this statistic. d_{12} also performed poorly in detecting the signatures of a soft sweep from standing variation (Figure C.11) and performed poorly overall compared to other neutrality test statistics. This was another indication that the haplogroup classification method should be redesigned.

4.3 Combining information from multiple tests for positive selection

4.3.1 Materials and Methods

4.3.1.1 The Cumulative Selection Score (CSS)

A previous approach, the Composite of Multiple Signals (CMS), was developed to identify overlapping signatures of positive selection from multiple neutrality test statistics (Grossman *et al.* 2010; Grossman *et al.* 2013). The CMS at site *j* can be calculated from *n* neutrality test statistics (s_i) using the following equation:

$$CMS_j = \prod_{i=1}^{n} \frac{P(s_i|selected)\times Prior(selection)}{P(s_i|selected)\times Prior(selection) + P(s_i|neutral)\times Prior(neutral)} \quad \textbf{Eq. 4.4}$$

CMS can be interpreted as the posterior probability that site *j* is a target of positive selection. The probability of a given score s_i under neutrality or selection is obtained from simulations, and CMS is therefore wholly dependent on the specification of appropriate models of human demography under neutrality and selection. Grossman *et al.* implemented CMS on candidate regions identified in previous genome-wide scans, meaning it has not been implemented in a genome-wide scan itself.

To improve upon this methodology, I developed the Cumulative Selection Score (CSS), which is the product of the empirical p values for *n* neutrality test statistics at a site, *j*:

$$CSS_j = \prod_{i=1}^{n} \Pr(s_i | data) \qquad \textbf{Eq. 4.5}$$

When all of the empirical p values for a site are transformed to upper-tailed p values, CSS is highest when all of the component p values are also high. To assess the performance of CSS under known conditions, I used the simulated sequence datasets described in section 4.2.1.1.

4.3.1.2 Identifying correlations between neutrality test statistics

CSS relies on the assumption that each component statistic is independently distributed under neutrality. If this assumption holds, the co-occurrence of high empirical p values in the component statistics can be attributed to selection. If, however, the random variation in the component statistics is also correlated, then high CSS values would result from random chance. To test this assumption, I calculated the Spearman's rank correlation coefficient (ρ) between each pair of neutrality test statistics in neutral simulated datasets using the R package `Hmisc` (Figure 4.8).

The correlation coefficients presented in Figure 4.8 fall into three major groups: population differentiation statistics (ΔDAF, F_{ST}, d_{12}, & d_{xy}); linkage disequilibrium statistics (H_{12}, H_2/H_1, $\Delta iEHH_D$, $iEHH_S$, iHS, xp-$iEHH_S$, & xp-$iEHH_D$); and site frequency spectrum statistics (H_{FW}, D*, F*, D_{Tajima}). Unsurprisingly, these correlated groups correspond with the basic categories that define these statistics (Table 4.2). These correlation coefficients can guide the selection of neutrality test statistics to include in the calculation of CSS. For example, one could include one statistic from each of the three categories, or the two least correlated statistics from each category, *etc.*

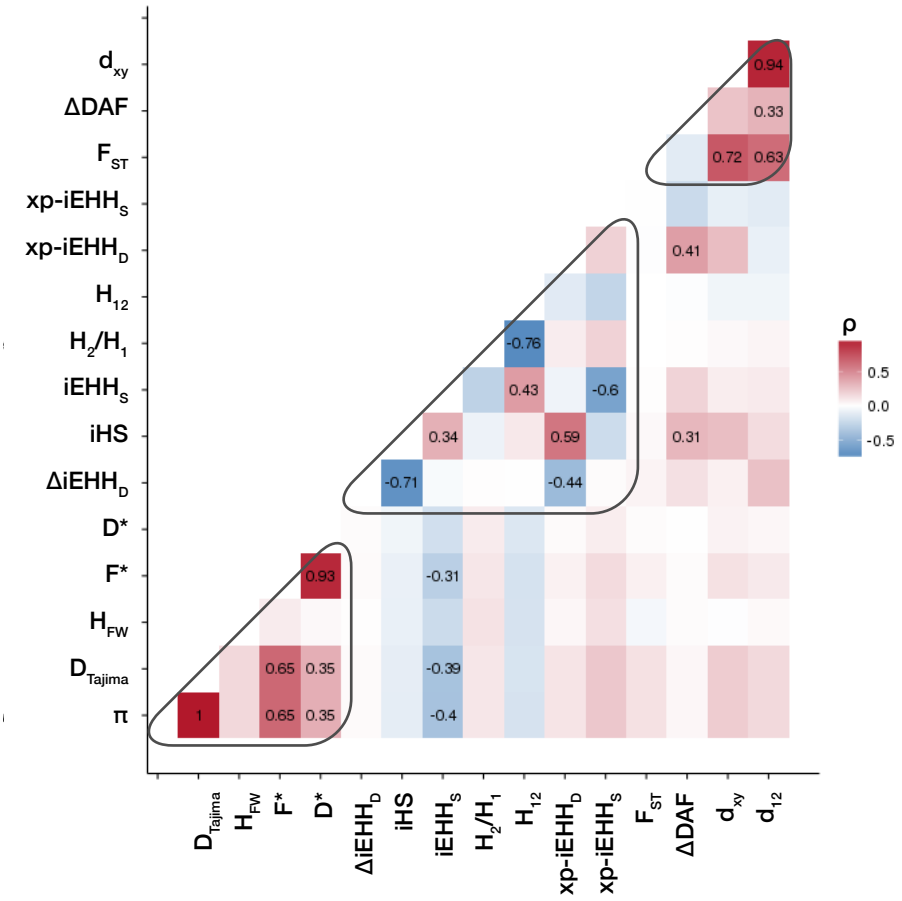


Figure 4.8 Neutrality test statistics fall into three major groups based on correlations with one another in a neutral simulation
Each square in row *i*, column *j*, is colored by the value of the Spearman's rank correlation coefficient, ρ, between statistics *i* and *j*. The statistics fall into three groups, which are mostly correlated within, but not between, groups. These three groups are marked with gray outlines.

4.3.1.3 Identifying optimal combinations of statistics to include in calculating the CSS

The effectiveness of CSS in detecting the signatures of positive selection is largely dependent on the component statistics used to calculate it. To assess the usefulness of various combinations of statistics, I calculated several dozen different variations of CSS. I chose each combination of statistics qualitatively based on the height and definition of peaks in the binned mean empirical p values from simulated

data (see Figure 4.4, Figure 4.5, Figure 4.6, Figure 4.7, and supplemental Figure C.1, Figure C.2, Figure C.3, Figure C.4, Figure C.5, Figure C.6, Figure C.7, Figure C.8, Figure C.9, Figure C.10, and Figure C.11 row 2, column 2). I focused on the peaks found in models of selection with $s = 0.01$ and $f = 0.05$, because these values are most compatible with known cases of recent positive selection in humans. I also took into account the strength of correlations between the individual statistics under neutrality (Figure 4.8). After calculating these different versions of CSS, I plotted the binned mean of each CSS across the simulated sequence in a manner similar to the plots I generated for each individual statistic (*e. g.* Figure 4.9).

Table 4.3 Combinations of statistics included in CSS calculations

CSS name	Statistics included	Type of Signal
CSS_1	$\Delta iEHH_D$, $iEHH_S$, ΔDAF, F_{ST}, $xp\text{-}iEHH_S$	Narrow peak, low CI
CSS_2	$\Delta iEHH_D$, iHS, ΔDAF, F_{ST}, H_{FW}, D_{Tajima}	Narrow peak, low CI
CSS_3	H_{FW}, F_{ST}, $iEHH_S$	Narrow peak, low CI
CSS_4	iHS, ΔDAF, H_{FW}, D_{Tajima}	High peak, high CI
CSS_5	ΔDAF, H_{FW}, iHS	High peak, high CI
CSS_6	ΔDAF, H_{FW}, $iEHH_S$	High peak, high CI

I assessed the peaks in binned mean CSS for each combination of statistics, also focusing on simulated datasets including selection with $s = 0.01$ and $f = 0.05$. From these results, I identified three combinations of statistics that result in strong signals of selection (high peaks at the selected site) and three combinations that result in more specific signals of selection (narrow peaks at the selected site and confidence intervals very close to the binned mean values). I chose to use these six variations of CSS (shown in Table 4.3) for my analyses of the usefulness of CSS in detecting recent positive selection.

To compare the behavior of these six variations of CSS, I plotted the binned mean CSS value against the starting frequency of the beneficial allele in my simulated datasets for varying values of the selection coefficient (Figure 4.11). Because these variations of CSS may have different numbers of component statistics, their unadjusted mean CSS values are not comparable. To account for the different number of component statistics in each CSS, I adjusted each binned mean CSS from Figure 4.11 using the maximum possible CSS if all component statistics are in the 95th percentile, according to the following equations:

$$CSS_{max} = 0.95^{c} \quad \textbf{Eq. 4.6}$$

$$CSS_{adj} = CSS/CSS_{max} \quad \textbf{Eq. 4.7}$$

4.3.2 Results

The binned mean Cumulative Selection Score (CSS) for each of six combinations of statistics is presented in Figure 4.9 & Figure 4.10 and in Supplemental Figure C.12, Figure C.13, Figure C.14, and Figure C.15. In general, CSS compared favorably to the previously developed CMS. CSS values displayed stronger signatures of selection than did the individual statistics used to calculate each CSS. Binned mean CSS values exhibited narrower peaks around the selected site and had much narrower confidence intervals. In fact, CSS appeared to have narrower confidence intervals and a higher peak at the selected site than did CMS (Grossman *et al.* 2010), though direct comparison is difficult due to methodological differences. CSS appears to be more useful for a genome-wide scan for selection than any individual statistic, and may have greater resolution for detecting individual causal variants.

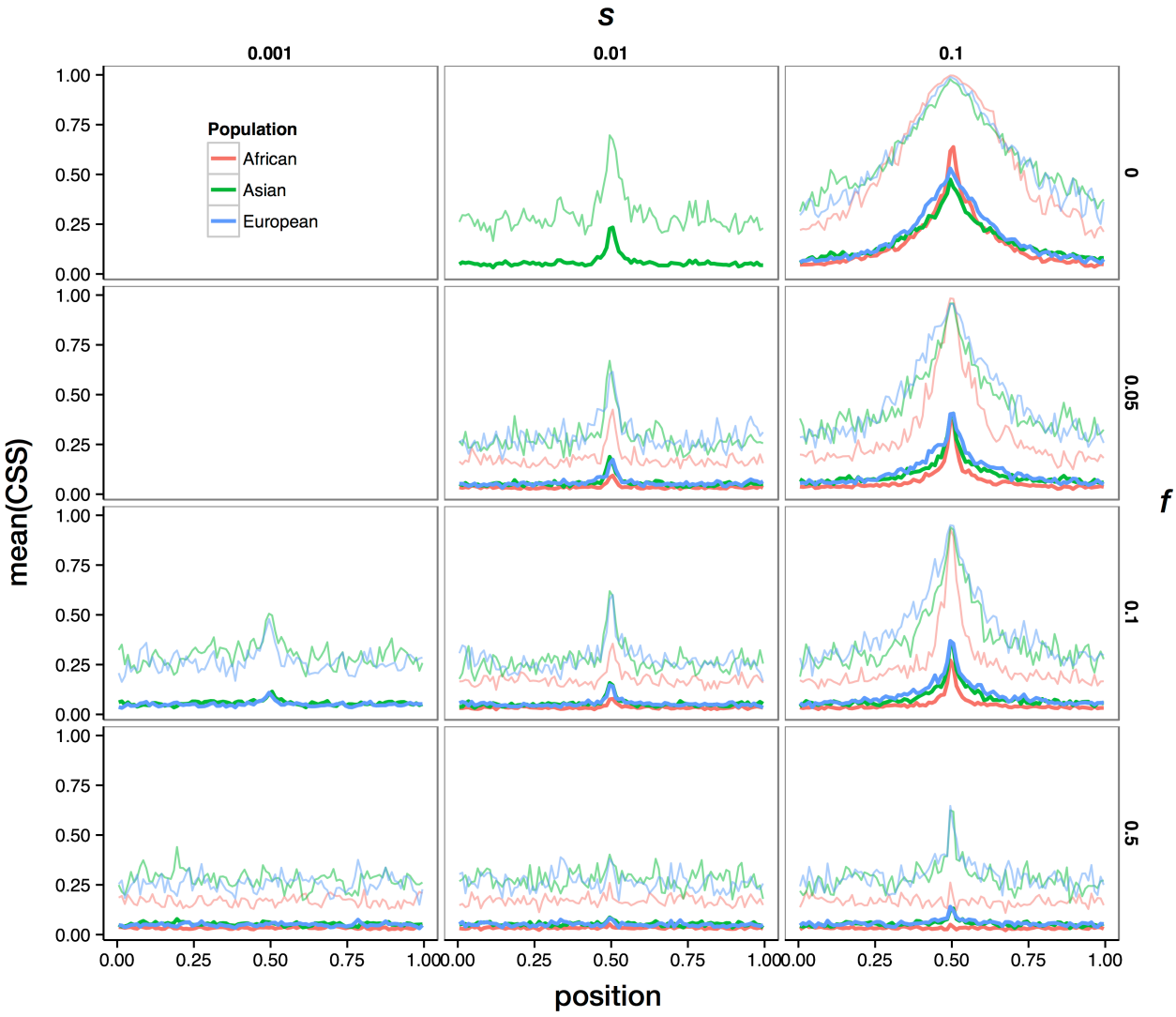


Figure 4.9 Binned mean CSS_4 displays a marked peak at the selected site in simulations including selection

For each combination of *s* and *f*, the average CSS_4 binned by position is plotted in thick, darker lines. Thin, lighter lines indicate the binned 95% confidence interval. CSS_4 includes p values from iHS, ΔDAF, H_{FW}, and D_{Tajima}. Simulations including selection in each of three populations are shown in each panel where available. Populations are colored according to the legend at top left.

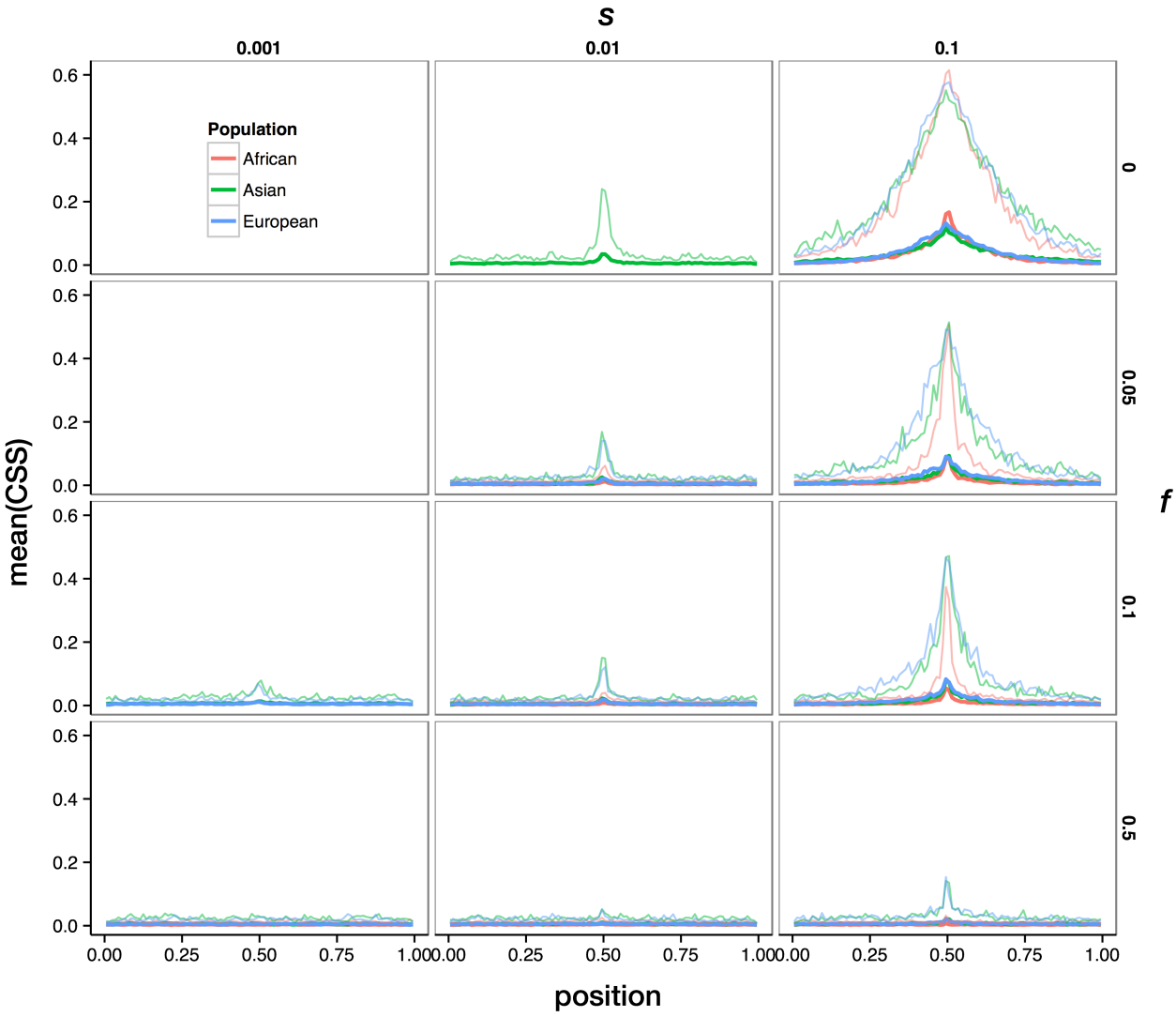


Figure 4.10 Binned mean CSS_2 displays an isolated peak at the selected site in simulations including selection

For each combination of *s* and *f*, the average CSS_4 binned by position is plotted in thick, darker lines. Thin, lighter lines indicate the binned 95% confidence interval. CSS_2 includes p values from $\Delta iEHH_D$, iHS, ΔDAF, F_{ST}, H_{FW}, D_{Tajima}. Simulations including selection in each of three populations are shown in each panel where available. Populations are colored according to the legend at top left.

I identified six combinations of statistics that produced either high peaks with little noise, or narrow peaks with narrow confidence intervals (Table 4.3). These six variations of CSS demonstrated that the statistics used to calculate CSS could be tuned for different purposes. The CSS variations with high peaks (*e. g.* CSS_4, Figure 4.9) could be good for detecting candidate regions. The CSS variations with narrow peaks (*e. g.* CSS_2, Figure 4.10) and narrow confidence intervals could be used to detect causal variants within those regions.

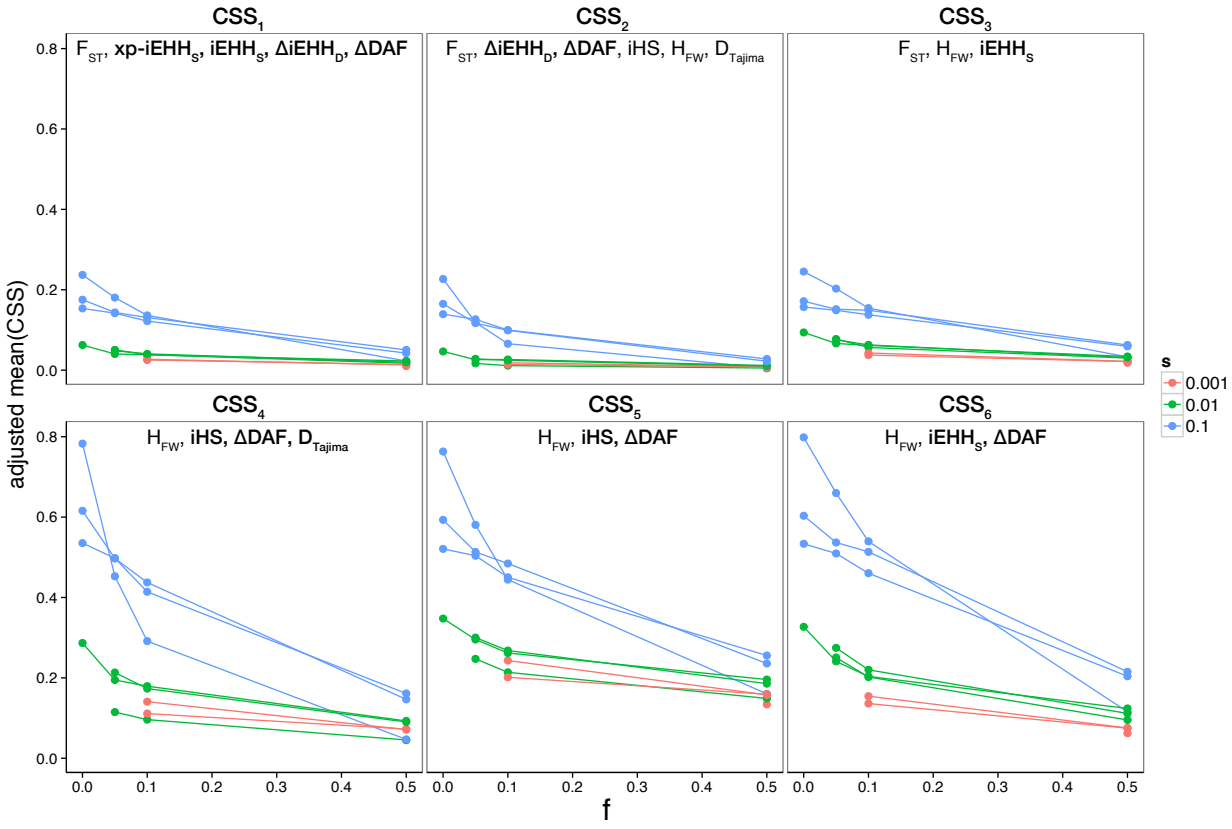


Figure 4.11 Mean CSS values at the selected site vary with *s* and *f* in simulations including selection
Each panel displays points marking the mean CSS value for the bin including the selected site. The mean CSS was adjusted to account for different numbers of statistics included in CSS, as described above. Points and lines are colored by the selection coefficient (*s*) for each simulation, according to the legend at right.

All of the variations of CSS that I tested displayed noticeable peaks in most selection models, including soft sweeps from standing variation. None of the variations of CSS displayed peaks at the selected site in simulated soft sweep datasets with s = 0.001 and f = 0.50. CSS_2, CSS_4, and CSS_5 additionally did not display peaks at the selected site in simulated soft sweep datasets with s = 0.001 and f = 0.50, while CSS_1, CSS_3, and CSS_6 exhibited peaks at the selected site under all of the other simulated models of selection. CSS was about as good at detecting the signatures of soft sweeps from standing variation as its component statistics were.

Figure 4.11 shows how the peak value of each CSS was affected by the parameters s and f in the simulated datasets including selection. The adjusted peak CSS value decreased with s and with f. These results demonstrate that higher values of f make the signatures of a soft sweep from standing variation very difficult to detect, even when the selection coefficient is very high.

4.4 Application to population genomic data from chromosome 2

4.4.1 Materials and Methods

4.4.1.1 Analysis of 1,000 Genomes variant data

I obtained low-coverage single nucleotide variant (SNV) data from chromosome 2 for 50 randomly-selected individuals each from the ASN, CEU, and YRI populations from the 1,000 Genomes phase 1 dataset (1000 Genomes Project Consortium *et al.* 2012), courtesy of Peter Sudmant. I obtained a genetic map for chromosome 2 (http://hapmap.ncbi.nlm.nih.gov/downloads/recombination/2011-01_phaseII_B37/), which was generated from the HapMap project phase 2 data (International HapMap Consortium *et al.* 2007). Using this genetic map, I converted physical coordinates for

every variant site to genetic coordinates using linear interpolation. These genetic distances are important for calculating statistics based on EHH and for defining window sizes for SFS-based statistics. Finally, I filtered out SNVs that were invariant in the 150 individuals in my dataset. I separated the chromosome into smaller segments, each containing approximately 15,000 variant sites. I analyzed each of these smaller segments separately to reduce the computational time required for the analysis.

Using the methodology described in sections 4.2.1.2, 4.2.1.4, and 4.2.1.5, I calculated neutrality test statistics at every common variant site in the dataset, using a cutoff of 5% to define common variants. For LD-based statistics, I used a homozygote counting algorithm to calculate each statistic from unphased genotype data for each individual, avoiding the error associated with computationally phased haplotypes (see section 4.4.1.2). Rather than using the binned mean and s.d. from neutral simulations to standardize each LD-based statistic, I used the mean and s.d. obtained from the chromosome 2 data itself. To calculate empirical p values for each neutrality test statistic, I similarly used the empirical cumulative distribution function of that statistic across chromosome 2. This prevents inaccuracies in the neutral simulation model from affecting the ability to detect the signatures of selection.

Finally, I followed the procedures described in sections 4.3.1.1 and 4.3.1.3 to calculate CSS across the entirety of chromosome 2. I used CSS_4 to identify candidate signatures of selection across the chromosome, assuming that binned mean CSS values in the 98th percentile were the targets of recent positive selection. Then I used CSS_2 to investigate the *LCT* region in detail.

4.4.1.2 *Comparing phased and unphased versions of LD-based neutrality test statistics*

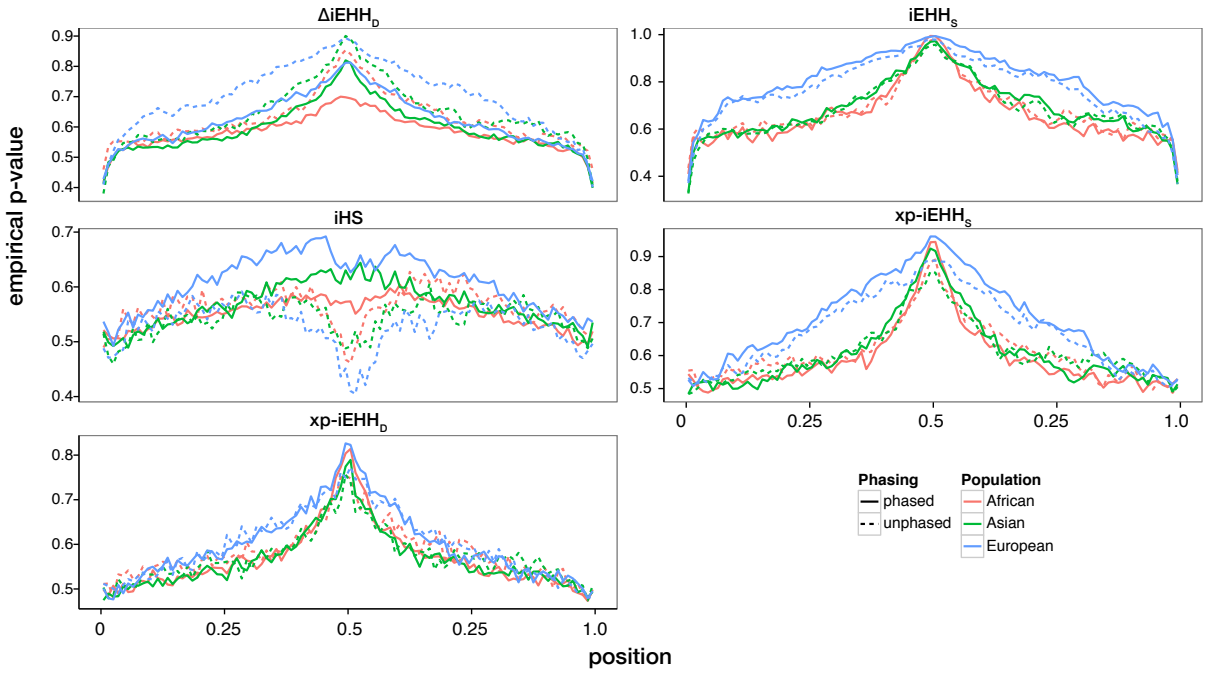


Figure 4.12 LD-based statistics can be estimated from either phased haplotype or unphased genotype data
Each panel displays the binned mean empirical p value of each LD or haplotype structure statistic in simulations including selection with $s = 0.1$ and $f = 0.1$. Estimates from unphased genotype data are shown as dashed lines and those from phased haplotype data are shown as solid lines. Lines are colored according to the population under selection per the legend in the bottom right. Correlation coefficients for each statistic in each population are displayed in Table 4.4.

In the simulated datasets I have discussed so far, the sequence data I was working with was perfectly phased. This made the calculation of LD-based statistics possible without any need to account for unknown phasing of the variants within each individual. In a real human population sample, however, perfectly phased data is not available. Unphased genotype data can be computationally phased to produce predicted haplotypes given the data; however, there is an associated amount of error introduced by computational phasing. Tang *et al.* developed the homozygote counting

algorithm for calculating the site-specific (rather than allele-specific) iEHH in unphased genotype data (Tang *et al.* 2007). I generalized this method to calculating iHS, xp-iEHH, and ΔiEHHD as well.

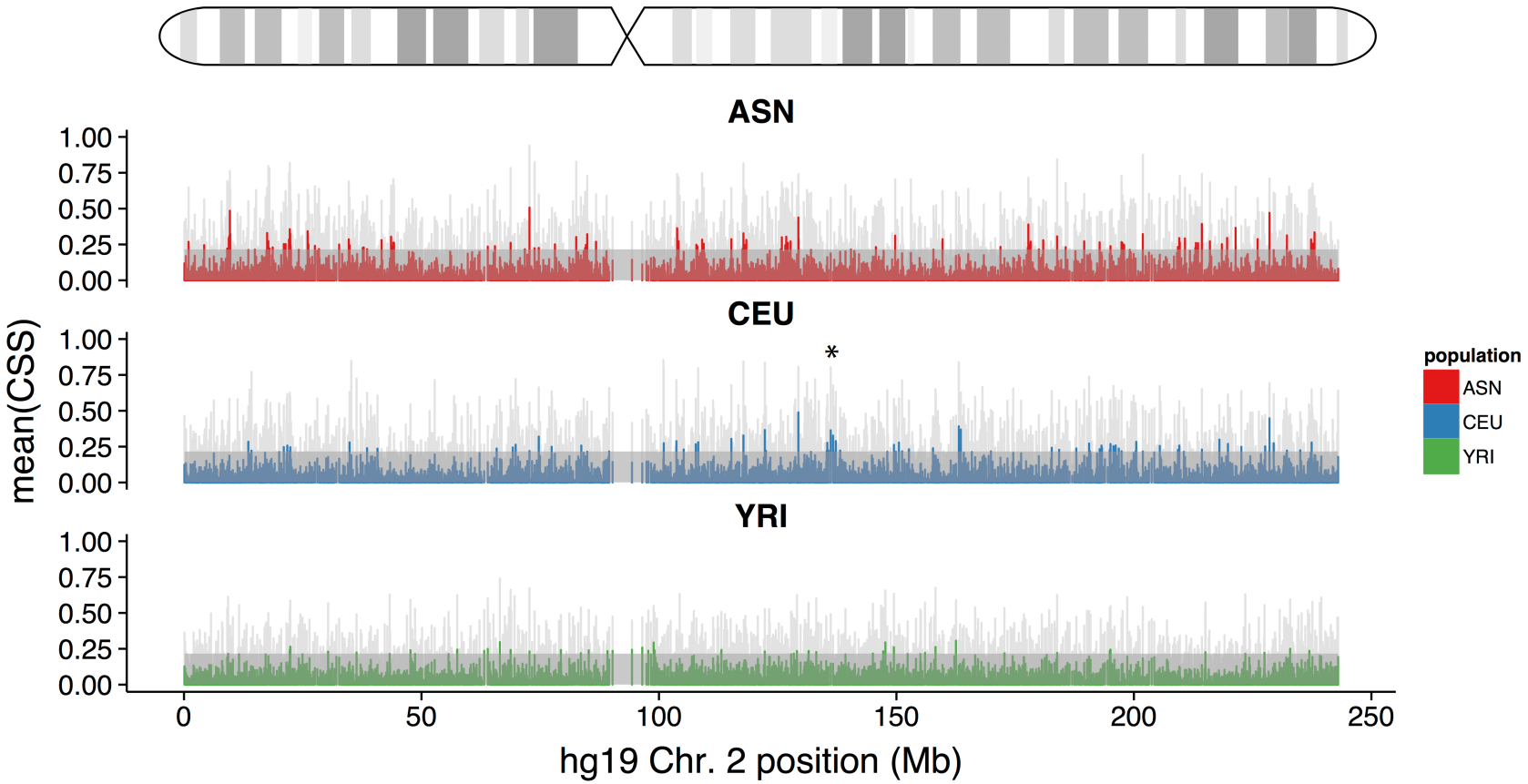
For each of the LD-based statistics, the phased and unphased versions of the statistic in one representative simulated dataset including selection are presented in Figure 4.12. The phased and unphased versions of each LD-based neutrality test statistic were significantly correlated in simulated datasets including selection (Table 4.4). The height of the peak in binned mean empirical p values for each statistic is of a similar magnitude between phased and unphased versions as well. By using the unphased versions of these statistics to analyze whole genome variant data from a human population sample, I removed one possible source of error from my analysis.

Table 4.4 Correlation coefficients between phased and unphased LD statistics[a]

Statistic	African R^2	European R^2	Asian R^2
ΔiEHHD,	0.40	0.53	0.55
iEHHS	0.95	0.69	0.97
iHS	0.49	0.65	0.60
xp-iEHHD	0.61	0.86	0.84
xp-iEHHS	0.80	0.72	0.64

a: all correlation coefficients were significant at $p < 0.0001$

Figure 4.13 Signals of selection across human chromosome 2, measured by CSS_4 in the 1,000 Genomes ASN, CEU, and YRI populations.
Each horizontal panel plots the mean value of CSS_4 in 0.075 cM bins across the length of chromosome 2. Mean CSS_4 values are colored according to population per the legend at right. Light gray bars indicate the 95% confidence interval within each bin. The darker gray box shading the lower portion of each panel marks the 98th percentile of CSS_4 values in all populations combined. Values above this gray box are candidate signatures of selection. The asterisk in the CEU panel is the approximate location of *LCT*.


ASN
CEU
YRI
1.00
0.75
0.50
0.25
0.00
mean(CSS)
*
0
50
100
150
200
250
hg19 Chr. 2 position (Mb)
population
ASN
CEU
YRI

4.4.2 Results

The results of calculating CSS_4 across the entirety of chromosome 2 are displayed in Figure 4.13. Because chromosome 2 is one of the longest chromosomes in the human genome, this served as a proof-of-concept for applying CSS in a genome-wide scan for the signatures of positive selection. I reduced the computational time of the analysis by splitting the chromosome into smaller segments, but the analysis of entire chromosome arms at once is computationally possible on a platform with enough available RAM, especially for the smaller chromosomes. The application of this method to the entire human genome is clearly warranted.

My method used a 98th percentile cutoff value for determining which binned mean CSS_4 values were high enough to indicate the signatures of recent positive selection. While this is an arbitrary cutoff value, it identified several interesting candidate selection regions. Most notably, the region surrounding *LCT* had an unusually high CSS_4 value (asterisk in Figure 4.13) in the European-descended CEU population. I examined the region in more detail to determine if the signature of selection could be further refined. I chose to use CSS_2 for the detailed analysis of the LCT region because it had relatively distinct peaks around the selected region in simulated data (Figure 4.10).

The entire region had an excess of high CSS_2 values in the CEU population, indicating a strong signature of positive selection. A previously-identified enhancer variant was observed in the region of very high CSS_2 values, but the highest Cumulative Selection Score in the region was shared by the two variants rs11898588 (A/G) and rs11903319 (G/T), which are only 11 bp apart. Interestingly, rs11903319 is

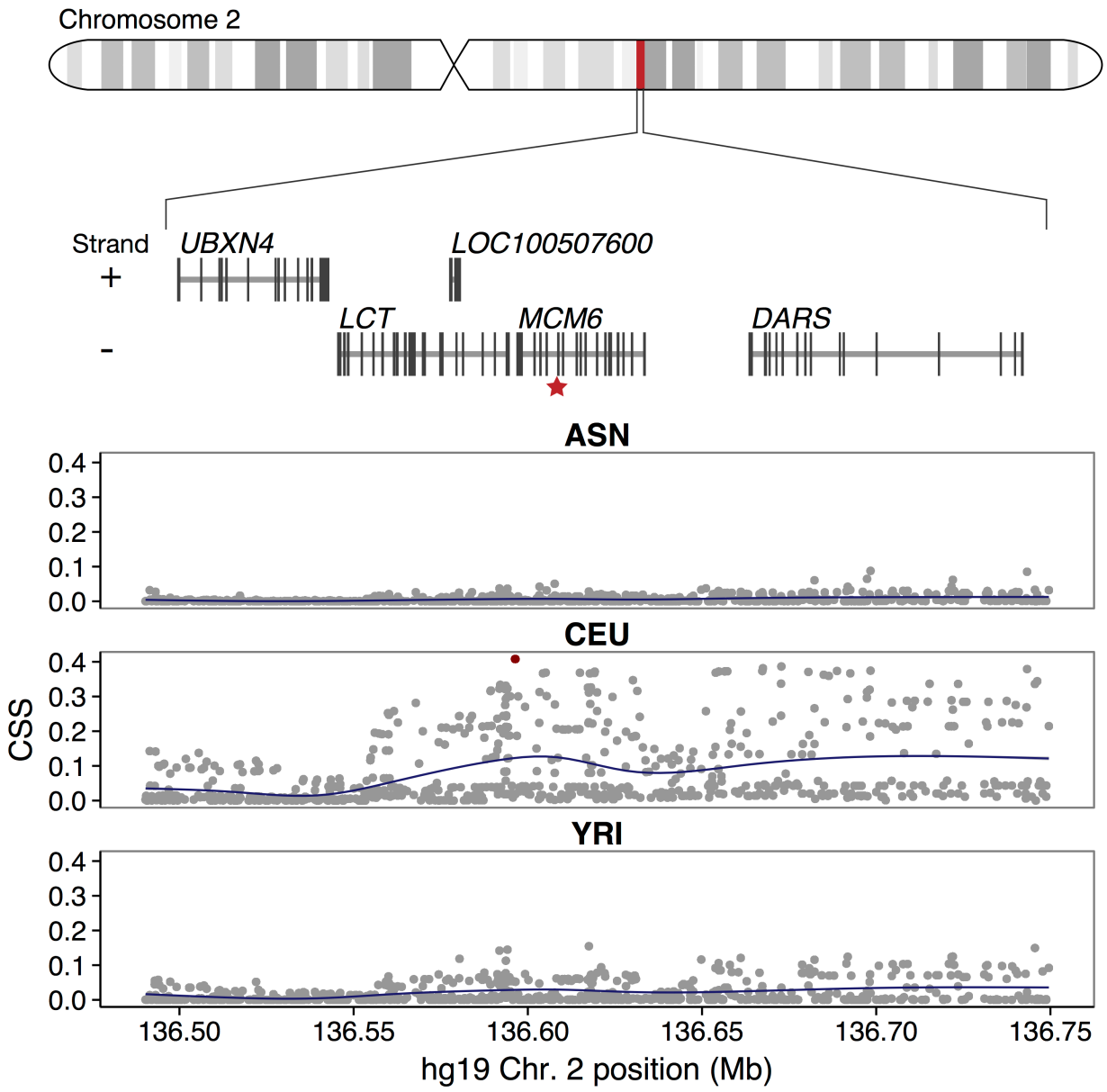


Figure 4.14 Signatures of selection in the *LCT* region, measured by CSS_2.
Each horizontal panel plots the CSS_2 value at every variant site in the *LCT* region for each population. Blue lines on each plot are smoothed lines from a generalize additive model. Two points plotted in red (and only 11 bp apart) share the highest CSS value in the entire region. The chromosome ideogram and gene locations from Figure 4.1 are shown at top.

fixed for the derived G allele in the CEU population, while this allele is around 70% in YRI and ASN populations. These two variants are in the intergenic region between *MCM6* and *LCT*, ~12 kb downstream of the known enhancer variant. While there are many variants with high CSS_2 scores within this region, that the highest CSS value was within 12 kb of a site with known functional consequences for selection is encouraging. In contrast, previous genome-wide scans for selection using SNP data and one or two neutrality test statistics routinely identified candidate regions of several Mb in length, often containing dozens of genes. The identification of a signature of positive selection near *LCT* is further evidence that CSS will improve the resolution and accuracy of genome-wide scans for selection.

4.5 Discussion

The methods developed in this chapter can still be improved in various ways. First and foremost, the impact of these results would be increased by the addition of a wider range of parameters in the simulation models (Table 4.1). It is particularly important to examine the performance of individual neutrality test statistics and of CSS in models of a soft sweep from standing variation with starting frequencies between 0.1 and 0.5, to determine at which value of *f* the signal of selection becomes difficult to detect.

Including only those simulated datasets for which the beneficial allele reached a frequency of 80% also limits the applicability of these results. A more comprehensive approach would be to obtain a specified number of simulated datasets across a range of beneficial allele frequencies. This would enable exploration of the effect of the beneficial allele frequency on the detection ability of each neutrality test statistic. In

effect, this would add another parameter to the simulation models specified in Table 4.1.

The genome-wide approach for the Cumulative Selection Score (CSS) can be further improved by determining a strict criterion for calling a cluster of high CSS values as a candidate selection region. Peak-finding algorithms and clustering approaches could be useful for this goal. The CSS approach particularly needs a better way to adjust scores for comparison between CSS versions with different numbers of contributing statistics. Methods based on Fisher's combined probability or the weighted Z method may be helpful for this purpose (Whitlock 2005).

While the CSS is qualitatively performing as well as the previously described CMS approach, a direct comparison between the two on the same simulated datasets would confirm this in a formal way. Similarly, the qualitative difference in the peaks of binned mean CSS and binned mean empirical p values for individual neutrality test statistics should be confirmed quantitatively. The resolution of CSS appears favorable compared to individual statistics and this should be confirmed more systematically.

Despite these shortcomings, I can make several interesting conclusions from the work described in this chapter. One somewhat surprising observation is that existing neutrality test statistics that were designed to detect the signatures of a classic hard sweep also perform well at detecting the signatures of a soft sweep from standing variation. In general, the best neutrality test statistics for detecting the signatures of a hard sweep are also the best for detecting a soft sweep. For small values of *f*, several statistics do just as well detecting soft sweeps as they do detecting the hard sweeps they were designed for. At the same time, the incalcitrance of soft sweep models with *f*

= 0.5 supports the conclusion that some selective sweeps will always be beyond our detection ability.

I also compared LD-based statistics calculated from unphased genotype data with those calculated from phased haplotype data in simulated datasets. The high correlation between the phased statistics and their unphased counterparts suggests that using the unphased versions will not lose much information. This can remove another source of error by avoiding the use of computationally phased data. To confirm this, I plan to compare the reduction in signal that results from using unphased LD statistics with the reduction in signal due to error in computational phasing.

I have also demonstrated that CSS is a good modification of the CMS for application to genome-wide data. I identified a class of contributing statistic combinations that had high peak CSS values but also high levels of noise. These CSSs are appropriate for a first-pass genome-wide scan, with a goal of a low false negative rate. Another CSS class exhibited lower peak values, but very distinct peaks with narrow confidence intervals. These CSSs are best used in a detailed second-pass analysis within each candidate selected region. This approach worked well in human variation data on chromosome 2, to identify the signature of selection at *LCT* and then identify variants within the region that are close to a putatively causal variant. Application of this CSS approach to the whole genome promises to provide further clarity on the question of which alleles have experienced recent positive selection in the human genome.

5. Concluding Remarks

In this dissertation I have described several approaches for making inferences regarding recent human evolutionary history from population genomic data. First, I demonstrated that mitochondrial lineage tests capture only part of the available information regarding an individual's genetic ancestry. Especially in recently admixed populations, it is common for a person's autosomal ancestry to be more complex than indicated by their mitochondrial ancestry.

Second, I resolved a controversy in the literature by showing that extant variation on the X chromosome and the autosomes is consistent with a female bias in effective population size in ancestral humans, as well as with a later male bias associated with the Out of Africa migration and bottleneck. Previously conflicting results can be explained by methodological differences – specifically the use of different effective sex ratio estimators that measure different time periods in human history.

Finally, I evaluated the detection ability of the most frequently used neutrality test statistics in models of selection with varying selection coefficients and selective sweep models. I found that neutrality tests designed to detect the signatures of a hard sweep are also relatively good for detecting the signatures of a soft sweep from standing variation. I also developed the Cumulative Selection Score (CSS) for combining the information from multiple neutrality tests and extended this into an approach for genome-wide scans for selection. Using carefully selected combinations of statistics to calculate CSS, I can first identify regions as candidate targets of positive

selection and then identify variants within these regions that may be the specific site subject to selection.

The results from these disparate but related projects have yielded enticing information about recent evolution in humans, but we have much still to learn. The record of human evolution left in our patterns of genetic variation has revealed much about the distribution of genetic ancestry, varying sex ratios over time, and specific genetic adaptations. How much evolutionary history we can still extract from population genomic data remains an open question. The advent of affordable population-scale whole genome sequencing and the sampling of more diverse human populations are sure to facilitate new and unexpected discoveries.

In particular, the past century has seen great strides in our understanding of the impact of selection on human genetic variation. The principles of population genetics have provided explanations for why a disease may be more prevalent in one population than another or, indeed, why a disease allele is maintained at all. Additionally, the exceptional cases of extreme population differentiation caused by directional selection have proved the rule of unusually close genetic bonds between the many populations of our far-flung species.

Recent studies have yielded an immense body of knowledge concerning the genetic signatures of positive selection on a genome-wide scale. Positive selection has been a strong force acting on a small but significant portion of the human genome and has left its marks across all chromosomes and all human populations, affecting many varied gene processes and pathways. There are no strikingly singular adaptations

evident in the genetic record of human evolution, showing that human populations have adapted to their respective environments in ways numerous and diverse. The underlying message from these findings is that the effects of selection cannot be discounted for any genomic location, classes of functional elements, or populations (Hahn 2008).

A comprehensive understanding of the effects of selection on the human genome has never seemed more attainable than it does today; however, many important questions remain. First, what are the practical limits of detection for the signatures of selection? Second, how much of the genome has been subject to positive selection? Third, what is the distribution of selection coefficients among the targets of positive selection in the human genome?

The limits of detection can be determined with coalescent simulations under widely varying models of selection, but our detection ability may improve with advances in detection methods. In particular, the development of methods for detecting polygenic selection may vastly improve our ability to detect the subtle shifts in allele frequency that may underlie many human adaptations. Also important will be models of selection on copy number variants and microsatellites. As detection methods improve, our estimate of the proportion of the genome that has been a target of selection will also necessarily improve. Selection coefficients estimated from identified targets of positive selection can give some indication of the distribution of effect sizes of an advantageous allele, but complementary approaches such as deep mutational scanning of protein variants and experimental evolution of model organisms

are already providing valuable information on the average effect of a beneficial mutation.

Once we are armed with a detailed map of the candidate targets of selection across the human genome, it will be important to validate these candidates and determine their functional importance. Using comprehensive functional datasets such as ENCODE, we can prioritize candidate selected regions for detailed molecular investigation and validation. If we can make this important connection between a genomic signature of positive selection and its phenotypic consequences, we will be able to make more accurate predictions of genetic risk for individuals, and gain insights into the connection between human populations and their evolutionary environments.

6. References

1000 Genomes Project Consortium, Abecasis GR, Auton A, *et al.* 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65.

Achilli A, Perego UA, Bravi CM, Coble MD, Kong Q-P, Woodward SR, Salas A, Torroni A, Bandelt H-J. 2008. The phylogeny of the four pan-American MtDNA haplogroups: implications for evolutionary and disease studies. PLoS ONE 3:e1764.

Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, Kruglyak L. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. Plos Biol 2:e286.

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. 2002. Interrogating a high-density SNP map for signatures of natural selection. Genome Research 12:1805–1814.

Akey JM. 2009. Constructing genomic maps of positive selection in humans: where do we go from here? Genome Research 19:711–722.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Research 19:1655–1664.

Bamshad M, Wooding S, Salisbury BA, Stephens JC. 2004. Deconstructing the relationship between genetics and race. Nat Rev Genet 5:598–609.

Bamshad MJ, Mummidi S, Gonzalez E, *et al.* 2002. A strong signature of balancing selection in the 5' cis-regulatory region of CCR5. Proceedings of the National Academy of Sciences 99:10539–10544.

Barbujani G, Colonna V. 2010. Human genome diversity: frequently asked questions. Trends in Genetics 26:285–295.

Barbujani G, Magagni A, Minch E, Cavalli-Sforza LL. 1997. An apportionment of human DNA diversity. Proceedings of the National Academy of Sciences 94:4516–4519.

Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. 2008. Natural selection has driven population differentiation in modern humans. Nat Genet 40:340–345.

Barreiro LB, Quintana-Murci L. 2010. From evolutionary genetics to human immunology: how selection shapes host defence genes. Nat Rev Genet 11:17–30.

Barton NH. 1998. The effect of hitch-hiking on neutral genealogies. Genetics Research 72:123–133.

Begun D, Whitley P. 2000. Reduced X-linked nucleotide polymorphism in Drosophila simulans. Proceedings of the National Academy of Sciences 97:5960–5965.

Behar DM, Rosset S, Blue-Smith J, *et al.* 2007. The Genographic Project public participation mitochondrial DNA database. PLoS Genet 3:e104.

Benn M, Schwartz M, Nordestgaard BG, Tybjærg-Hansen A. 2008. Mitochondrial haplogroups: ischemic cardiovascular disease, other diseases, mortality, and longevity in the general population. Circulation 117:2492–2501.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. 2004. Genetic signatures of strong recent positive selection at the lactase gene. American Journal of Human Genetics 74:1111–1120.

Biswas S, Akey JM. 2006. Genomic insights into positive selection. Trends in Genetics 22:437–446.

Boattini A, Martínez-Cruz B, Sarno S, *et al.* 2013. Uniparental markers in Italy reveal a sex-biased genetic structure and different historical strata. PLoS ONE 8:e65441.

Bollongino R, Nehlich O, Richards MP, Orschiedt J, Thomas MG, Sell C, Fajkosová Z, Powell A, Burger J. 2013. 2000 years of parallel societies in Stone Age Central Europe. Science 342:479–481.

Bolnick DA, Fullwiley D, Duster T, *et al.* 2007. Genetics. The science and business of genetic ancestry testing. Science 318:399–400.

Brandt G, Haak W, Adler CJ, *et al.* 2013. Ancient DNA Reveals Key Stages in the Formation of Central European Mitochondrial Genetic Diversity. Science 342:257–261.

Bustamante C, Ramachandran S. 2009. Evaluating signatures of sex-specific processes in the human genome. Nat Genet 41:8–10.

Cann HM, de Toma C, Cazes L, *et al.* 2002. A human genome diversity cell line panel. Science 296:261–262.

Cardena MMSG, Ribeiro-Dos-Santos A, Santos S, Mansur AJ, Pereira AC, Fridman C. 2013. Assessment of the relationship between self-declared ethnicity, mitochondrial haplogroups and genomic ancestry in Brazilian individuals. PLoS ONE 8:e62005.

Casto AM, Li JZ, Absher DM, Myers R, Ramachandran S, Feldman MW. 2010. Characterization of X-linked SNP genotypic variation in globally distributed human populations. Genome Biol 11:R10.

Cavalli-Sforza LL, Feldman M. 2003. The application of molecular genetic approaches

to the study of human evolution. Nat Genet 33:266–275.

Cavalli-Sforza LL. 1966. Population structure and human evolution. Proceedings of the Royal Society B: Biological Sciences 164:362–379.

Cavalli-Sforza LL. 2007. Human evolution and its relevance for genetic epidemiology. Annu. Rev. Genom. Human. Genet. 8:1–15.

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

Chen H, Patterson N, Reich D. 2010. Population differentiation as a test for selective sweeps. Genome Research 20:393–402.

Chen YS, Torroni A, Excoffier L, Santachiara-Benerecetti AS, Wallace DC. 1995. Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. American Journal of Human Genetics 57:133–149.

Danecek P, Auton A, Abecasis GR, *et al.* 2011. The variant call format and `VCFtools`. Bioinformatics (Oxford, England) 27:2156–2158.

Deshpande O, Batzoglou S, Feldman M. 2009. A serial founder effect model for human settlement out of Africa. Proceedings of the Royal Society B: Biological Sciences 276:291–300.

Douadi MI, Gatti S, Levrero F, Duhamel G, Bermejo M, Vallet D, Menard N, Petit EJ. 2007. Sex-biased dispersal in western lowland gorillas (Gorilla gorilla gorilla). Mol Ecol 16:2247–2259.

Duggan AT, Stoneking M. 2013. A highly unstable recent mutation in human mtDNA. American Journal of Human Genetics 92:279–284.

Enattah NS, Jensen TGK, Nielsen M, *et al.* 2008. Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. American Journal of Human Genetics 82:57–72.

Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Järvelä I. 2002. Identification of a variant associated with adult-type hypolactasia. Nat Genet 30:233–237.

Eriksson J, Siedel H, Lukas D, Kayser M, Erler A, Hashimoto C, Hohmann G, Boesch C, Vigilant L. 2006. Y-chromosome analysis confirms highly sex-biased dispersal and suggests a low male effective population size in bonobos (Pan paniscus). Mol Ecol 15:939–949.

Ewing G, Hermisson J. 2010. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus.

Bioinformatics (Oxford, England) 26:2064–2065.

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. Genetics 155:1405–1413.

Felsenstein, J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. *Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.*

Frudakis T. 2008. The legitimacy of genetic ancestry tests. Science 319:1039–40–authorreply1039–40.

Fu Q, Mittnik A, Johnson PLF, *et al.* 2013. A revised timescale for human evolution based on ancient mitochondrial genomes. Current Biology 23:553–559.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. Genetics 133:693–709.

Gallego Romero I, Basu Mallick C, Liebert A, *et al.* 2012. Herders of Indian and European cattle share their predominant allele for lactase persistence. Molecular Biology and Evolution 29:249–260.

Galvani A, Slatkin M. 2003. Evaluating plague and smallpox as historical selective pressures for the CCR5-Δ32 HIV-resistance allele. Proceedings of the National Academy of Sciences 100:15276–15279.

Garud NR, Messer PW, Buzbas EO, Petrov DA. 2013. Soft selective sweeps are the primary mode of recent adaptation in Drosophila melanogaster.

Gluckman PD, Low FM, Buklijas T, Hanson MA, Beedle AS. 2011. How evolutionary principles improve the understanding of human health and disease. Evolutionary Applications 4:249–263.

Goldstein DB, Chikhi L. 2002. Human migrations and population structure: what we know and why it matters. Annu. Rev. Genom. Human. Genet. 3:129–152.

Goudet J, Perrin N, Waser P. 2002. Tests for sex-biased dispersal using bi-parentally inherited genetic markers. Mol Ecol 11:1103–1114.

Greenwood P. 1980. Mating systems, philopatry and dispersal in birds and mammals. Animal Behaviour 28:1140–1162.

Grossman SR, Andersen KG, Shlyakhter I, *et al.* 2013. Identifying recent adaptations in large-scale genomic data. Cell 152:703–713.

Grossman SR, Shlyakhter I, Shylakhter I, *et al.* 2010. A composite of multiple signals distinguishes causal variants in regions of positive selection. Science 327:883–886.

Hahn M. 2008. Toward a selection theory of molecular evolution. Evolution 62:255–

265.

Hamilton G, Stoneking M. 2005. Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilocal populations. Proceedings of the National Academy of Sciences 102:7476–7480.

Hammer M, Garrigan D, Wood E, Wilder J, Mobasher Z, Bigham AW, Krenz JG, Nachman MW. 2004. Heterogeneous patterns of variation among multiple human X-linked loci: the possible role of diversity-reducing selection in non-Africans. Genetics 167:1841–1853.

Hammer MF, Mendez FL, Cox MP, Woerner AE, Wall JD. 2008. Sex-biased evolutionary forces shape genomic patterns of human diversity. PLoS Genet 4:e1000202.

Hammer MF, Woerner AE, Mendez FL, Watkins JC, Cox MP, Wall JD. 2010. The ratio of human X chromosome to autosome diversity is positively correlated with genetic distance from genes. Nat Genet 42:830–831.

Handley LJL, Manica A, Goudet J, Balloux F. 2007. Going the distance: human population genetics in a clinal world. Trends in Genetics 23:432–439.

Harris EE, Meyer D. 2006. The molecular signature of selection underlying human adaptations. Am. J. Phys. Anthropol. Suppl 43:89–130.

Hawks J, Wang ET, Cochran GM, Harpending HC, Moyzis RK. 2007. Recent acceleration of human adaptive evolution. Proceedings of the National Academy of Sciences 104:20753–20758.

Hedrick P. 2007. Sex: differences in mutation, recombination, selection, gene flow, and genetic drift. Evolution 61:2750–2771.

Hedrick PW, Verrelli BC. 2006. “Ground truth” for selection on CCR5-Delta32. Trends in Genetics 22:293–296.

Hermisson J, Pennings P. 2005. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics 169:2335–2352.

Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, 1000 Genomes Project, Sella G, Przeworski M. 2011. Classic selective sweeps were rare in recent human evolution. Science 331:920–924.

Herrnstadt C, Elson JL, Fahy E, *et al.* 2002. Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups. American Journal of Human Genetics 70:1152–1171.

Hudson R. 2002. Generating samples under a Wright-Fisher neutral model of genetic

variation. Bioinformatics (Oxford, England) 18:337–338.

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

Ingram CJE, Mulcare CA, Itan Y, Thomas MG, Swallow DM. 2009. Lactose digestion and the evolutionary genetics of lactase persistence. Hum Genet 124:579–591.

International HapMap Consortium, Frazer KA, Ballinger DG, *et al.* 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861.

Jinam TA, Hong L-C, Phipps ME, Stoneking M, Ameen M, Edo J, HUGO Pan-Asian SNP Consortium, Saitou N. 2012. Evolutionary history of continental southeast Asians: "early train" hypothesis based on genetic analysis of mitochondrial and autosomal DNA data. Molecular Biology and Evolution 29:3513–3527.

Jones BL, Raga TO, Liebert A, *et al.* 2013. Diversity of lactase persistence alleles in ethiopia: signature of a soft selective sweep. American Journal of Human Genetics 93:538–544.

Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT, Batzer MA. 2000. The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. American Journal of Human Genetics 66:979–988.

Kaessmann H, Wiebe V, Weiss G, Pääbo S. 2001. Great ape DNA sequences reveal a reduced diversity and an expansion in humans. Nat Genet 27:155–156.

Kaplan NL, Hudson RR, Langley CH. 1989. The "hitchhiking effect" revisited. Genetics 123:887–899.

Keinan A, Mullikin JC, Patterson N, Reich D. 2009. Accelerated genetic drift on chromosome X during the human dispersal out of Africa. Nat Genet 41:66–70.

Keinan A, Reich D. 2010. Can a sex-biased human demography account for the reduced effective population size of chromosome X in non-Africans? Molecular Biology and Evolution 27:2312–2321.

Kivisild T, Rootsi S, Metspalu M, *et al.* 2003. The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. American Journal of Human Genetics 72:313–332.

Kong Q-P, Yao Y-G, Liu M, Shen S-P, Chen C, Zhu C-L, Palanichamy MG, Zhang Y-P. 2003. Mitochondrial DNA sequence polymorphisms of five ethnic populations from northern China. Hum Genet 113:391–405.

Labuda D, Lefebvre J, Roy-Gagnon M. 2010. Response to Lohmueller *et al.* American

Journal of Human Genetics 86:980.

Labuda D, Lefebvre J-F, Nadeau P, Roy-Gagnon M-H. 2010. Female-to-male breeding ratio in modern humans-an analysis based on historical recombinations. American Journal of Human Genetics 86:353–363.

Lambert CA, Connelly CF, Madeoy J, Qiu R, Olson MV, Akey JM. 2010. Highly punctuated patterns of population structure on the X chromosome and implications for African evolutionary history. American Journal of Human Genetics 86:34–44.

Langergraber K, Siedel H, Mitani J, Wrangham R, Reynolds V, Hunt K, Vigilant L. 2007. The genetic signature of sex-biased migration in patrilocal chimpanzees and humans. PLoS ONE 2:e973.

Lee DY, Hayes JJ, Pruss D, Wolffe AP. 1993. A positive role for histone acetylation in transcription factor access to nucleosomal DNA. Cell 72:73–84.

Lewontin RC. 1995. The apportionment of human diversity. :381–398.

Li H. 2011. `Tabix`: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics (Oxford, England) 27:718–719.

Li J, Li H, Jakobsson M, Li S, Sjödin P, Lascoux M. 2012. Joint analysis of demography and selection in population genetics: where do we stand and where could we go? Mol Ecol 21:28–44.

Li JZ, Absher DM, Tang H, *et al.* 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100–1104.

Li Y, Vinckenbosch N, Tian G, *et al.* 2010. Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. Nat Genet.

Libert F, Cochaux P, Beckman G, *et al.* 1998. The ΔCCR5 mutation conferring protection against HIV-1 in Caucasian populations has a single and recent origin in Northeastern Europe. Human Molecular Genetics 7:399–406.

Lohmueller KE, Albrechtsen A, Li Y, *et al.* 2011. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. PLoS Genet 7:e1002326.

Lohmueller KE, Degenhardt JD, Keinan A. 2010. Sex-averaged recombination and mutation rates on the X chromosome: a comment on Labuda *et al.* American Journal of Human Genetics 86:978–980; authorreply980–authorreply981.

Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genet Res 23:23–35.

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

Melnick D, Hoelzer G. 1992. Differences in male and female macaque dispersal lead to contrasting distributions of nuclear and mitochondrial DNA variation. Int J Primatol 13:379–393.

Moreno-Estrada A, Gravel S, Zakharia F, *et al.* 2013. Reconstructing the Population Genetic History of the Caribbean. PLoS Genet 9:e1003925.

Musters H, Huntley M, Singh R. 2006. A genomic comparison of faster-sex, faster-X, and faster-male evolution between Drosophila melanogaster and Drosophila pseudoobscura. Journal of molecular evolution 62:693–700.

O'Reilly PF, Birney E, Balding DJ. 2008. Confounding between recombination and selection, and the Ped/Pop method for detecting selection. Genome Research 18:1304–1313.

Olds LC, Sibley E. 2003. Lactase persistence DNA variant enhances lactase promoter activity in vitro: functional role as a cis regulatory element. Human Molecular Genetics 12:2333–2340.

Oota H, Settheetham-Ishida W, Tiwawech D, Ishida T, Stoneking M. 2001. Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. Nat Genet 29:20–21.

Poetsch M, Wiegand A, Harder M, Blöhm R, Rakotomavo N, Freitag-Wolf S, Wurmb-Schwark von N. 2013. Determination of population origin: a comparison of autosomal SNPs, Y-chromosomal and mtDNA haplogroups using a Malagasy population as example. Eur J Hum Genet.

Polanski A, Kimmel M, Chakraborty R. 1998. Application of a time-dependent coalescence process for inferring the history of population size changes from DNA sequence data. Proceedings of the National Academy of Sciences 95:5456–5461.

Pool J, Nielsen R. 2007. Population size changes reshape genomic patterns of diversity. Evolution 61:3001–3006.

Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Current Biology 20:R208–R215.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Prugnolle F, Meeus TD. 2002. Inferring sex-biased dispersal from population genetic tools: a review. Heredity 88:161–165.

Przeworski M, Coop G, Wall JD. 2005. The signature of positive selection on standing genetic variation. Evolution 59:2312–2323.

Purcell S, Neale B, Todd-Brown K, *et al.* 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81:559–575.

Quinlan R. 2008. Human pair-bonds: Evolutionary functions, ecological variation, and adaptive development. Evolutionary Anthropology 17:227–238.

Ramachandran S, Deshpande O, Roseman C. 2005. Support from the relationship of genetic and geographic distance in human populations for a …. Proceedings of the National Academy of Sciences 102:15942–15947.

Ramachandran S, Rosenberg N, Feldman M, Wakeley J. 2008. Population differentiation and migration: Coalescence times in a two-sex island model for autosomal and X-linked loci. Theoretical Population Biology 74:291–301.

Ramachandran S, Rosenberg N, Zhivotovsky L, Feldman M. 2004. Robustness of the inference of human population structure: a comparison of X-chromosomal and autosomal microsatellites. Human Genomics 1:87–97.

Reed FA, Akey JM, Aquadro CF. 2005. Fitting background-selection predictions to levels of nucleotide variation and divergence along the human autosomes. Genome Research 15:1211–1221.

Ronald J, Akey JM. 2005. Genome-wide scans for loci under selection in humans. Human Genomics 2:113–125.

Rosenberg NA. 2006. Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann. Hum. Genet. 70:841–847.

Royal CD, Novembre J, Fullerton SM, Goldstein DB, Long JC, Bamshad MJ, Clark AG. 2010. Inferring genetic ancestry: opportunities, challenges, and implications. American Journal of Human Genetics 86:661–673.

Ruiz-Pesini E, Lott MT, Procaccio V, *et al.* 2007. An enhanced MITOMAP with a global mtDNA mutational phylogeny. Nucleic Acids Research 35:D823–D828.

Ruiz-Pesini E, Mishmar D, Brandon M, Procaccio V, Wallace DC. 2004. Effects of purifying and adaptive selection on regional variation in human mtDNA. Science 303:223–226.

Sabeti PC, Reich DE, Higgins JM, *et al.* 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837.

Sabeti PC, Varilly P, Fry B, *et al.* 2007. Genome-wide detection and characterization of positive selection in human populations. Nature 449:913–918.

Sabeti PC, Walsh E, Schaffner SF, *et al.* 2005. The case for selection at CCR5-Delta32. Plos Biol 3:e378.

Salas A, Acosta A, Alvarez-Iglesias V, Cerezo M, Phillips C, Lareu MV, Carracedo A. 2008. The mtDNA ancestry of admixed Colombian populations. Am J Hum Biol 20:584–591.

Sarata AK. 2008. Genetic ancestry testing: CRS report for Congress. Available from: http://webcache.googleusercontent.com/search?q=cache:48huLKl2GNsJ:research.policyarchive.org/18866.pdf+&cd=1&hl=en&ct=clnk&gl=us

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. Genome Research 15:1576–1583.

Scheinfeldt LB, Biswas S, Madeoy J, Connelly CF, Schadt EE, Akey JM. 2009. Population genomic analysis of ALMS1 in humans reveals a surprisingly complex evolutionary history. Molecular Biology and Evolution 26:1357–1367.

Seielstad M, Minch E, Cavalli-Sforza LL. 1998. Genetic evidence for a higher female migration rate in humans. Nat Genet 20:278–280.

Ségurel L, Martínez-Cruz B, Quintana-Murci L, *et al.* 2008. Sex-specific genetic structure and social organization in Central Asia: Insights from a multi-locus study. PLoS Genet 4:e1000200.

Shriver MD, Kittles RA. 2004. Genetic ancestry and the search for personalized genetic histories. Nat Rev Genet 5:611–618.

Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, Deka R, Ferrell RE. 1997. Ethnic-affiliation estimation by use of population-specific DNA markers. American Journal of Human Genetics 60:957–964.

Simonson TS, Xing J, Barrett R, *et al.* 2011. Ancestry of the Iban is predominantly Southeast Asian: genetic evidence from autosomal, mitochondrial, and Y chromosomes. PLoS ONE 6:e16338.

Singh N, Macpherson J, Jensen J, Petrov DA. 2007. Similar levels of X-linked and autosomal nucleotide variation in African and non-African populations of Drosophila melanogaster. BMC Evolutionary … 7:202.

Stajich JE, Hahn MW. 2005. Disentangling the effects of demography and selection in human history. Molecular Biology and Evolution 22:63–73.

Stephan W. 2010. Genetic hitchhiking versus background selection: the controversy and its implications. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 365:1245–1253.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.

Tang H, Peng J, Wang P, Risch NJ. 2005. Estimation of individual admixture: analytical and study design considerations. Genet Epidemiol 28:289–301.

Tang K, Thornton KR, Stoneking M. 2007. A new approach for using genome scans to detect recent positive selection in the human genome. Plos Biol 5:e171.

Tennessen JA, Bigham AW, O'Connor TD, *et al.* 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337:64–69.

Tennessen JA, Madeoy J, Akey JM. 2010. Signatures of positive selection apparent in a small sample of human exomes. Genome Research 20:1327–1334.

The American Society of Human Genetics. 2008. Ancestry testing statement. www.ashg.org [Internet]. Available from: http://www.ashg.org/pdf/ASHGAncestryTestingStatement_FINAL.pdf

Tishkoff SA, Reed FA, Friedlaender FR, *et al.* 2009. The genetic structure and history of Africans and African Americans. Science 324:1035–1044.

Tishkoff SA, Reed FA, Ranciaro A, *et al.* 2007. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39:31–40.

Tishkoff SA, Verrelli BC. 2003. Patterns of human genetic diversity: implications for human evolutionary history and disease. Annu. Rev. Genom. Human. Genet. 4:293–340.

Torroni A, Huoponen K, Francalacci P, Petrozzi M, Morelli L, Scozzari R, Obinu D, Savontaus ML, Wallace DC. 1996. Classification of European mtDNAs from an analysis of three European populations. Genetics 144:1835–1850.

Torroni A, Schurr TG, Cabell MF, Brown MD, Neel JV, Larsen M, Smith DG, Vullo CM, Wallace DC. 1993. Asian affinities and continental radiation of the four founding Native American mtDNAs. American Journal of Human Genetics 53:563–590.

Torroni A, Sukernik RI, Schurr TG, Starikorskaya YB, Cabell MF, Crawford MH, Comuzzie AG, Wallace DC. 1993. mtDNA variation of aboriginal Siberians reveals distinct genetic affinities with Native Americans. American Journal of Human Genetics 53:591–608.

Underhill P, Kivisild T. 2007. Use of Y chromosome and mitochondrial DNA population

structure in tracing human migrations. Annual Review of Genetics 41:539–564.

van Oven M, Kayser M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum. Mutat. 30:E386–E394.

Venables WN, Ripley BD. 2002. Modern Applied Statistics with S. 4 ed. New York: Springer Available from: http://www.stats.ox.ac.uk/pub/MASS4

Via M, Ziv E, Burchard EG. 2009. Recent advances of genetic ancestry testing in biomedical research and direct to consumer testing. Clin. Genet. 76:225–235.

Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC. 1991. African populations and the evolution of human mitochondrial DNA. Science 253:1503–1507.

Vitti JJ, Cho MK, Tishkoff SA, Sabeti PC. 2012. Human evolutionary genomics: ethical and interpretive issues. Trends in Genetics 28:137–145.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. Plos Biol 4:e72.

Wagner JK, Cooper JD, Sterling R, Royal CD. 2012. Tilting at windmills no longer: a data-driven discussion of DTC DNA ancestry tests. Genet. Med. 14:586–593.

Wall J, Cox M, Mendez F, Woerner A, Severson T, Hammer M. 2008. A novel DNA sequence database for analyzing human demographic history. Genome Research 18:1354–1361.

Watkins WS, Xing J, Huff C, Witherspoon DJ, Zhang Y, Perego UA, Woodward SR, Jorde LB. 2012. Genetic analysis of ancestry, admixture and selection in Bolivian and Totonac populations of the New World. BMC Genet 13:39.

Weir BS. 1996. Genetic Data Analysis II. Sunderland, MA: Sinauer

Wen B, Xie X, Gao S, *et al.* 2004. Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. American Journal of Human Genetics 74:856–865.

Whitlock MC. 2005. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. J Evolution Biol 18:1368–1373.

Wilder J, Kingan S, Mobasher Z, Pilkington M, Hammer MF. 2004. Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by migration rate of females versus males. Nat Genet 36:1122–1125.

Wilder J, Mobasher Z, Hammer M. 2004. Genetic evidence for unequal effective population sizes of human females and males. Molecular Biology and Evolution

21:2047–2056.

Wilkins J, Marlowe F. 2006. Sex-biased migration in humans: what should we expect from genetic data? Bioessays 28:290–300.

Wong C, Li Y, Lee C, Huang C-H. 2011. Ensemble learning algorithms for classification of mtDNA into haplogroups. Briefings in bioinformatics 12:1–9.

Wood E, Stover D, Ehret C, Destro-Bisol G. 2005. Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. Eur J Hum Genet 13:867–876.

Wright S. 1950. Genetical structure of populations. Nature.

Zhong M, Lange K, Papp JC, Fan R. 2010. A powerful score test to detect positive selection in genome-wide scans. Eur J Hum Genet 18:1148–1159.

Zhong M, Zhang Y, Lange K, Fan R. 2011. A Cross-Population Extended Haplotype-based Homozygosity Score Test to Detect Positive Selection in Genome-wide Scans. Statistics and its Interface 4:51–63.

Appendix A. Supplementary material for Chapter 2

A.1 Supplementary Figures

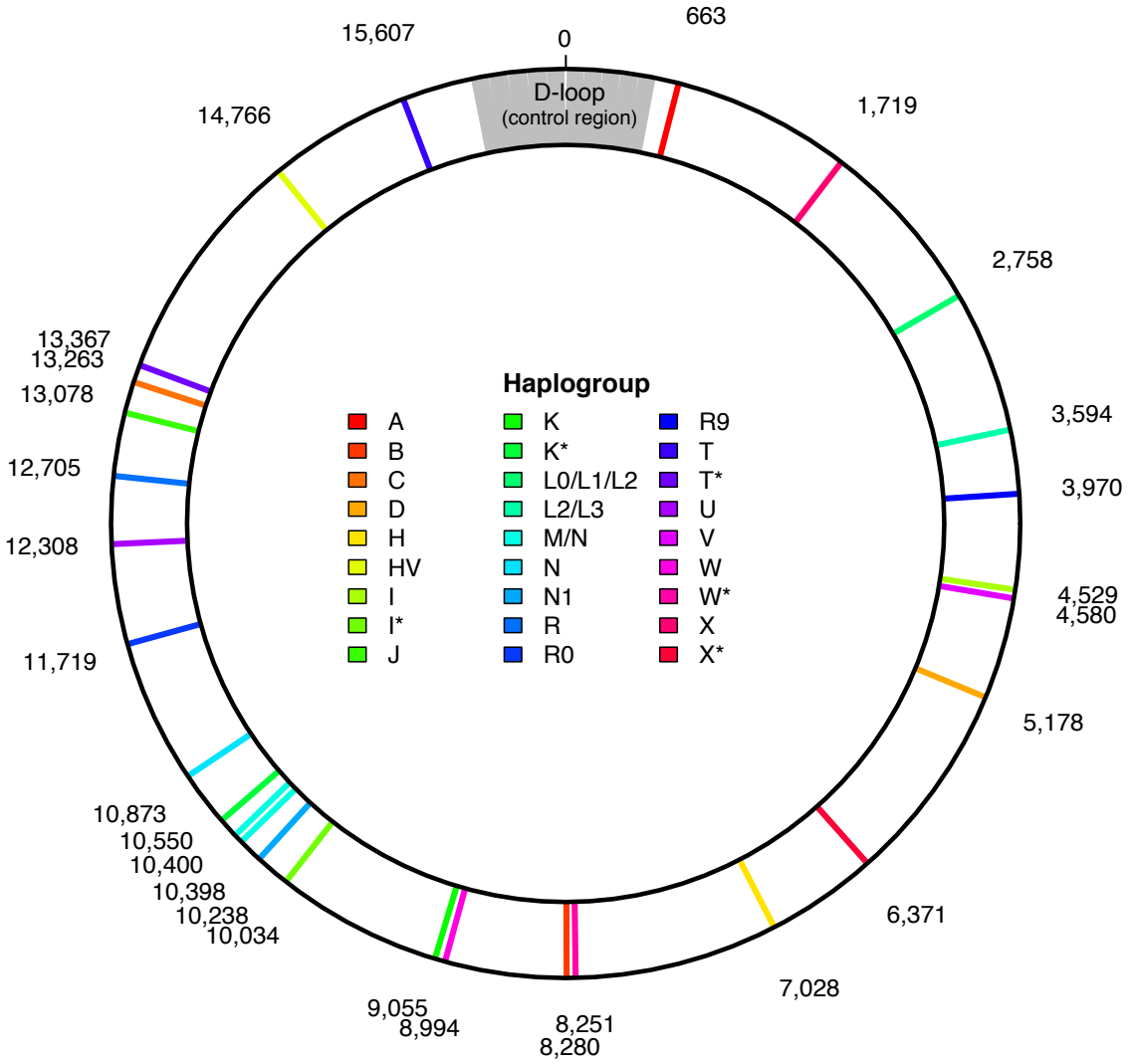


Figure A.1 mtDNA haplogroup diagnostic SNP locations
Each SNP used to diagnose particular mtDNA haplogroups is plotted in context on the mtDNA genome and colored by the haplogroup it is used to diagnose. Each SNP is labeled with its mtDNA genomic coordinate.

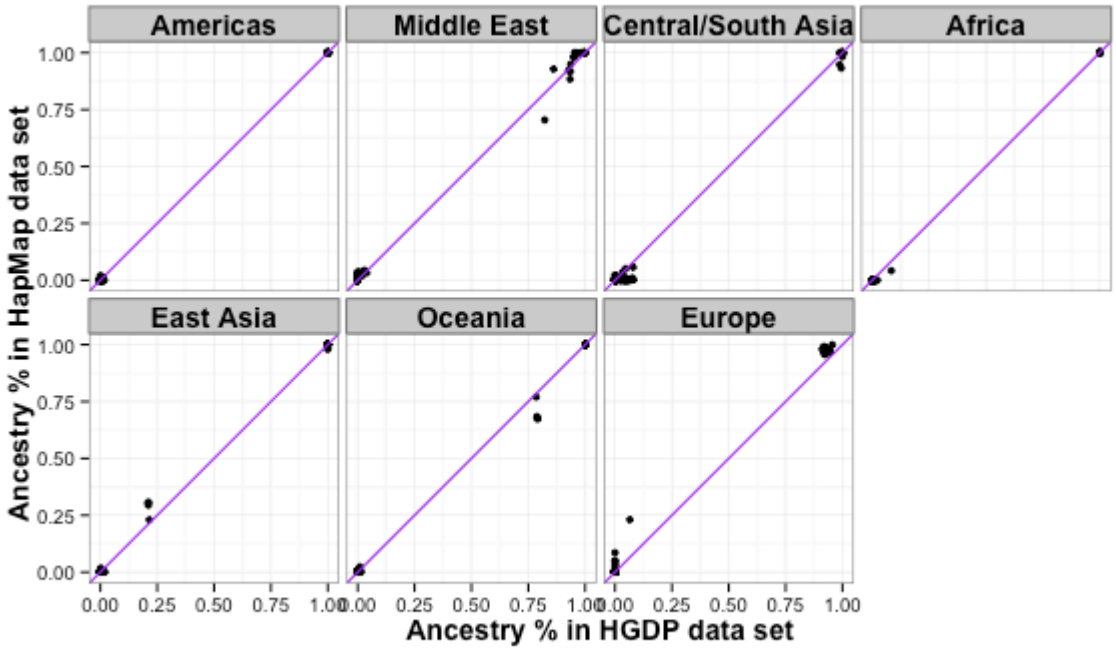


Figure A.2 Correlation between paired continental ancestry estimates in "pseudo-ancestors" from the HGDP dataset.
For the 140 pseudo-ancestors identified from the HGDP, continental ancestry was estimated twice – once in with the HGDP data run in ADMIXTURE and once with the 1,000 Genomes data run in ADMIXTURE. Ancestry estimates from the HGDP run are plotted on the x-axis and those from the 1,000 Genomes are plotted on the y-axis. Each panel is the plot for ancestry percentages in a single continental region. Purples lines are best-fit lines for the data. R^2 values for each continental region: Americas - 0.9999, Middle East – 0.9989, Central/South Asia – 0.9983, Africa – 0.9999, East Asia – 0.9999, Oceania – 0.9989, Europe – 0.9987. p-values for all continental regions are < 0.00001

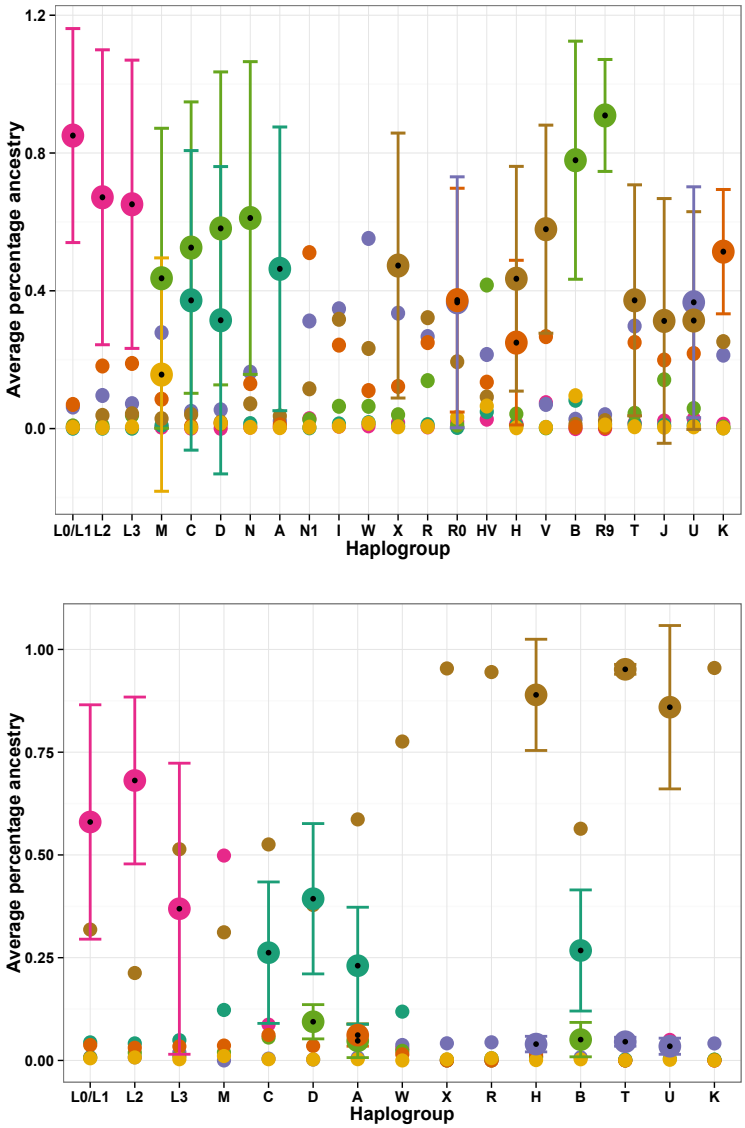


Figure A.3 Associations between mtDNA haplogroups and high continental ancestry percentages
Colored circles are the average continental ancestry percentages within each mtDNA haplogroup. Larger colored circles with black points denote average continental ancestry percentages that are higher within the haplogroup than expected by chance, according to a permutation test shuffling haplogroup labels (1,000 repetitions). These significant points also have bars indicating one standard deviation on either side of the average. Top: HGDP dataset Bottom: 1,000 Genomes dataset

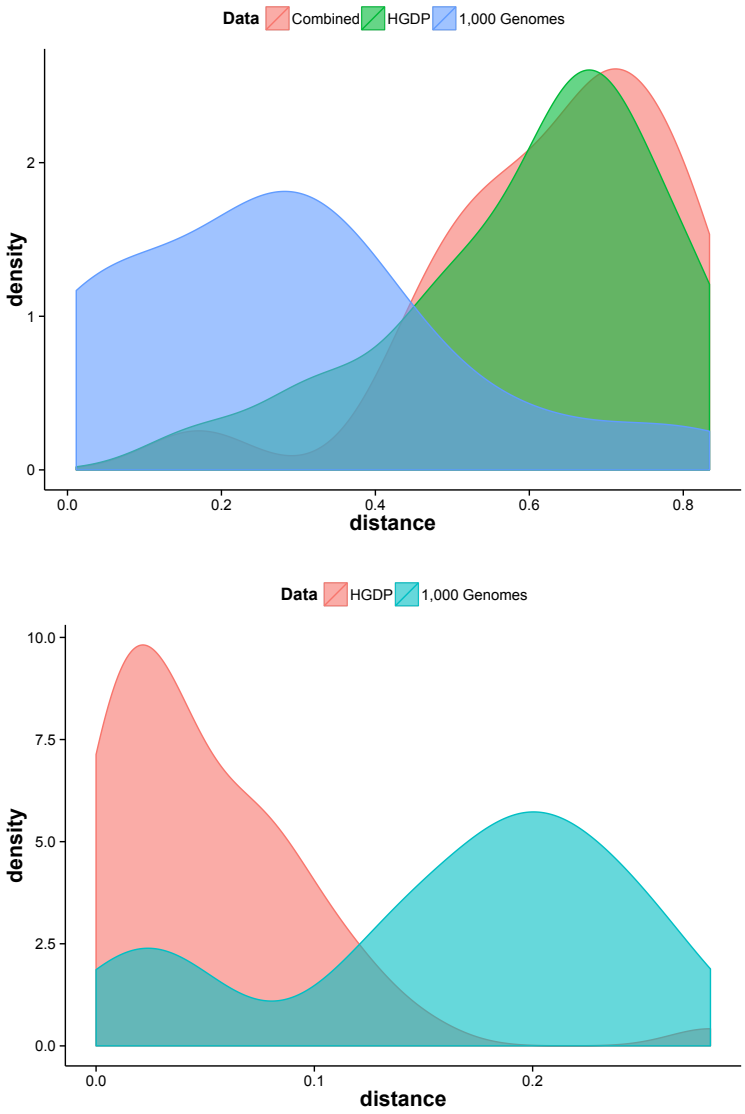


Figure A.4 Mean pairwise Euclidean ancestry distances compared between datasets, by haplogroup and by population

Each colored curve is the density distribution for the average pairwise continental ancestry distances within the dataset of the corresponding color. For within-haplogroup distances, both datasets can be combined (red). Unique populations in each dataset prevent a similar aggregation of within-population distances. Top: distances are calculated within each mtDNA haplogroup. Bottom: distances are calculated within each population.

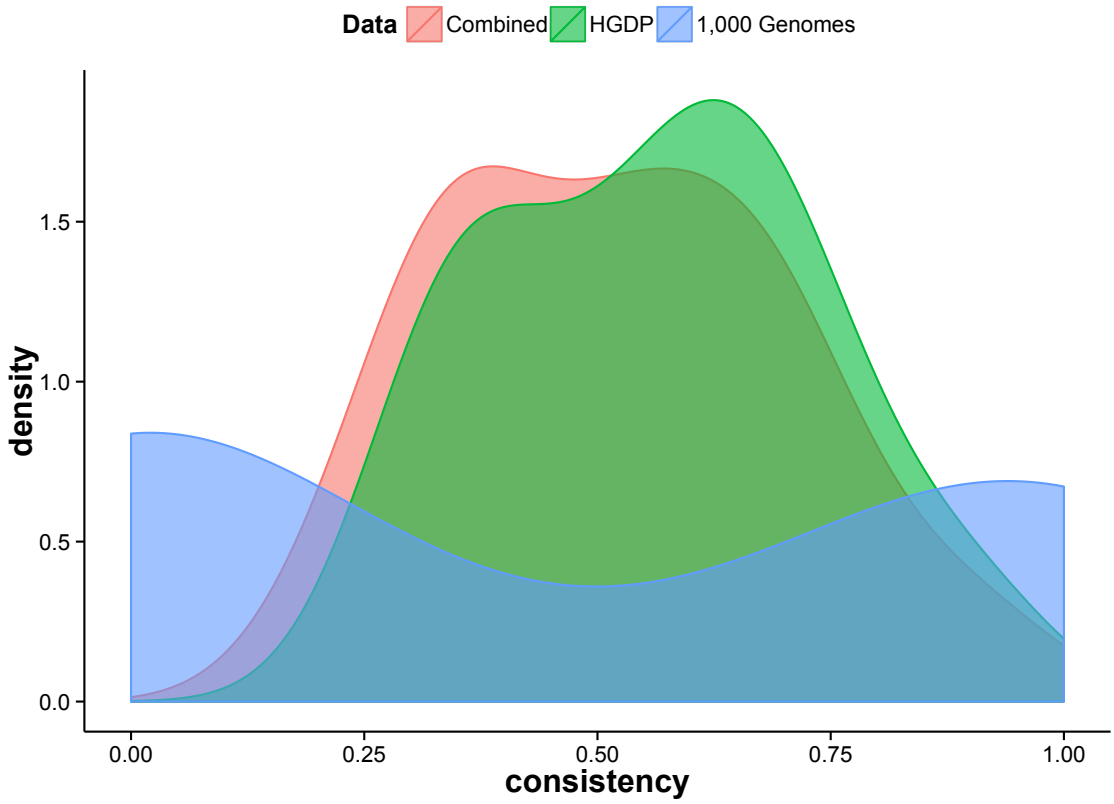


Figure A.5 Distribution of consistency scores for each mtDNA haplogroup, compared between datasets.
Each colored curve is the density distribution of the consistency scores within each of the datasets, or both datasets combined. Consistency is the proportion of people within the haplogroup whose highest continental ancestry percentage is the same as the haplogroup's highest average continental ancestry in the HGDP.

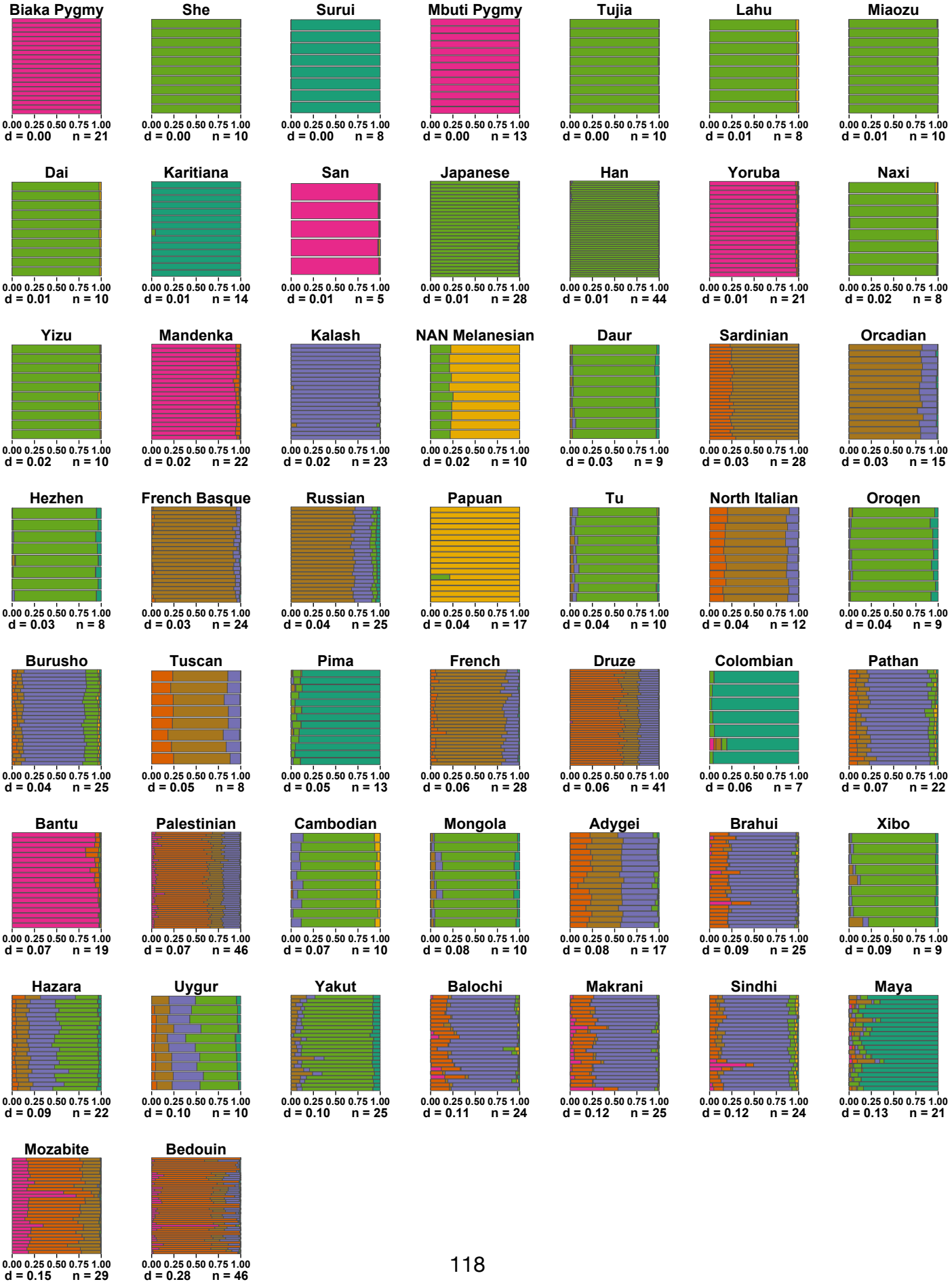

Biaka Pygmy
0.00 0.25 0.50 0.75 1.00
d = 0.00 n = 21
She
0.00 0.25 0.50 0.75 1.00
d = 0.00 n = 10
Surui
0.00 0.25 0.50 0.75 1.00
d = 0.00 n = 8
Mbuti Pygmy
0.00 0.25 0.50 0.75 1.00
d = 0.00 n = 13
Tujia
0.00 0.25 0.50 0.75 1.00
d = 0.00 n = 10
Lahu
0.00 0.25 0.50 0.75 1.00
d = 0.01 n = 8
Miaozu
0.00 0.25 0.50 0.75 1.00
d = 0.01 n = 10
Dai
0.00 0.25 0.50 0.75 1.00
d = 0.01 n = 10
Karitiana
0.00 0.25 0.50 0.75 1.00
d = 0.01 n = 14
San
0.00 0.25 0.50 0.75 1.00
d = 0.01 n = 5
Japanese
0.00 0.25 0.50 0.75 1.00
d = 0.01 n = 28
Han
0.00 0.25 0.50 0.75 1.00
d = 0.01 n = 44
Yoruba
0.00 0.25 0.50 0.75 1.00
d = 0.01 n = 21
Naxi
0.00 0.25 0.50 0.75 1.00
d = 0.02 n = 8
Yizu
0.00 0.25 0.50 0.75 1.00
d = 0.02 n = 10
Mandenka
0.00 0.25 0.50 0.75 1.00
d = 0.02 n = 22
Kalash
0.00 0.25 0.50 0.75 1.00
d = 0.02 n = 23
NAN Melanesian
0.00 0.25 0.50 0.75 1.00
d = 0.02 n = 10
Daur
0.00 0.25 0.50 0.75 1.00
d = 0.03 n = 9
Sardinian
0.00 0.25 0.50 0.75 1.00
d = 0.03 n = 28
Orcadian
0.00 0.25 0.50 0.75 1.00
d = 0.03 n = 15
Hezhen
0.00 0.25 0.50 0.75 1.00
d = 0.03 n = 8
French Basque
0.00 0.25 0.50 0.75 1.00
d = 0.03 n = 24
Russian
0.00 0.25 0.50 0.75 1.00
d = 0.04 n = 25
Papuan
0.00 0.25 0.50 0.75 1.00
d = 0.04 n = 17
Tu
0.00 0.25 0.50 0.75 1.00
d = 0.04 n = 10
North Italian
0.00 0.25 0.50 0.75 1.00
d = 0.04 n = 12
Oroqen
0.00 0.25 0.50 0.75 1.00
d = 0.04 n = 9
Burusho
0.00 0.25 0.50 0.75 1.00
d = 0.04 n = 25
Tuscan
0.00 0.25 0.50 0.75 1.00
d = 0.05 n = 8
Pima
0.00 0.25 0.50 0.75 1.00
d = 0.05 n = 13
French
0.00 0.25 0.50 0.75 1.00
d = 0.06 n = 28
Druze
0.00 0.25 0.50 0.75 1.00
d = 0.06 n = 41
Colombian
0.00 0.25 0.50 0.75 1.00
d = 0.06 n = 7
Pathan
0.00 0.25 0.50 0.75 1.00
d = 0.07 n = 22
Bantu
0.00 0.25 0.50 0.75 1.00
d = 0.07 n = 19
Palestinian
0.00 0.25 0.50 0.75 1.00
d = 0.07 n = 46
Cambodian
0.00 0.25 0.50 0.75 1.00
d = 0.07 n = 10
Mongola
0.00 0.25 0.50 0.75 1.00
d = 0.08 n = 10
Adygei
0.00 0.25 0.50 0.75 1.00
d = 0.08 n = 17
Brahui
0.00 0.25 0.50 0.75 1.00
d = 0.09 n = 25
Xibo
0.00 0.25 0.50 0.75 1.00
d = 0.09 n = 9
Hazara
0.00 0.25 0.50 0.75 1.00
d = 0.09 n = 22
Uygur
0.00 0.25 0.50 0.75 1.00
d = 0.10 n = 10
Yakut
0.00 0.25 0.50 0.75 1.00
d = 0.10 n = 25
Balochi
0.00 0.25 0.50 0.75 1.00
d = 0.11 n = 24
Makrani
0.00 0.25 0.50 0.75 1.00
d = 0.12 n = 25
Sindhi
0.00 0.25 0.50 0.75 1.00
d = 0.12 n = 24
Maya
0.00 0.25 0.50 0.75 1.00
d = 0.13 n = 21
Mozabite
0.00 0.25 0.50 0.75 1.00
d = 0.15 n = 29
Bedouin
0.00 0.25 0.50 0.75 1.00
d = 0.28 n = 46

Figure A.6 Individual continental ancestry percentages within HGDP populations For each HGDP population, individual samples are plotted as a barplot of continental ancestry percentages, colored according to the labels in main text Fig. 1A. Under the x-axis the average pairwise ancestry distance (d) and sample size (n) of the population are shown. Populations are sorted in ascending order by mean pairwise Euclidean distance, from left to right and top to bottom.

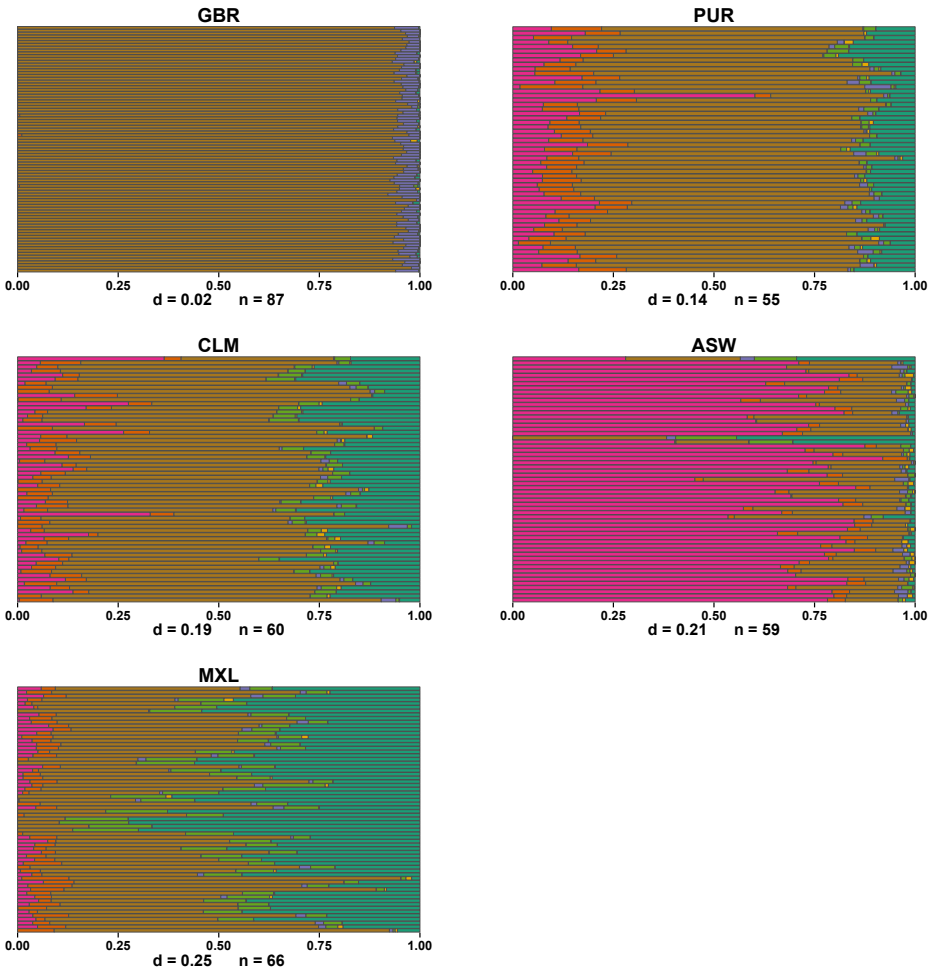


Figure A.7 Individual continental ancestry percentages within 1,000 Genomes populations

For each 1,000 Genomes population, individual samples are plotted as a barplot of continental ancestry percentages, colored according to the labels in main text Fig. 1A. Under the x-axis the average pairwise ancestry distance (d) and sample size (n) of the population are shown. Populations are sorted in ascending distance order, from left to right and top to bottom.

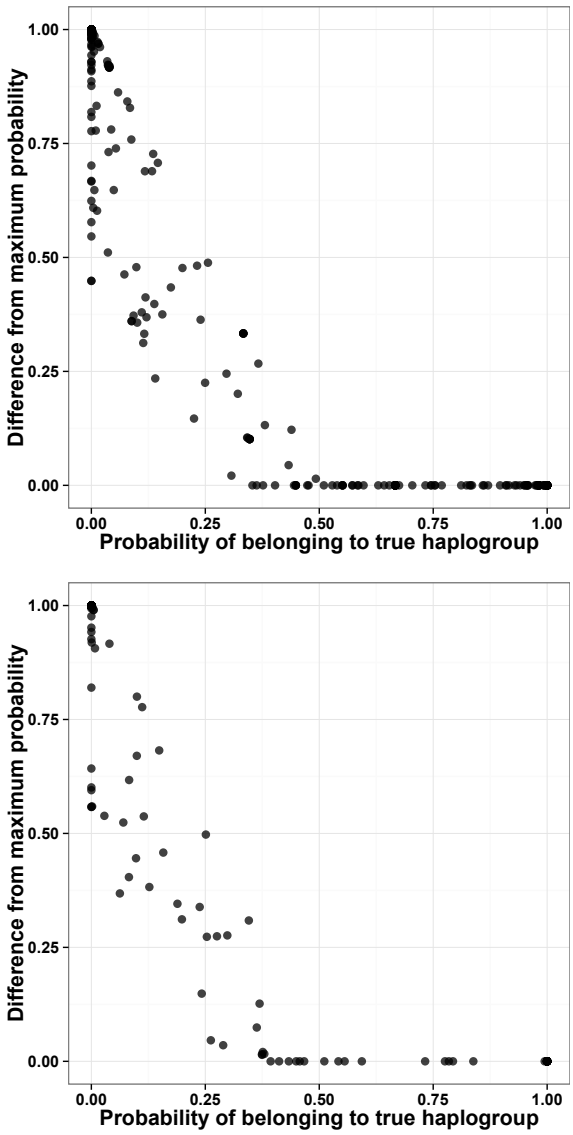


Figure A.8 Strong false predictions and strong true predictions in classification probabilities from the fitted logit model

Each point is a sample from the HGDP (top) or 1,000 Genomes (bottom) dataset. A point's x coordinate is the fitted logit classification probability for the sample's experimentally determined "true" mtDNA haplogroup. A point's y coordinate is the difference between this true classification probability and the sample's highest fitted logit classification probability (the prediction). If a sample's predicted haplogroup is the same as its true haplogroup, then the y-axis value will be zero. Higher y values indicate higher misclassification probabilities. Points in the lower right corner of the plot are strong correct predictions (most desirable). Points in the upper left corner are strong false predictions (least desirable).

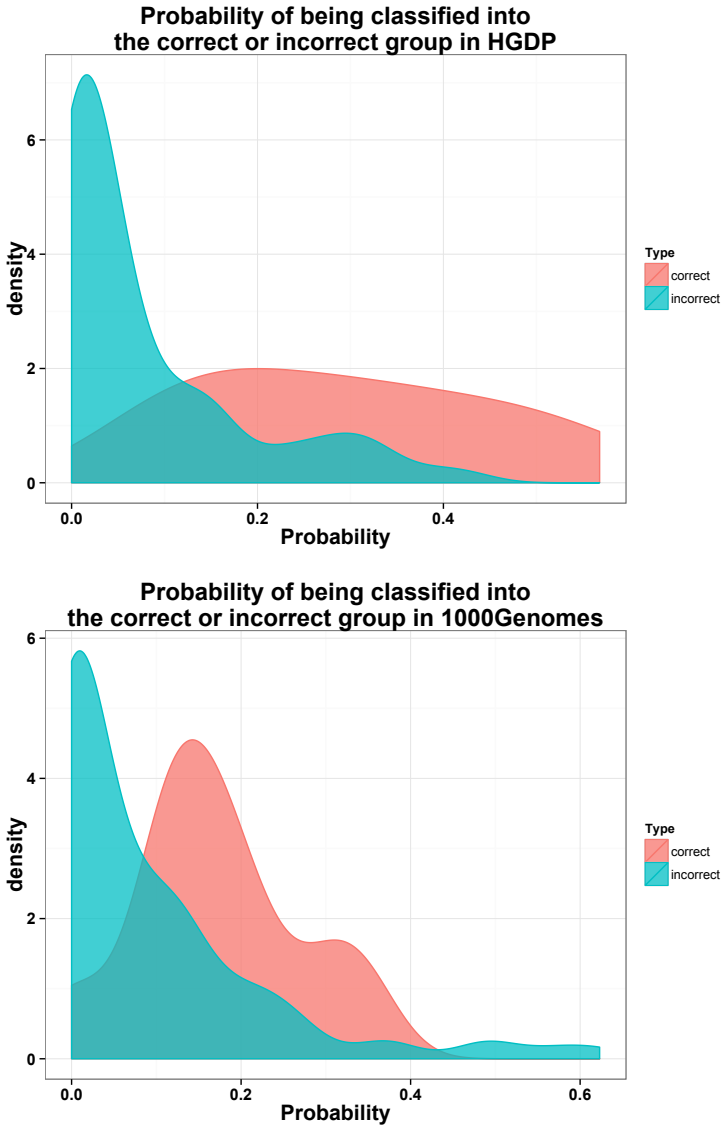


Figure A.9 Correct vs. incorrect classification probabilities compared between datasets
Each colored curve is the density distribution of all fitted logit classification probabilities for each sample in the HGDP (top) or in the 1,000 Genomes (bottom). Probabilities for a sample's true haplogroup are the correct probabilities, shown in red. Probabilities for all other haplogroups are the incorrect probabilities, shown in turquoise. The distribution of correct probabilities is significantly different from the incorrect probabilities in the HGDP (t test, $p = 0.001$) and in the 1,000 Genomes (t test, $p = 0.02$).

A.2 Supplementary Tables

Table A.1 Average continental ancestry proportions

A. HGDP

Haplogroup	Americas	Middle.East	Central.South.Asia	Africa	East.Asia	Oceania	Europe
L0\L1	0.11%	7.00%	6.26%	85.06%	0.78%	0.37%	0.42%
L2	0.12%	18.17%	9.61%	67.15%	0.88%	0.24%	3.83%
L3	0.12%	18.91%	7.24%	65.12%	3.82%	0.49%	4.31%
M	0.98%	8.54%	27.86%	0.42%	43.64%	15.68%	2.89%
C	37.23%	0.70%	5.00%	0.12%	52.54%	0.32%	4.08%
D	31.45%	1.40%	5.46%	0.08%	58.11%	1.65%	1.86%
N	1.44%	13.06%	16.38%	0.43%	61.13%	0.37%	7.20%
A	46.38%	1.12%	2.21%	0.41%	45.94%	0.28%	3.67%
N1	0.24%	51.08%	31.23%	2.88%	2.61%	0.41%	11.55%
I	1.36%	24.23%	34.78%	0.71%	6.48%	0.72%	31.73%
W	1.73%	11.04%	55.20%	0.80%	6.46%	1.51%	23.26%
X	0.71%	12.22%	33.51%	1.68%	4.02%	0.52%	47.33%
R	1.14%	24.96%	26.71%	0.50%	13.92%	0.57%	32.20%
R0	0.33%	37.29%	36.66%	2.02%	1.03%	3.27%	19.40%
HV	4.88%	13.48%	21.51%	2.73%	41.67%	6.55%	9.19%
H	0.88%	24.96%	24.75%	1.42%	4.24%	0.25%	43.50%
V	0.32%	26.70%	7.01%	7.52%	0.22%	0.35%	57.89%
B	8.22%	0.39%	2.71%	0.00%	77.91%	9.54%	1.23%
R9	1.37%	0.34%	4.03%	0.00%	90.90%	1.04%	2.31%
T	1.35%	25.02%	29.78%	1.64%	4.44%	0.54%	37.24%
J	0.99%	19.90%	31.07%	2.20%	14.18%	0.41%	31.26%
U	1.00%	21.80%	36.67%	2.79%	5.84%	0.52%	31.37%
K	0.23%	51.36%	21.32%	1.26%	0.41%	0.17%	25.25%

B. 1,000 Genomes

Haplogroup	Americas	Middle.East	Central.South.Asia	Africa	East.Asia	Oceania	Europe
L0\L1	4.39%	3.74%	0.80%	58.02%	0.68%	0.52%	31.84%
L2	4.15%	3.09%	0.76%	68.12%	1.88%	0.73%	21.27%
L3	4.93%	3.42%	1.97%	36.91%	1.08%	0.29%	51.39%
M	12.28%	3.62%	0.00%	49.85%	1.99%	1.07%	31.18%
C	26.22%	6.20%	0.44%	8.71%	5.58%	0.30%	52.56%
D	39.34%	3.53%	0.22%	9.44%	9.41%	0.24%	37.82%
A	23.05%	6.16%	0.70%	6.36%	4.77%	0.31%	58.65%
W	11.88%	1.49%	3.76%	2.95%	2.34%	0.00%	77.60%
X	0.18%	0.00%	4.17%	0.00%	0.00%	0.27%	95.38%
R	0.54%	0.00%	4.41%	0.00%	0.00%	0.55%	94.50%
H	3.60%	1.60%	3.98%	0.93%	0.86%	0.09%	88.95%
V	27.24%	6.72%	0.64%	3.03%	4.22%	0.00%	58.15%
B	26.75%	5.66%	0.75%	5.10%	5.07%	0.28%	56.38%
T	0.24%	0.00%	4.55%	0.00%	0.01%	0.03%	95.17%
U	2.39%	2.62%	3.46%	5.00%	0.42%	0.17%	85.94%
K	0.17%	0.00%	4.14%	0.00%	0.19%	0.00%	95.49%

Table A.2 The highest continental ancestry component in each haplogroup, by data set

Haplogroup	Highest in HGDP	Highest in 1,000 Genomes	Highest in all combined samples
A	Americas	Europe	Europe
B	East.Asia	Europe	East.Asia
C	East.Asia	Europe	East.Asia
D	East.Asia	Americas	East.Asia
H	Europe	Europe	Europe
HV	East.Asia	NA	East.Asia
I	Central.South.Asia	NA	Central.South.Asia
J	Europe	NA	Europe
K	Middle.East	Europe	Middle.East
L0\L1	Africa	Africa	Africa
L2	Africa	Africa	Africa
L3	Africa	Europe	Africa
M	East.Asia	Africa	East.Asia
N	East.Asia	NA	East.Asia
N1	Middle.East	NA	Middle.East
R	Europe	Europe	Europe
R0	Middle.East	NA	Middle.East
R9	East.Asia	NA	East.Asia
T	Europe	Europe	Europe
U	Central.South.Asia	Europe	Europe
V	Europe	Europe	Europe
W	Central.South.Asia	Europe	Central.South.Asia
X	Europe	Europe	Europe

Table A.3 The standard deviation of individual continental ancestry percentages for each continental region in each haplogroup.
Cells are colored by the value, with higher values in darker red.

A. HGDP

Haplogroup	Africa	Middle.East	Europe	Central.South.Asia	East.Asia	Oceania	Americas
L0/L1	0.31						
L2	0.43	0.26		0.21			
L3	0.42	0.26					
M		0.15		0.33	0.44	0.34	
C					0.42		0.43
D					0.45		0.45
N		0.25		0.27	0.45		
A					0.45		0.41
N1		0.31	0.10	0.26			
I		0.34	0.33	0.25			
W		0.17	0.27	0.25			
X		0.19	0.38	0.30			
R		0.32	0.35	0.27	0.28		
R0		0.32	0.25	0.36			
HV		0.21	0.17	0.28	0.44		
H		0.24	0.33	0.23			
V		0.25	0.30				
B					0.35	0.25	0.25
R9					0.16		
T		0.29	0.34	0.27			
J		0.27	0.36	0.36	0.30		
U		0.24	0.32	0.34			
K		0.18	0.21	0.14			

B. 1,000 Genomes

Haplogroup	Africa	Middle.East	Europe	Central.South.Asia	East.Asia	Oceania	Americas
L0/L1	0.29		0.21				
L2	0.20		0.12				
L3	0.35		0.34				
M	0.46		0.29				0.16
C	0.09		0.18				0.17
D	0.22		0.16		0.04		0.18
A			0.15				0.14
W			0.30				0.20
X			0.01				
R			0.01				
H			0.14				
V							
B			0.15				0.15
T			0.01				
U			0.20				
K			0.03				

Table A.4 Proportion of people with continental ancestry >50% matching the haplogroup's highest continental ancestry component in the HGDP

Haplogroup	Proportion > 50% in HGDP	Proportion > 50% in 1,000 Genomes	Proportion > 50% in all samples combined	HGDP - 1,000 Genomes
L0\L1	0.852	0.778	0.830	0.075
L2	0.674	0.857	0.719	-0.183
L3	0.656	0.435	0.526	0.221
M	0.394	0.000	0.386	0.394
C	0.537	0.000	0.328	0.537
D	0.616	0.000	0.529	0.616
N	0.636	NA	0.636	NA
A	0.559	0.057	0.221	0.502
N1	0.714	NA	0.714	NA
I	0.333	NA	0.333	NA
W	0.700	0.000	0.538	0.700
X	0.556	1.000	0.714	-0.444
R	0.278	1.000	0.381	-0.722
R0	0.500	NA	0.500	NA
HV	0.402	NA	0.402	NA
H	0.455	0.978	0.608	NA
V	0.600	1.000	0.636	NA
B	0.735	0.000	0.342	0.735
R9	0.926	NA	0.926	NA
T	0.395	1.000	0.511	-0.605
J	0.333	NA	0.333	NA
U	0.325	0.000	0.260	0.325
K	0.778	0.000	0.667	0.778

Appendix B. Supplementary material for Chapter 3

B.1 Derivation of formula for evaluating Q_π

Using a coalescent approach based on previous work (Polanski *et al.* 1998; Pool and Nielsen 2007), we can derive a formula for the expected value of π in a population with a given history. As shown in Supplementary Figure 4, the genealogy of a population can be partitioned into a series of discrete intervals, described by the parameters N_1, N_2, …,N_n and T_1, T_2, …,T_n, where N_i and T_i denote the population size and number of generations for the i^{th} interval, respectively. To derive an analytical formula for evaluating Q_π, we need to express π, the probability that two randomly sampled copies of a locus differ, as a function of the N_i's and T_i's. Assuming biallelic loci and an infinite sites model of evolution, π is equivalent to the probability that a single mutation occurred at some point in the genealogy of a locus (Polanski *et al.* 1998). The probability of a mutation occurring in a given genealogy is equal to $2\mu\tau$ where μ is the mutation rate per site per generation and τ is the time to coalescence for the two sampled loci.

Assuming an infinitely long lineage (*i. e.* the two sampled loci do coalesce eventually), then the total coalescent time, τ, can be determined by considering the contribution of each interval to τ. τ is a function of the coalescent times of each interval, denoted as $\tau_1, \tau_2, \ldots \tau_n$ and the probability, $P_c(i)$, that the two copies will coalesce in the i^{th} interval:

$$\tau = P_c(1)(T_2 + T_3 + ... + T_n + \tau_1) + P_c(2)(T_3 + T_4 + ... + T_n + \tau_2) + ... + P_c(n)\tau_n$$
$$= \sum_{i=1}^{n}\left[P_c(i)\left(\tau_i + \left(\sum_{j=i+1}^{n} T_j\right)\right)\right] \qquad \textbf{Eq. B.1}$$

Note that each interval's contribution to τ is the product of the probability that the two sampled copies coalesce in that interval and the expected coalescent time if coalescence does occur there.

Following standard coalescent theory, going backwards in time, $P_c(i)$ can be approximated as an exponential function, conditional on the probability of not coalescing in all previous segments:

$$P_c(i) = \left(e^{-\left(\sum_{j=i+1}^{n}\frac{T_j}{2N_j}\right)}\right)\left(1 - e^{-\frac{T_i}{2N_i}}\right) \qquad \textbf{Eq. B.2}$$

Note that the first term in Eq. B.2 is the probability of not coalescing in all previous intervals of the lineage and the second term is the probability of coalescing in the i^{th} interval given that coalescence has not occurred previously. For the most recent interval, n, the probability of not coalescing in all previous intervals becomes unity and therefore:

$$P_c(n) = \left(1 - e^{-\frac{T_n}{2N_n}}\right) \qquad \textbf{Eq. B.3}$$

We assume that $T_1 = \infty$ and therefore the probability of coalescing in interval 1 becomes unity, conditional on lack of coalescence in all previous sections and thus:

$$P_c(1) = \left(e^{-\left(\sum_{j=2}^{n} \frac{T_j}{2N_j} \right)} \right) \qquad \textbf{Eq. B.4}$$

Finally, we can express the expected coalescent time for the i^{th} interval, τ_i, in terms of the N and T parameters. Specifically, the expected value of τ_i is:

$$E(\tau_i) = \frac{\int_0^{T_i} x \left(e^{\frac{-x}{2N_i}} \right) dx}{\left(1 - e^{\frac{-T_i}{2Ni}} \right) 2N_i} \qquad \textbf{Eq. B.5}$$

Evaluating the integral, the expected value of τ_i becomes:

$$E(\tau_i) = 2N_i - T_i \left(\frac{e^{\frac{-T_i}{2N_i}}}{1 - e^{\frac{-T_i}{2N_i}}} \right) \qquad \textbf{Eq. B.6}$$

The final interval, where we have assumed that $T_1 = \infty$, yields the following limit:

$$\lim_{T_1 \to \infty} T_1 \left(\frac{e^{\frac{-T_1}{2N_1}}}{1 - e^{\frac{-T_1}{2N_1}}} \right) = 0 \qquad \textbf{Eq. B.7}$$

and therefore $\tau_1 \simeq 2N_1$. By substituting the expressions for τ_i and $P_c(i)$ into Eq. B.1, we obtain the final expression for π as provided in the main text (Eq. 3.4).

B.2 Derivation of formula for evaluating Q_{FST}

We extended the previous model to include two subpopulations that diverged from an ancestral population, which is described by a series of n discrete intervals that are characterized by the parameters T'_i and N'_i. In order to evaluate Q_{FST}, we first derived an expression for the average pairwise divergence, π_{12}, between the two subpopulations, which diverged t generations ago. When one copy of a locus is sampled from each subpopulation, π_{12} is also equivalent to the probability that a mutation occurred on the genealogy of the two sampled copies; however, because we assume no migration, coalescence can only occur in the ancestral population. The expected value of π_{12} is derived in the same way as that of π, but requires a modified form of Eq. B.1:

$$\tau = \sum_{i=1}^{n} \left[P_c(i) \left(2t + \tau_i + \left(\sum_{j=i+1}^{n} T'_j \right) \right) \right] \qquad \textbf{Eq. B.8}$$

Note that $2t$ is added to the expression for the coalescent time in each interval because the genealogy of the two sampled copies will always include a branch for subpopulation 1 and a branch for subpopulation 2, both of length t, where coalescence cannot occur. Substituting Eq. B.8 into Eq. B.1 yields the final equation for the expected value of π_{12} provided in the main text (Eq. 3.5).

Using the formulas for π and π_{12} derived above, we can obtain approximate expected values for F_{ST} between two subpopulations using the following formula from Hudson *et al.* (Hudson *et al.* 1992):

$$F_{ST} = 1 - \frac{H_w}{H_b} \qquad \textbf{Eq. B.9}$$

Note that H_w is the mean of nucleotide diversity in each subpopulation and H_b is the nucleotide diversity in the combined subpopulations:

$$\begin{aligned} H_w &= \tfrac{1}{2}\left(\pi_1 + \pi_2\right) \\ H_b &= \pi_{12} \end{aligned} \qquad \textbf{Eq. B.10}$$

Using these expressions for H_w and H_b we obtain the final formula for F_{ST} in terms of π that is presented as Eq. 3.6 in the main text.

B.3 Supplementary Figures

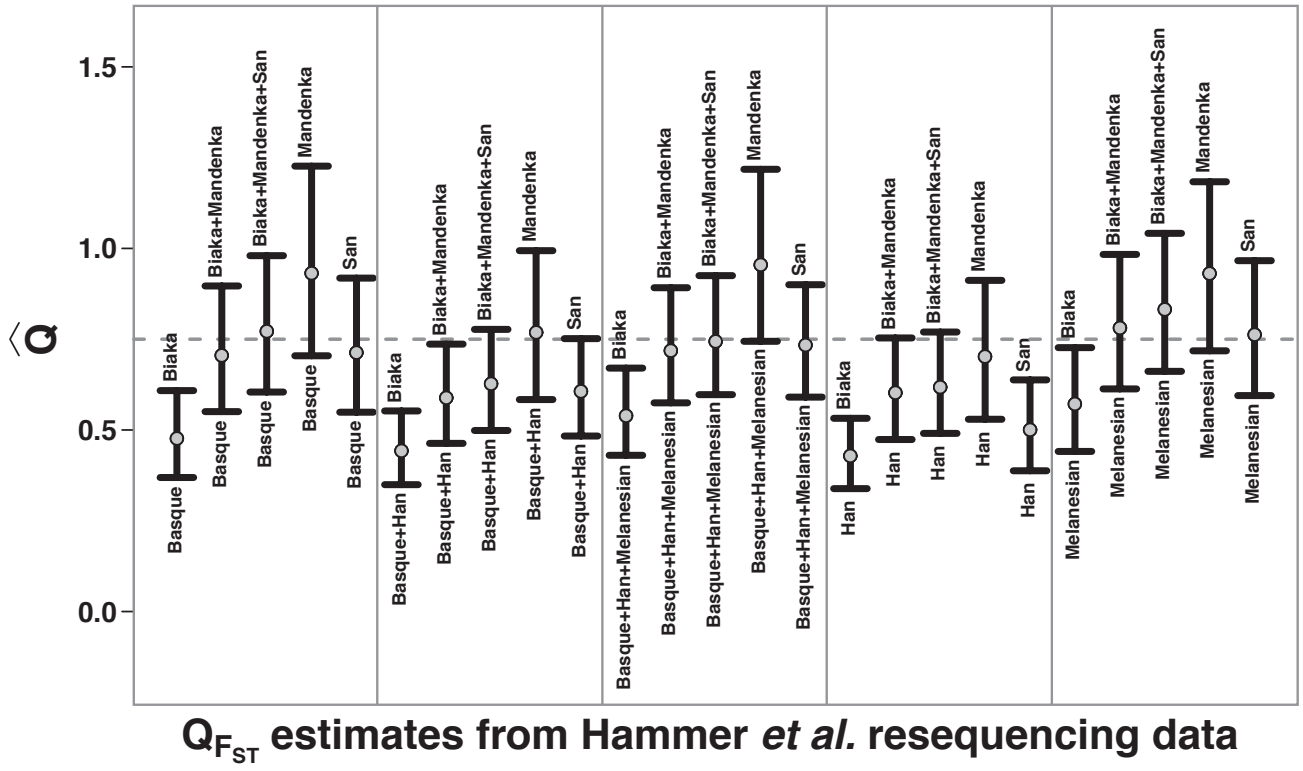
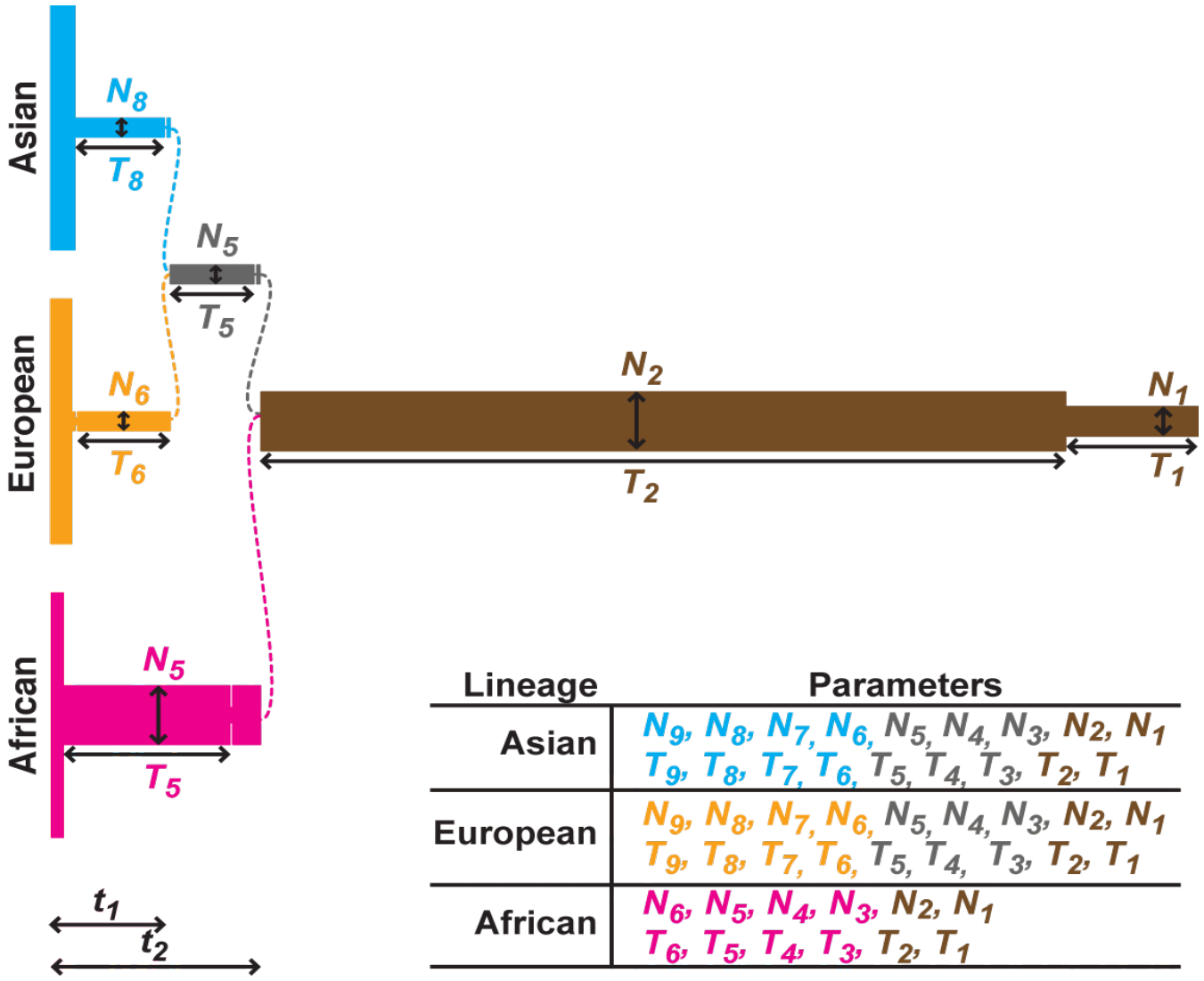


Figure B.1 Estimates of Q_{FST} in the Hammer *et al.* dataset
Nearly all comparisons between African and non-African populations are below or near the expectation of 0.75. Notable exceptions are that most comparisons to the Mandenka are higher than 0.75, and the Melanesian and Basque comparisons are higher than others. The x-axis separates the different comparisons. The two populations being compared are labeled on either end of the black bars. Black bars indicate 95% confidence intervals. The gray dashed line denotes the null expectation of 0.75.



Lineage	Parameters
Asian	$N_9, N_8, N_7, N_6, N_5, N_4, N_3, N_2, N_1$ $T_9, T_8, T_7, T_6, T_5, T_4, T_3, T_2, T_1$
European	$N_9, N_8, N_7, N_6, N_5, N_4, N_3, N_2, N_1$ $T_9, T_8, T_7, T_6, T_5, T_4, T_3, T_2, T_1$
African	$N_6, N_5, N_4, N_3, N_2, N_1$ $T_6, T_5, T_4, T_3, T_2, T_1$

Figure B.2 Coalescent Model of Human Evolution
Going backwards in time, the modern European (gold) and Asian (blue) populations coalesce to form the ancestral non-African population (gray). The African (pink) and non-African populations coalesce to form the ancestral human population (brown). The height of each section is proportional to the population size at that time, as taken from a best-fit model of human evolution (Schaffner *et al.* 2005). The length of each section is proportional to the number of generations spent at that population size. The coalescent history of each of the three populations is described by a set of parameters shown in the table at the bottom right. Each pair of *N* and *T* parameters corresponds to a single section of the coalescent model. Colors indicate which section each parameter describes. Some parameters are labeled for reference. The parameters $t_1$and t_2 are the divergence times, in generations, between the Asian & European and non-African & African populations respectively.

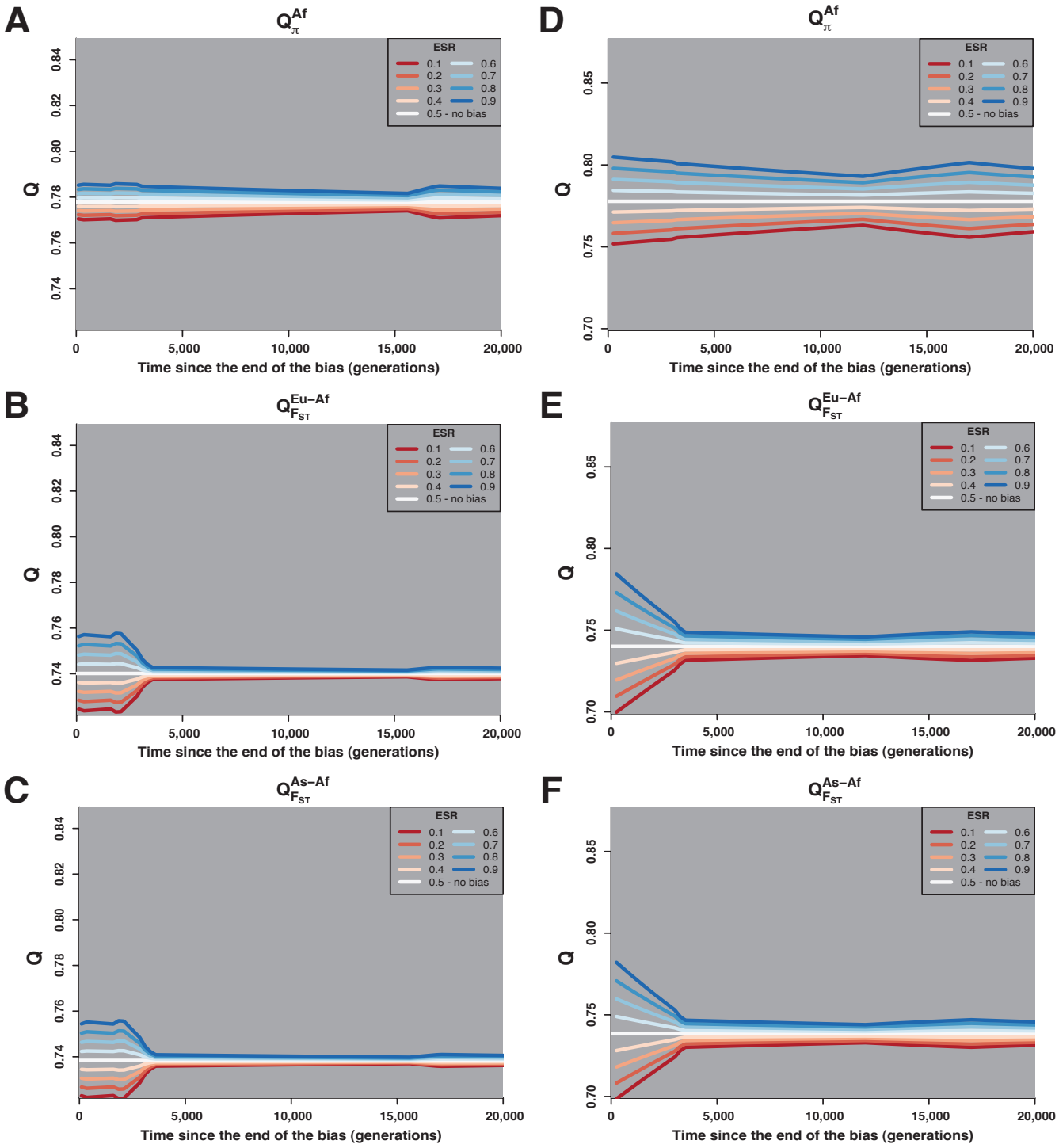


Figure B.3 Varying the Severity of the Sex Bias Introduced into Africans in the Theoretical Model

A sex bias with the ESR denoted by the line's color was introduced into the African lineage and ended *y* generations ago. Each panel shows the indicated *Q* estimator. (A – C) A bias lasting 1,400 generations. (D – F) A bias lasting 5,000 generations.

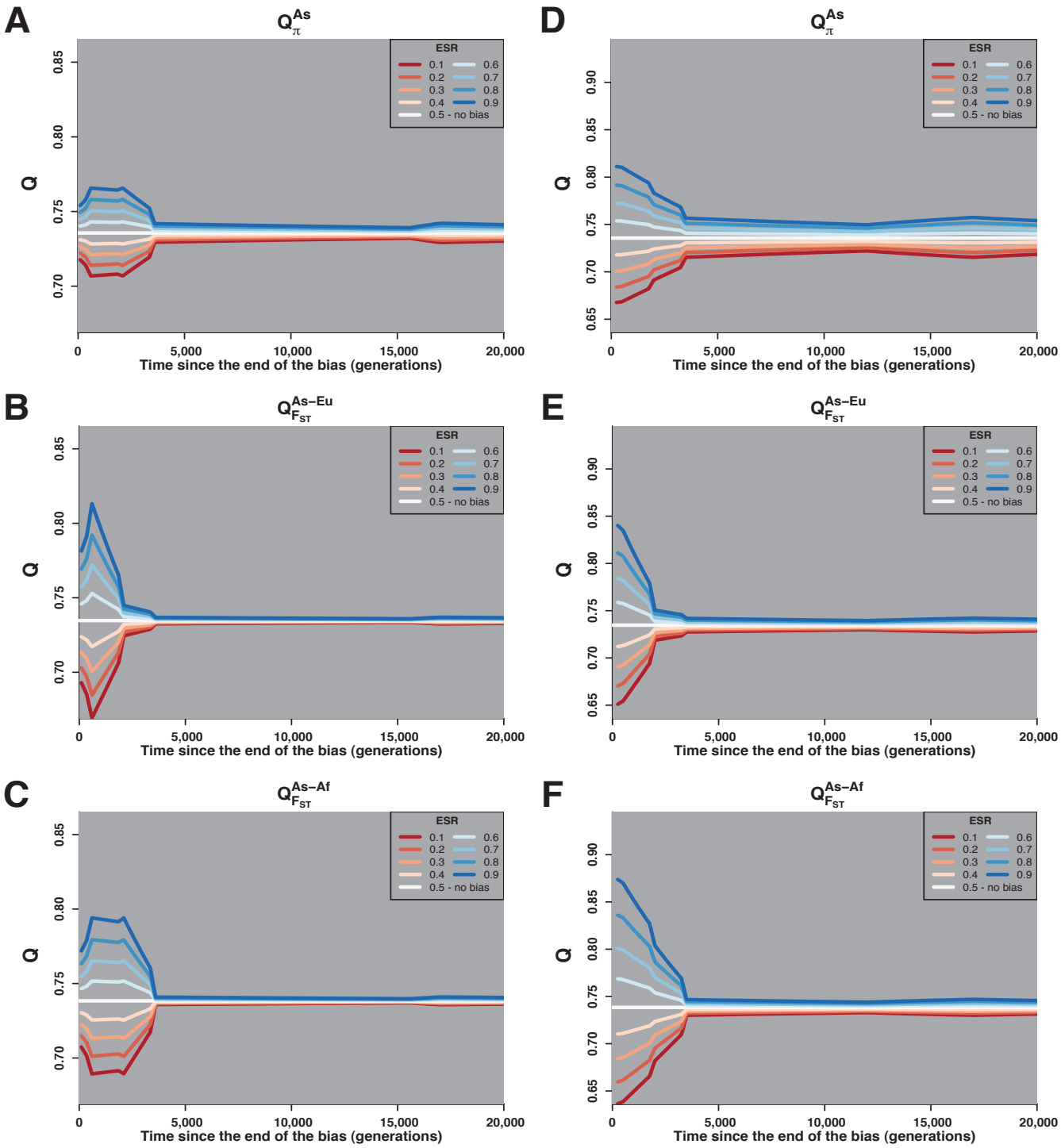


Figure B.4 Varying the Severity of the Sex Bias Introduced into Asians in the Theoretical Model

A sex bias with the ESR given by the line's color was introduced into the Asian lineage and ended *y* generations ago. Each panel shows the indicated *Q* estimator. (A – C) A bias lasting 1,400 generations. (D – F) A bias lasting 5,000 generations.

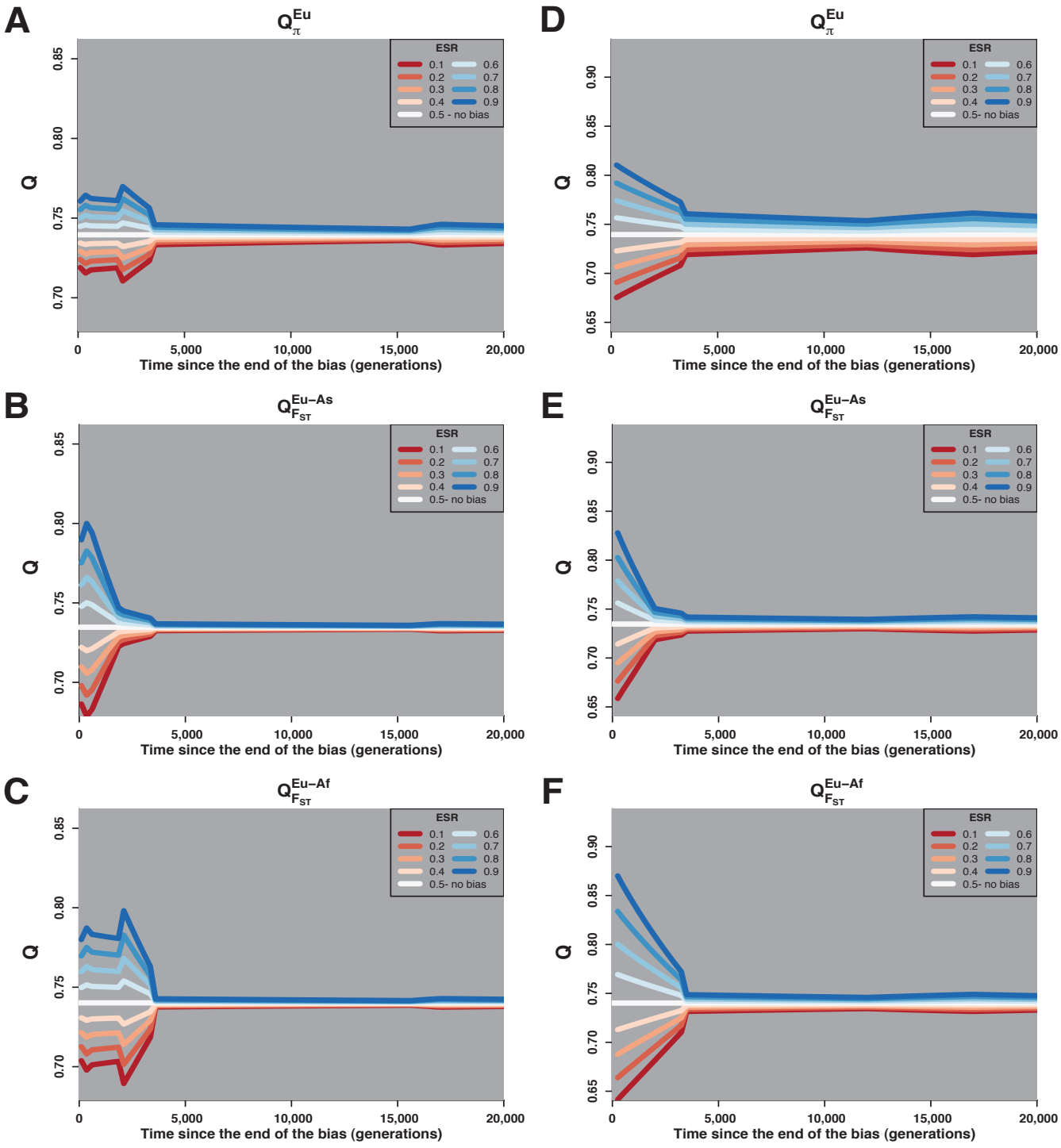


Figure B.5 Varying the Severity of the Sex Bias Introduced into Europeans in the Theoretical Model

A sex bias with the ESR given by the line's color was introduced into the European lineage and ended *y* generations ago. Each panel shows the indicated *Q* estimator. (A – C) A bias lasting 1,400 generations. (D – F) A bias lasting 5,000 generations.

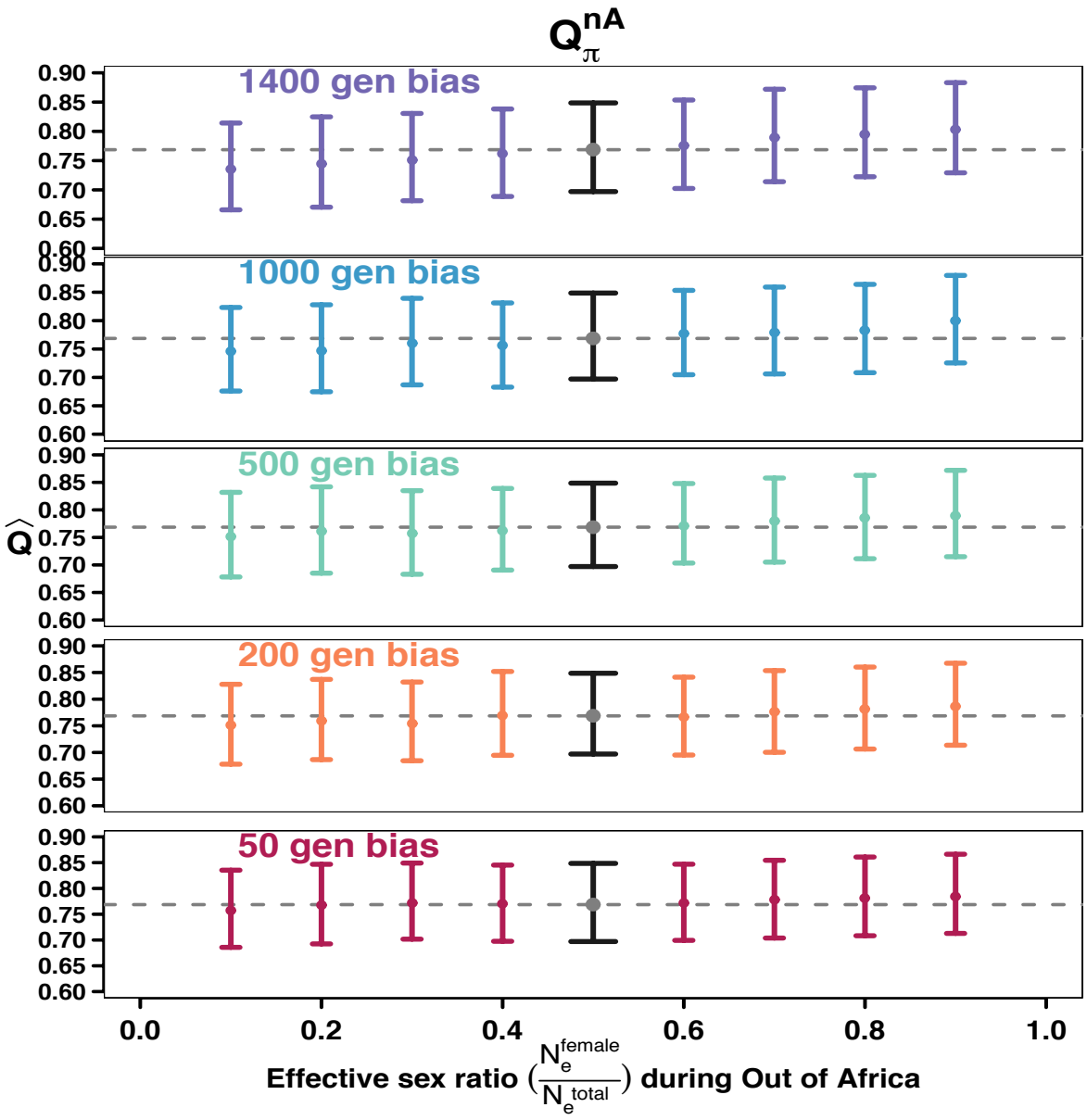


Figure B.6 Q_π in Non-Africans Does Not Detect Recent Sex Biases Associated with the Out of Africa Dispersal

A bias is introduced at the start of the Out of Africa bottleneck event and persists for one of five durations (indicated by different colors; measured in generations). The bias introduced is indicated on the x-axis, measured as the proportion of females in the population. Bars indicate the 95% confidence interval around the point estimate. Black bars and gray points display the estimate of *Q* under no sex bias, which is higher than 0.75 due to the effects of the bottleneck. The gray dashed line corresponds to the corrected expectation when there is no sex bias.

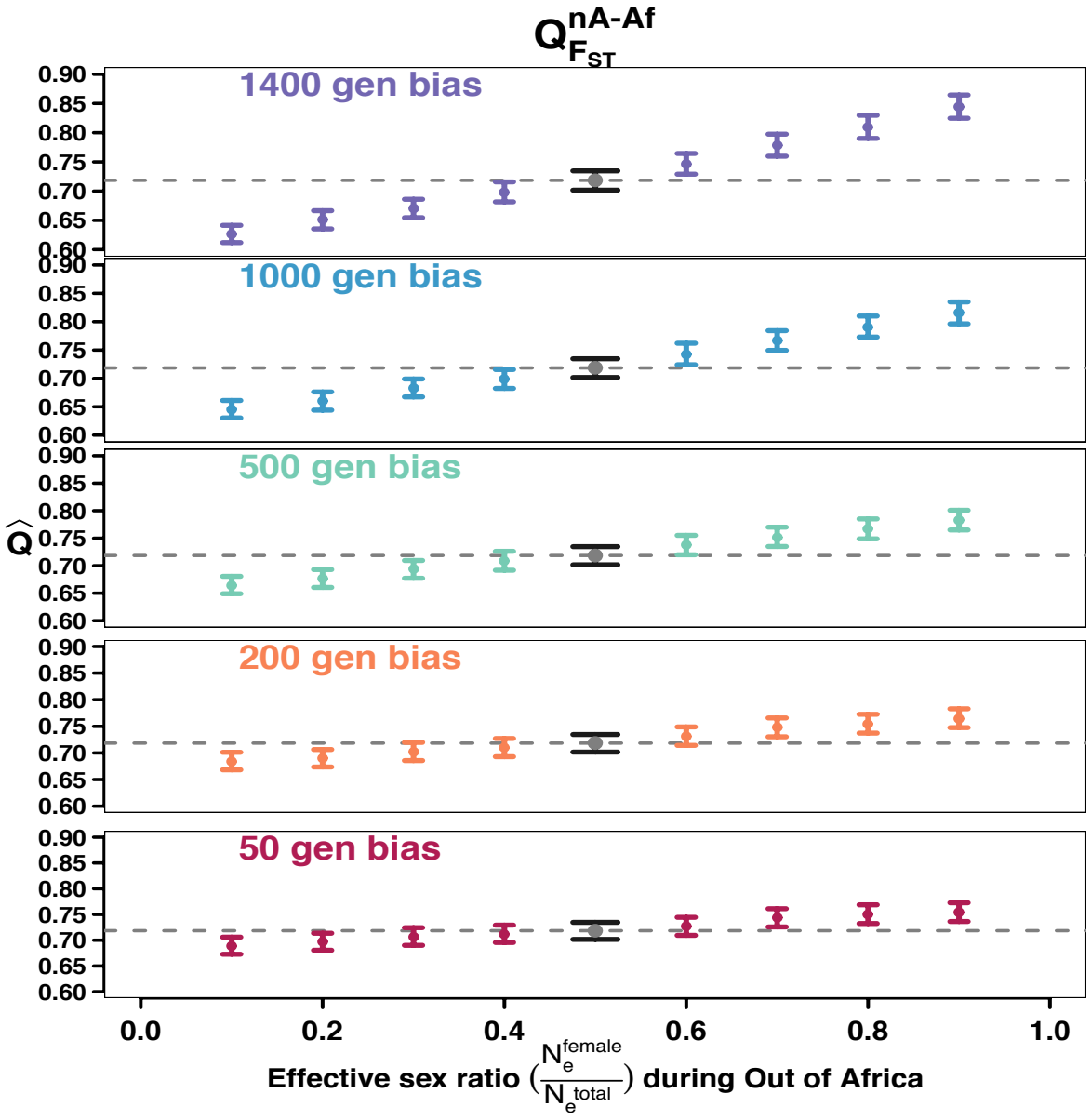


Figure B.7 Using F_{ST} to Estimate Q for the Comparison of Non-Africans to Africans Detects Recent Sex Biases if They Are Extreme

A bias is introduced at the start of the Out of Africa bottleneck event and persists for one of five durations (indicated by different colors; measured in generations; includes the bottleneck). The bias introduced is indicated on the x-axis, measured as the proportion of females in the population. Bars indicate the 95% confidence interval around the point estimate. Black bars and gray points display the estimate of Q under no sex bias, which is lower than 0.75 due to the effects of bottlenecks. The gray dashed line is the corrected expectation with no sex bias.

B.4 Supplementary Tables

Table B.1 Parameters used for demographic models, based on Schaffner *et al.* 2005

Parameter	Value Used
N_e ancestral [b]	12,500
N_e African [b]	24,000
Time of ancestral expansion within Africa	17,000 [a]
Time of Out of Africa split	3,500 [a]
N_e non-African [b]	7,700
Start time of Out of Africa bottleneck	3,450
N_e during Out of Africa bottleneck [b]	588
Duration of Out of Africa bottleneck	50 [a]
Start time of African bottleneck	3,050 [a]
N_e during African bottleneck [b]	6,250
Duration of African bottleneck	50 [a]
Start time of African expansion due to agriculture	200 [a]
Time of European/Asian split	2,000 [a]
Start time of Asian bottleneck	1,950 [a]
N_e during Asian bottleneck [b]	746
Duration of Asian bottleneck	50 [a]
Start time of Asian expansion due to agriculture	400 [a]
Start time of European bottleneck	450 [a]
N_e during European bottleneck [b]	2,500
Duration of European bottleneck	50 [a]
Start time of European expansion due to agriculture	350 [a]
Current N_e in all three populations [b]	100,000
Gene conversion (initiation prob/bp) [c]	4.5×10^{-9}
Mutation rate	1.5×10^{-8}
Recombination rate [c]	1.3 cM/Mb

[a] Time parameters are measured in generations ago
[b] These population sizes were scaled by 0.75 for the X chromosome demographic models
[c] These parameters are used only in the model for the coalescent simulations

Table B.2 Commands for Basic Simulation Models in `ms`

Autosomal resequence data, no sex bias:
`ms` 360 8000 -t 30.00 -r 44.791040 5000 -I 3 120 120 120 -ej 0.005000 3 2 -ej 0.008750 2 1 -en 0.000500 1 0.240000 -en 0.001000 3 0.077000 -en 0.000875 2 0.077000 -en 0.001125 2 0.077000 -en 0.001000 2 0.025000 -en 0.004875 3 0.077000 -en 0.004750 3 0.007460 -en 0.007500 1 0.062500 -en 0.008500 2 0.005880 -en 0.007625 1 0.240000 -en 0.008625 2 0.077000 -en 0.042500 1 0.125000
X-chromosomal resequence data, no sex bias:
`ms` 270 8000 -t 22.50 -r 16.796640 5000 -I 3 90 90 90 -ej 0.006667 3 2 -ej 0.011667 2 1 -en 0.000667 1 0.240000 -en 0.001333 3 0.077000 -en 0.001167 2 0.077000 -en 0.001500 2 0.077000 -en 0.001333 2 0.025000 -en 0.006500 3 0.077000 -en 0.006333 3 0.007460 -en 0.010000 1 0.062500 -en 0.011333 2 0.005880 -en 0.010167 1 0.240000 -en 0.011500 2 0.077000 -en 0.056667 1 0.125000
Autosomal SNP data, no sex bias:
`ms` 360 1500000 -s 1 -I 3 120 120 120 -ej 0.005000 3 2 -ej 0.008750 2 1 -en 0.000500 1 0.240000 -en 0.001000 3 0.077000 -en 0.000875 2 0.077000 -en 0.001125 2 0.077000 -en 0.001000 2 0.025000 -en 0.004875 3 0.077000 -en 0.004750 3 0.007460 -en 0.007500 1 0.062500 -en 0.008500 2 0.005880 -en 0.007625 1 0.240000 -en 0.008625 2 0.077000 -en 0.042500 1 0.125000
X-chromosomal SNP data, no sex bias:
`ms` 270 1500000 -s 1 -I 3 90 90 90 -ej 0.006667 3 2 -ej 0.011667 2 1 -en 0.000667 1 0.240000 -en 0.001333 3 0.077000 -en 0.001167 2 0.077000 -en 0.001500 2 0.077000 -en 0.001333 2 0.025000 -en 0.006500 3 0.077000 -en 0.006333 3 0.007460 -en 0.010000 1 0.062500 -en 0.011333 2 0.005880 -en 0.010167 1 0.240000 -en 0.011500 2 0.077000 -en 0.056667 1 0.125000

Appendix C. Supplementary material for chapter 4

C.1 Supplementary Figures

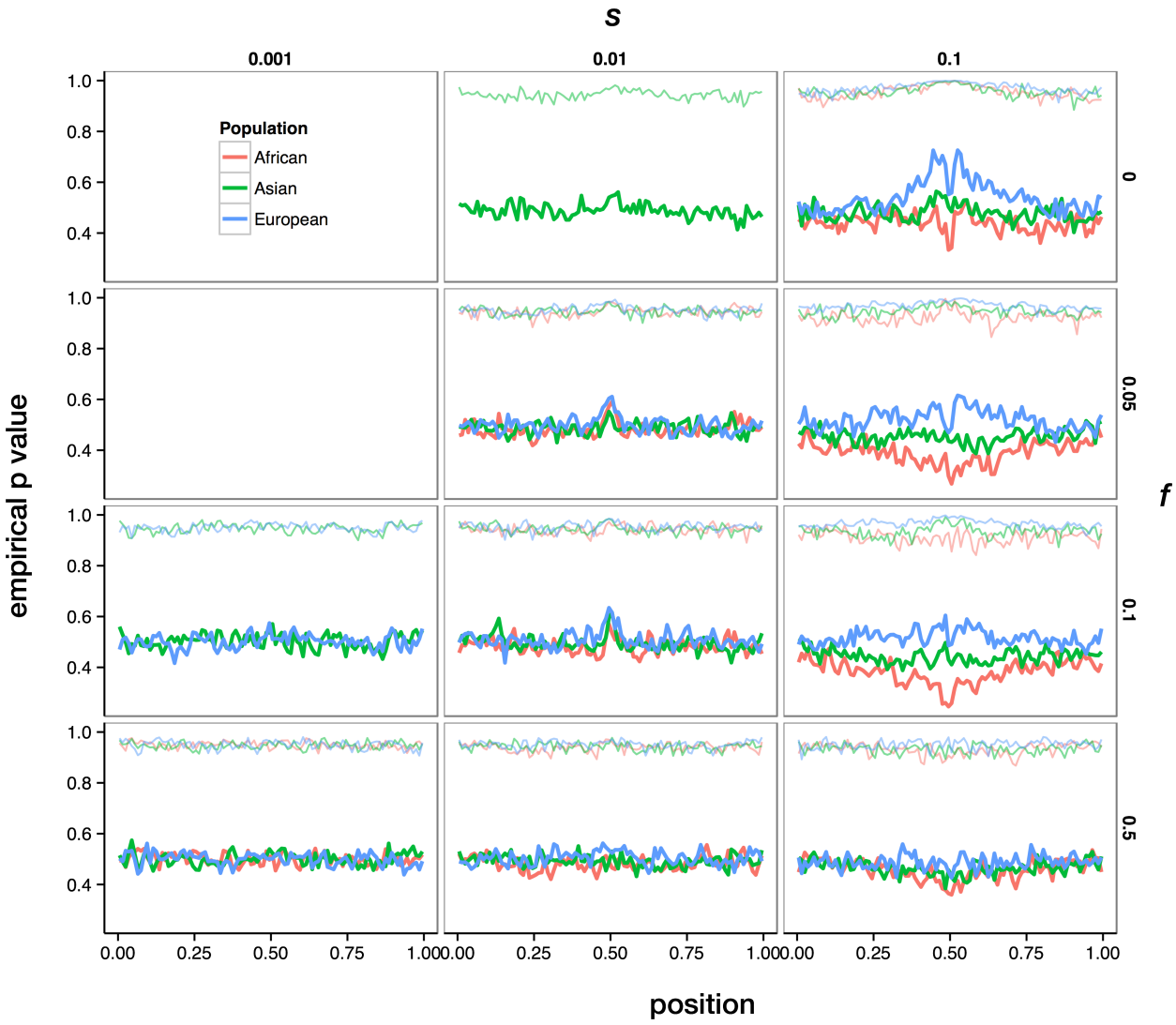


Figure C.1 Binned mean empirical p value for D* depends on *s* and *f* in simulations including selection

For each combination of *s* and *f*, the average empirical p value binned by position is plotted in thick, darker lines. Thin, lighter lines indicate the binned 95% confidence interval. Simulations including selection in each of three populations are shown in each panel where available. Populations are colored according to the legend at top left.

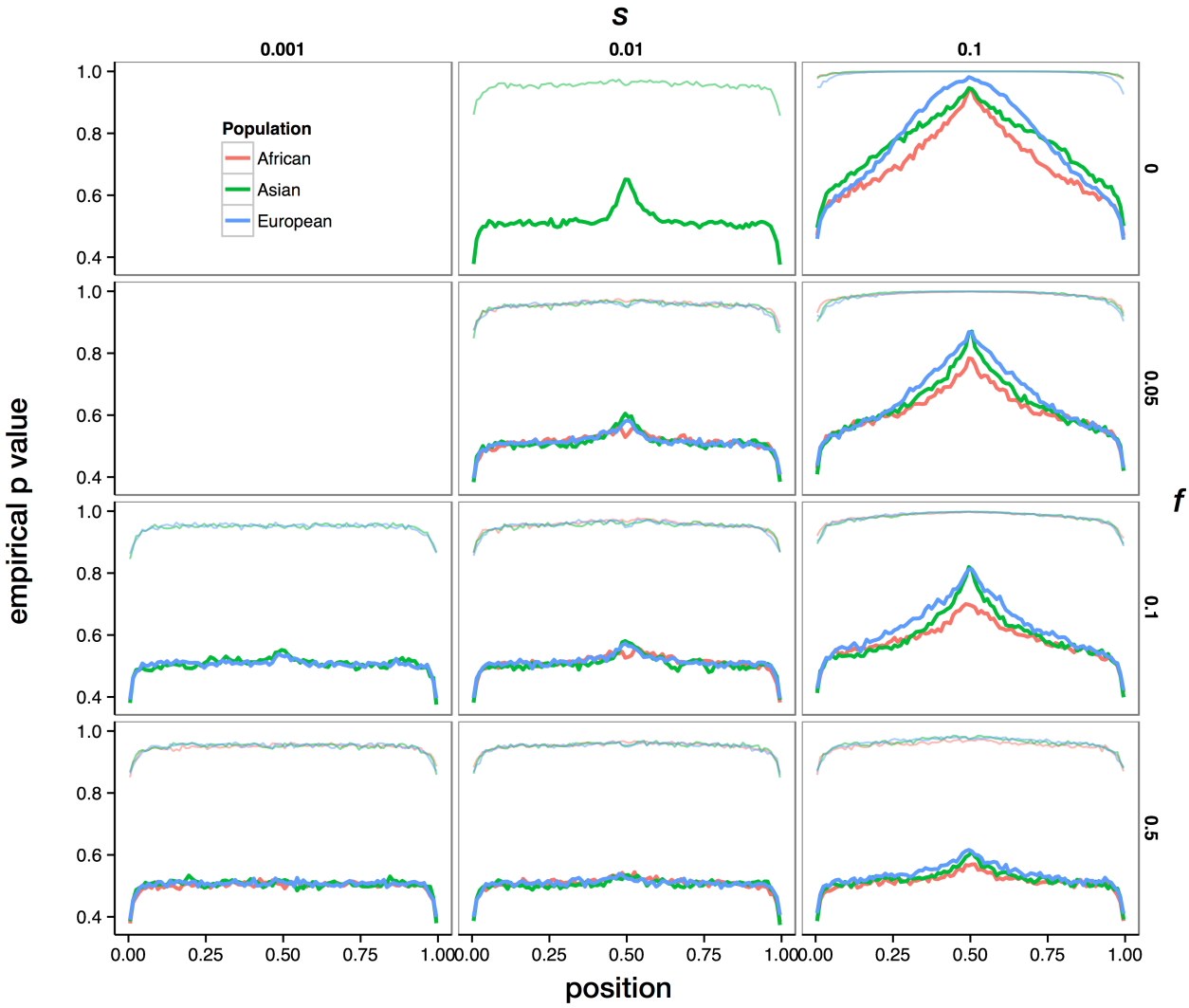


Figure C.2 Binned mean empirical p value for ΔiEHH$_D$ depends on *s* and *f* in simulations including selection

For each combination of *s* and *f*, the average empirical p value binned by position is plotted in thick, darker lines. Thin, lighter lines indicate the binned 95% confidence interval. Simulations including selection in each of three populations are shown in each panel where available. Populations are colored according to the legend at top left.

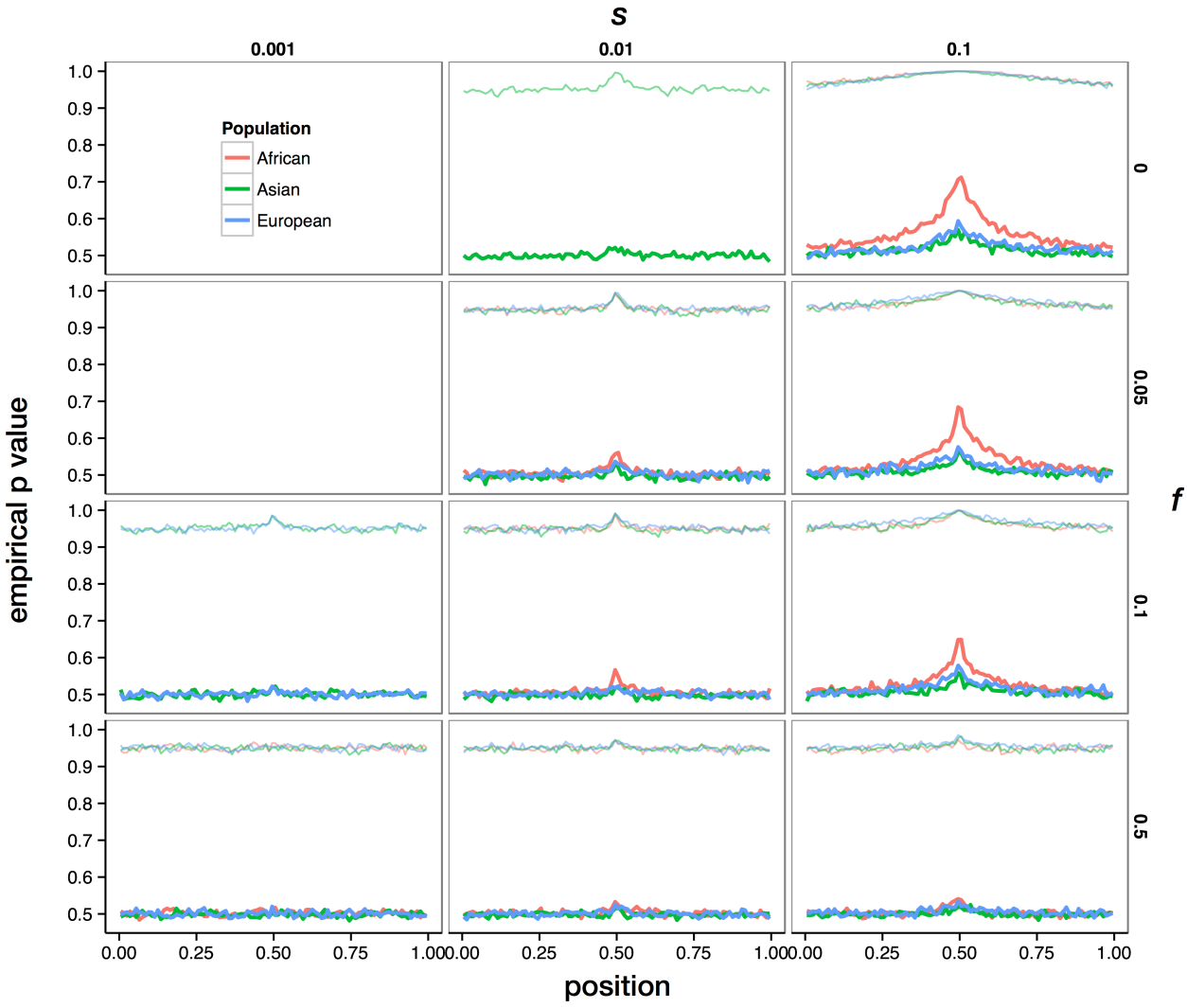


Figure C.3 Binned mean empirical p value for ΔDAF depends on *s* and *f* in simulations including selection

For each combination of *s* and *f*, the average empirical p value binned by position is plotted in thick, darker lines. Thin, lighter lines indicate the binned 95% confidence interval. Simulations including selection in each of three populations are shown in each panel where available. Populations are colored according to the legend at top left.

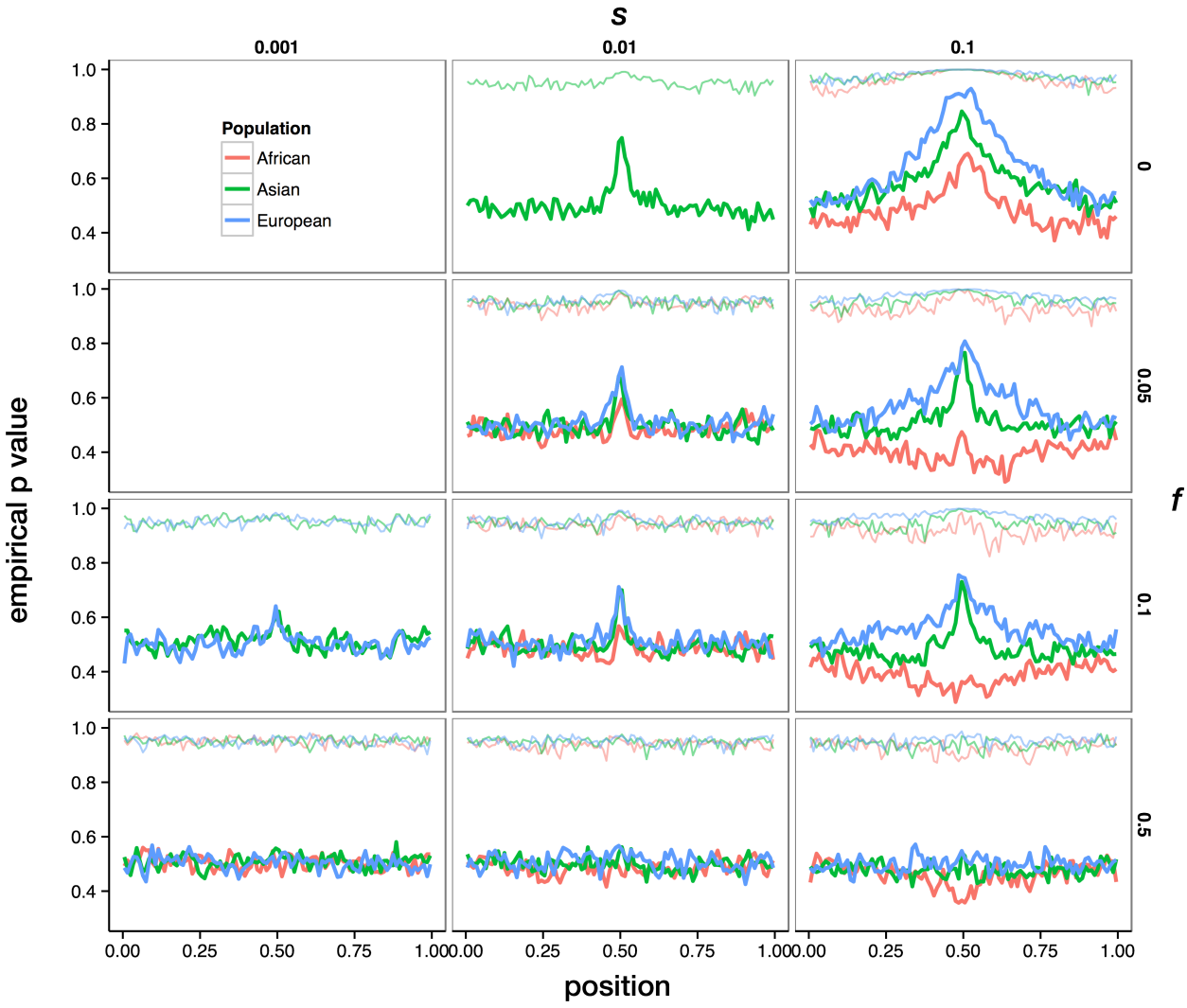


Figure C.4 Binned mean empirical p value for F* depends on *s* and *f* in simulations including selection

For each combination of *s* and *f*, the average empirical p value binned by position is plotted in thick, darker lines. Thin, lighter lines indicate the binned 95% confidence interval. Simulations including selection in each of three populations are shown in each panel where available. Populations are colored according to the legend at top left.

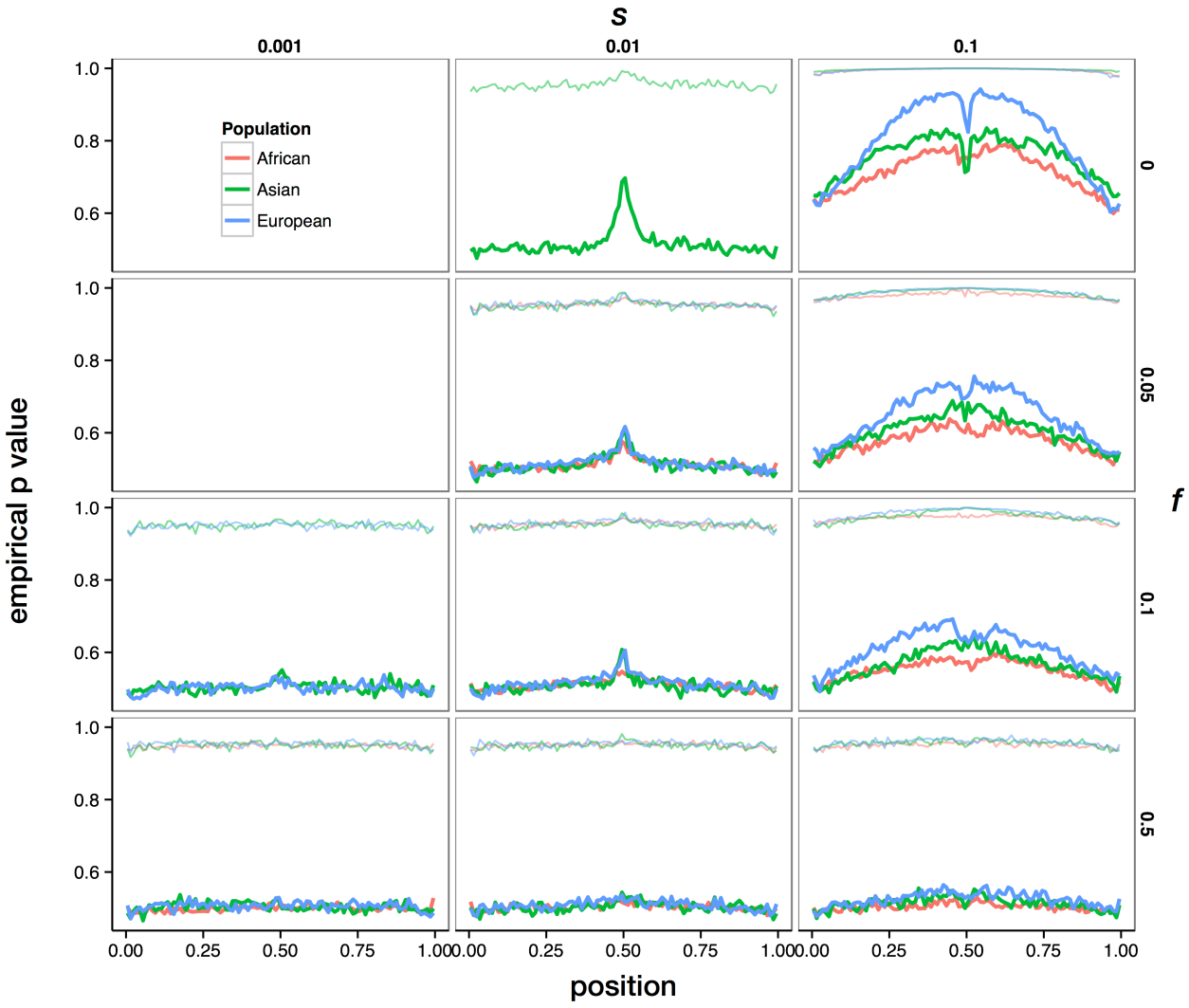


Figure C.5 Binned mean empirical p value for iHS depends on *s* and *f* in simulations including selection

For each combination of *s* and *f*, the average empirical p value binned by position is plotted in thick, darker lines. Thin, lighter lines indicate the binned 95% confidence interval. Simulations including selection in each of three populations are shown in each panel where available. Populations are colored according to the legend at top left.

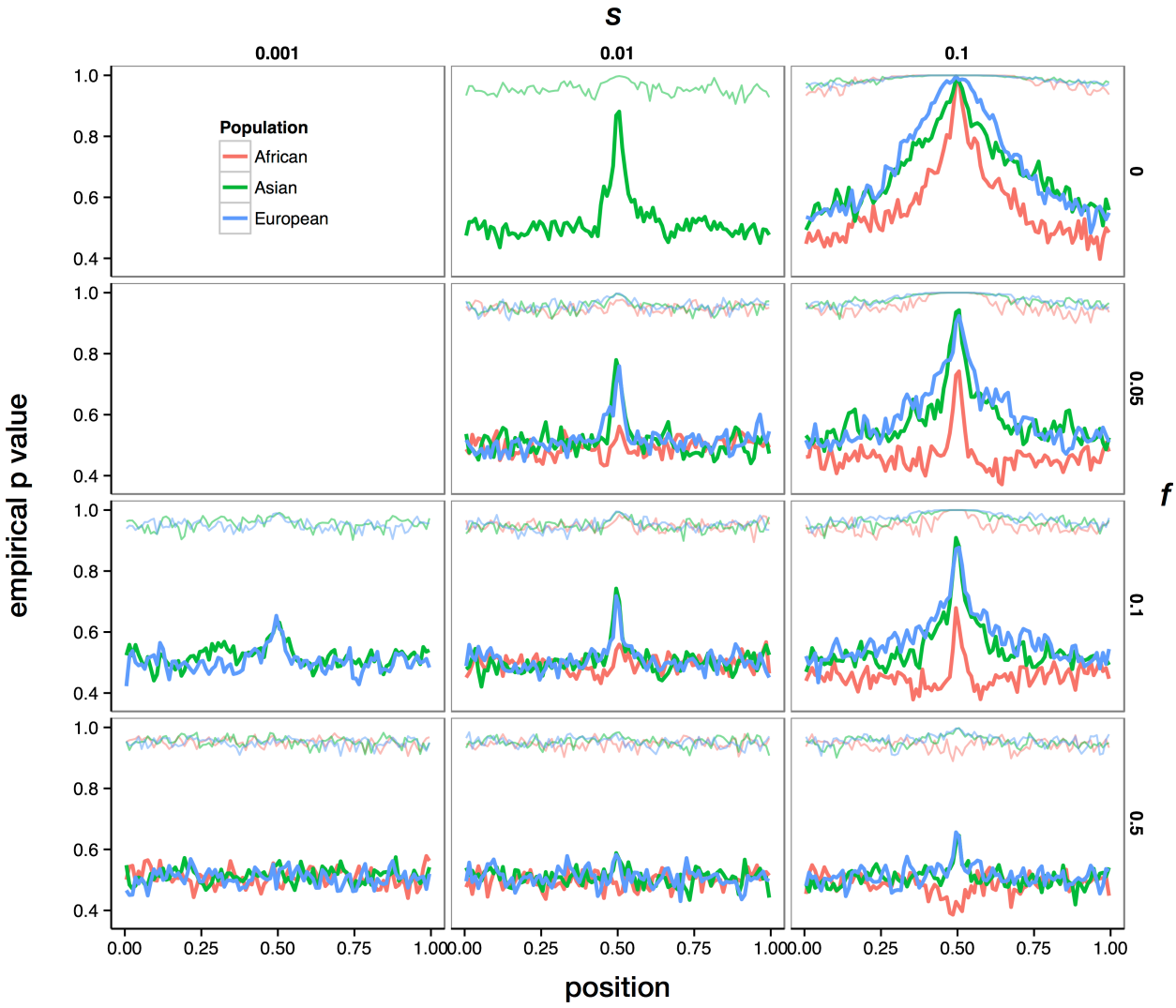


Figure C.6 Binned mean empirical p value for D_{Tajima} depends on *s* and *f* in simulations including selection

For each combination of *s* and *f*, the average empirical p value binned by position is plotted in thick, darker lines. Thin, lighter lines indicate the binned 95% confidence interval. Simulations including selection in each of three populations are shown in each panel where available. Populations are colored according to the legend at top left.

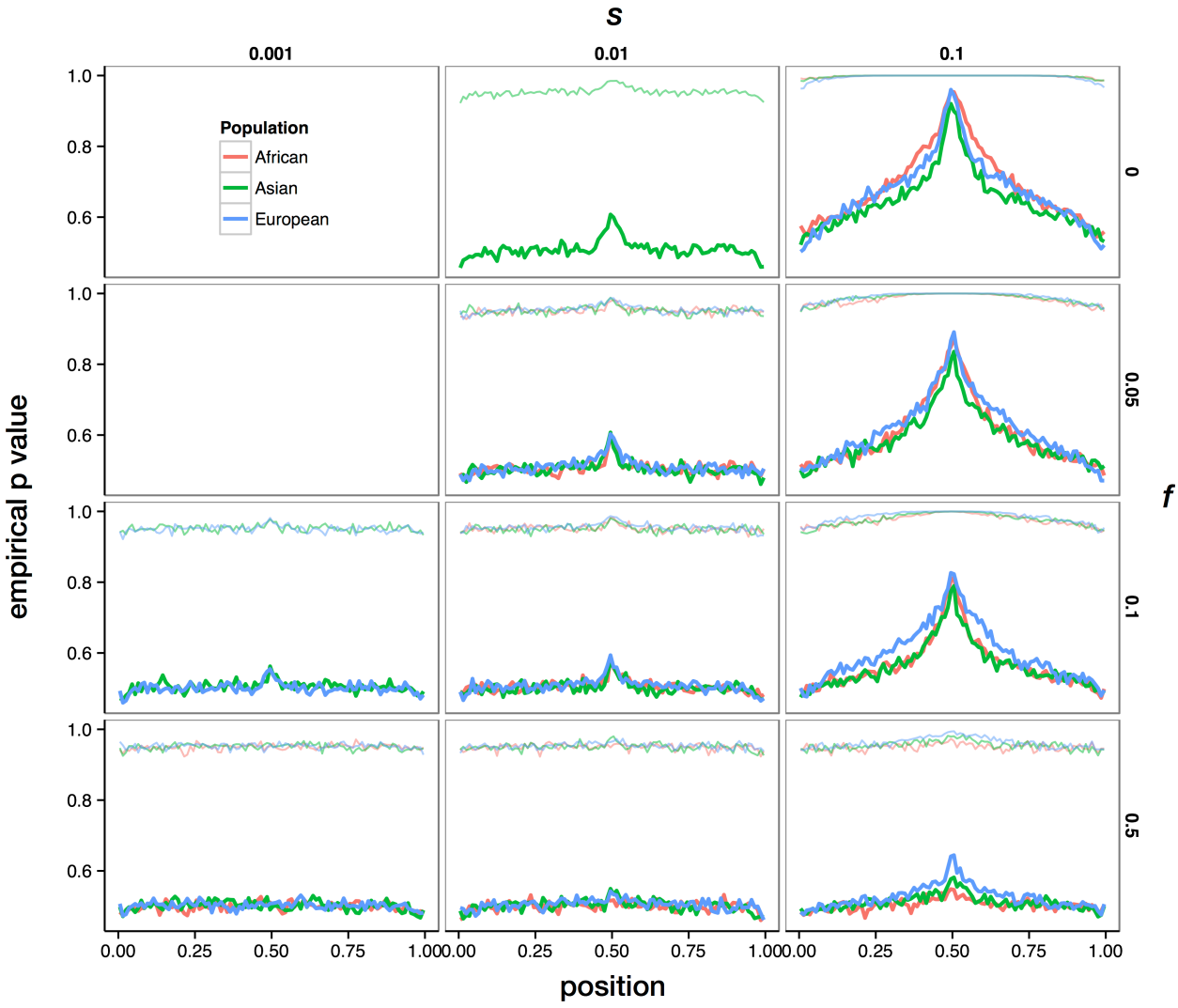


Figure C.7 Binned mean empirical p value for xp-iEHH$_D$ depends on *s* and *f* in simulations including selection

For each combination of *s* and *f*, the average empirical p value binned by position is plotted in thick, darker lines. Thin, lighter lines indicate the binned 95% confidence interval. Simulations including selection in each of three populations are shown in each panel where available. Populations are colored according to the legend at top left.

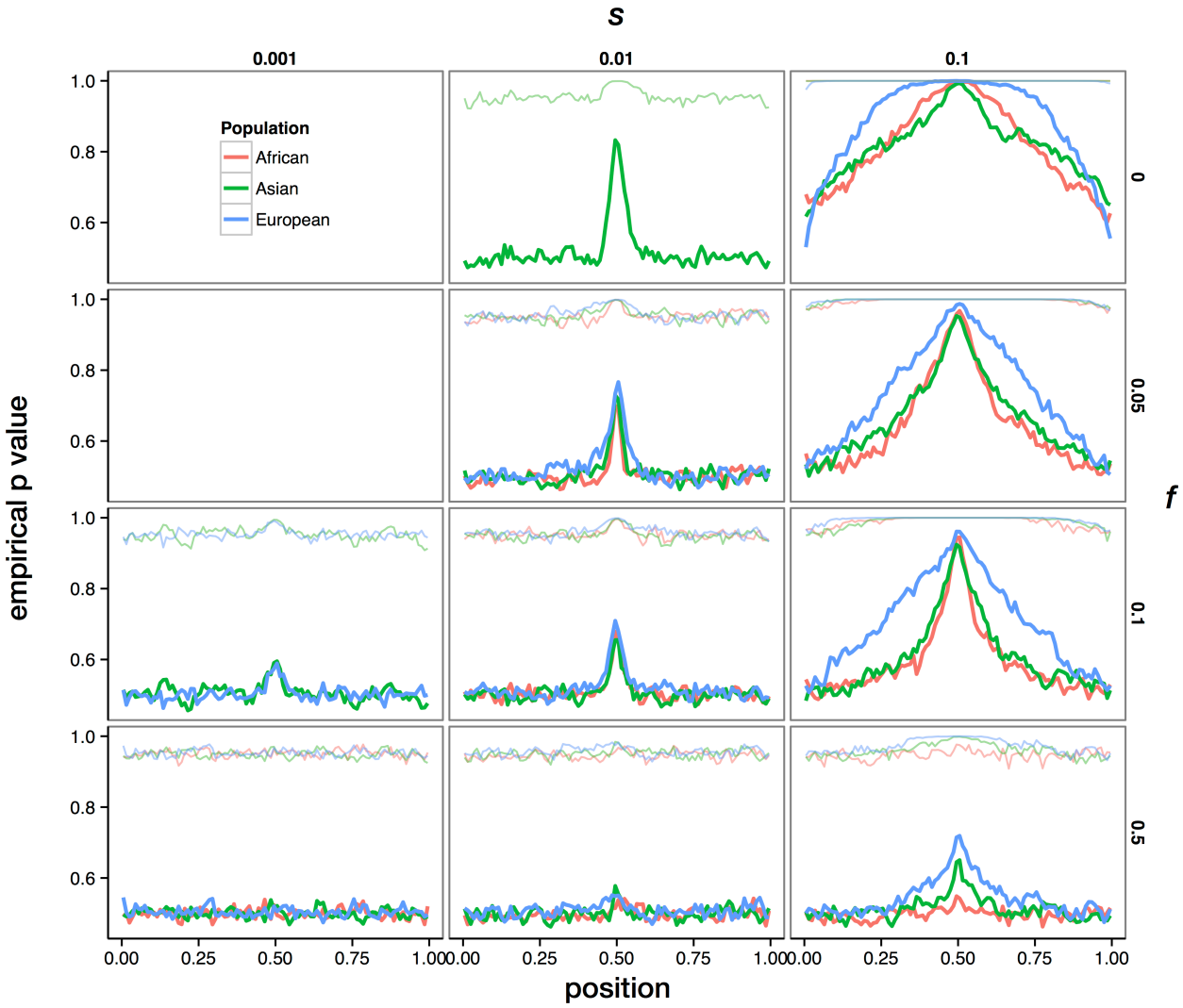


Figure C.8 Binned mean empirical p value for xp-iEHH$_S$ depends on *s* and *f* in simulations including selection

For each combination of *s* and *f*, the average empirical p value binned by position is plotted in thick, darker lines. Thin, lighter lines indicate the binned 95% confidence interval. Simulations including selection in each of three populations are shown in each panel where available. Populations are colored according to the legend at top left.

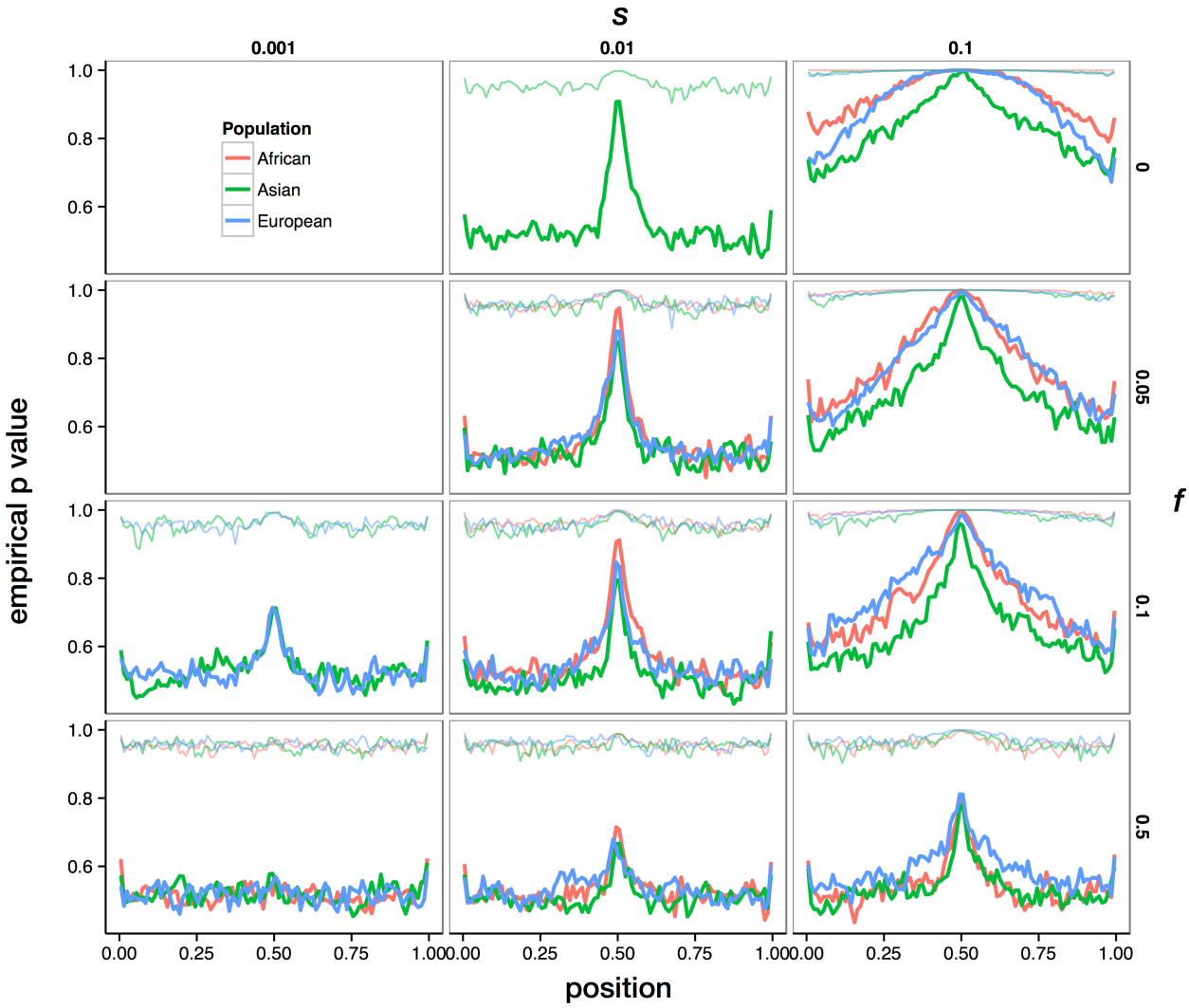


Figure C.9 Binned mean empirical p value for H_{12} depends on *s* and *f* in simulations including selection

For each combination of *s* and *f*, the average empirical p value binned by position is plotted in thick, darker lines. Thin, lighter lines indicate the binned 95% confidence interval. Simulations including selection in each of three populations are shown in each panel where available. Populations are colored according to the legend at top left.

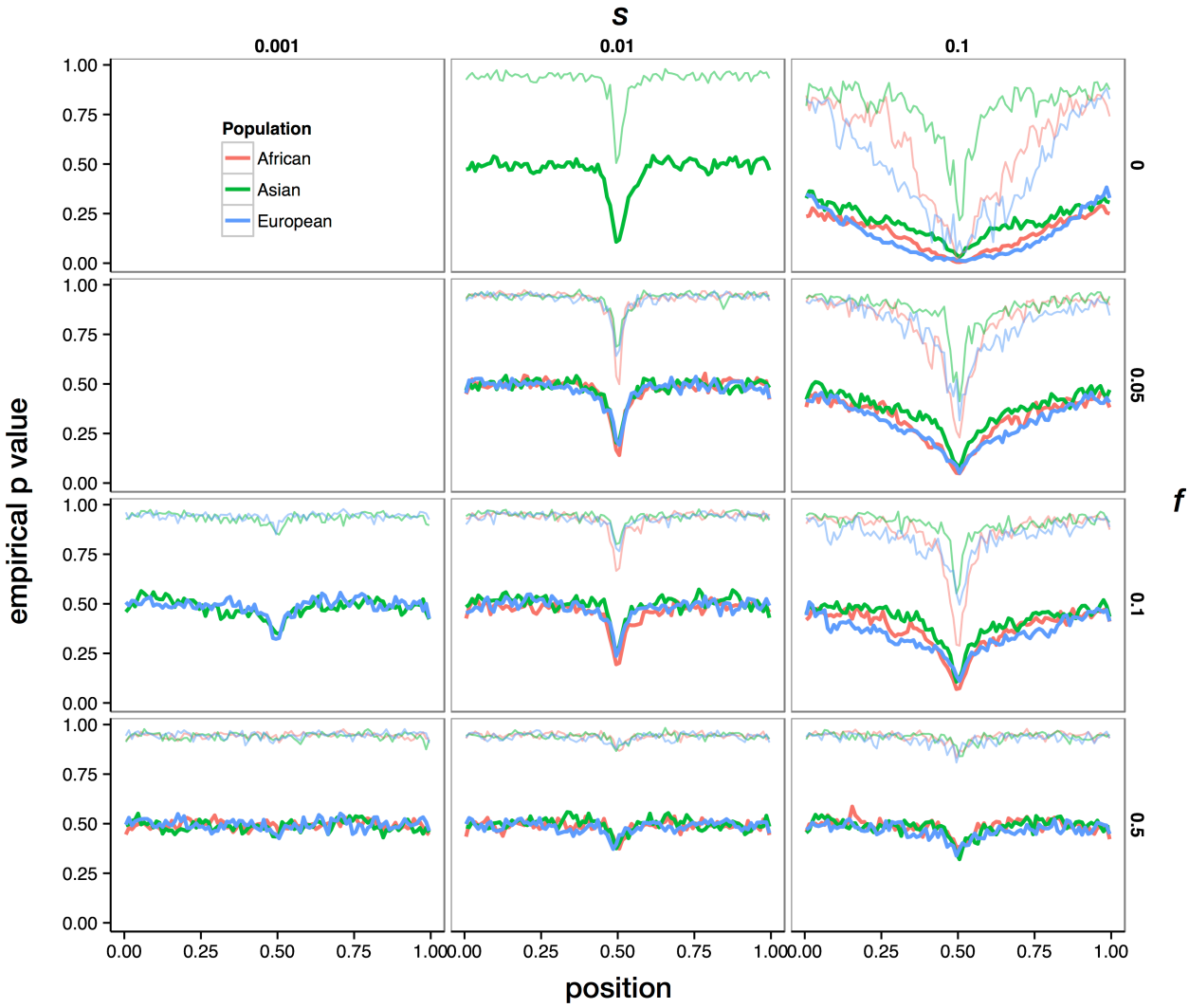


Figure C.10 Binned mean empirical p value for H_2/H_1 depends on *s* and *f* in simulations including selection

For each combination of *s* and *f*, the average empirical p value binned by position is plotted in thick, darker lines. Thin, lighter lines indicate the binned 95% confidence interval. Simulations including selection in each of three populations are shown in each panel where available. Populations are colored according to the legend at top left.

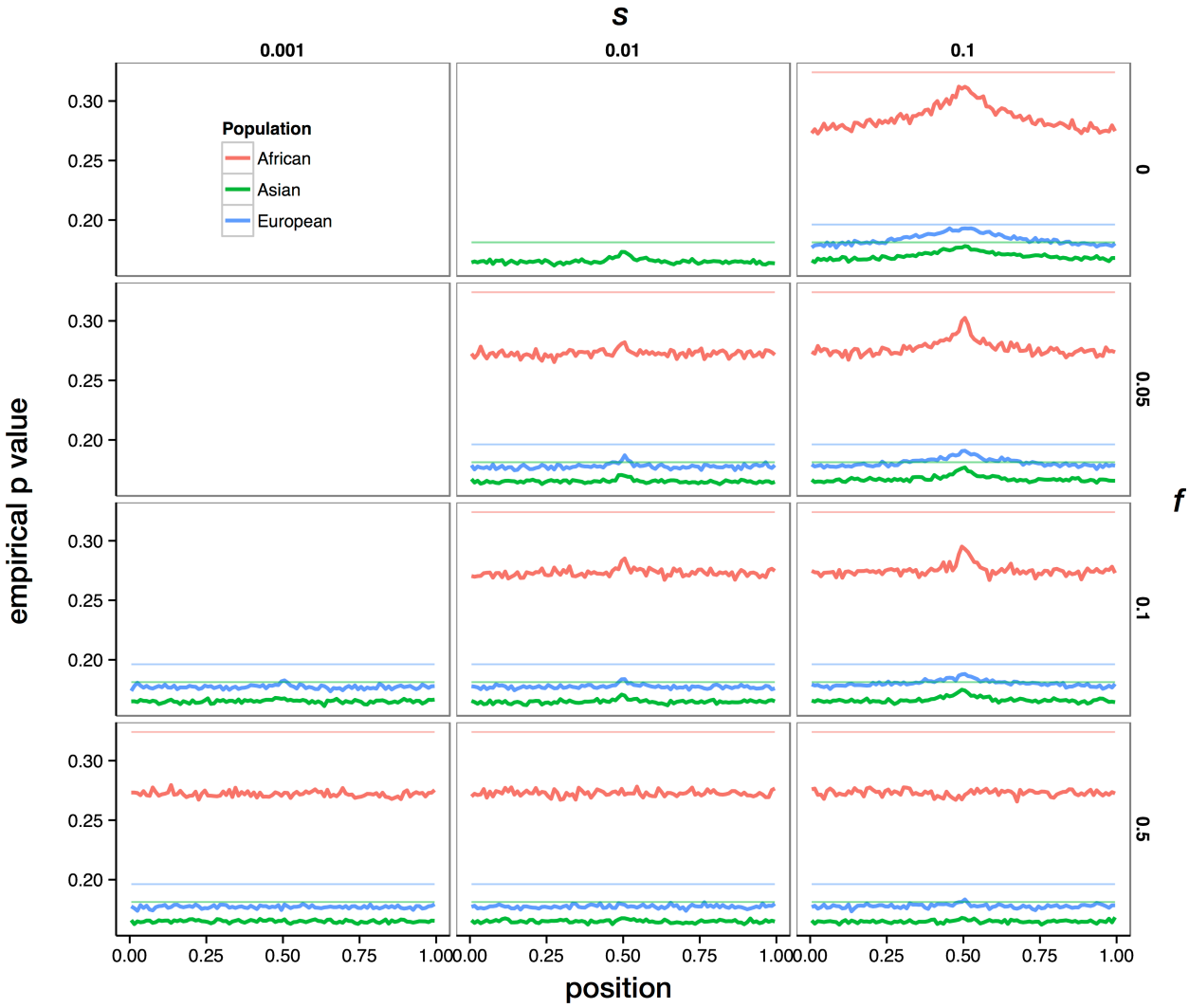


Figure C.11 Binned mean empirical p value for d_{12} varies little with s and f in simulations including selection

For each combination of s and f, the average empirical p value binned by position is plotted in thick, darker lines. Thin, lighter lines indicate the binned 95% confidence interval. Simulations including selection in each of three populations are shown in each panel where available. Populations are colored according to the legend at top left.

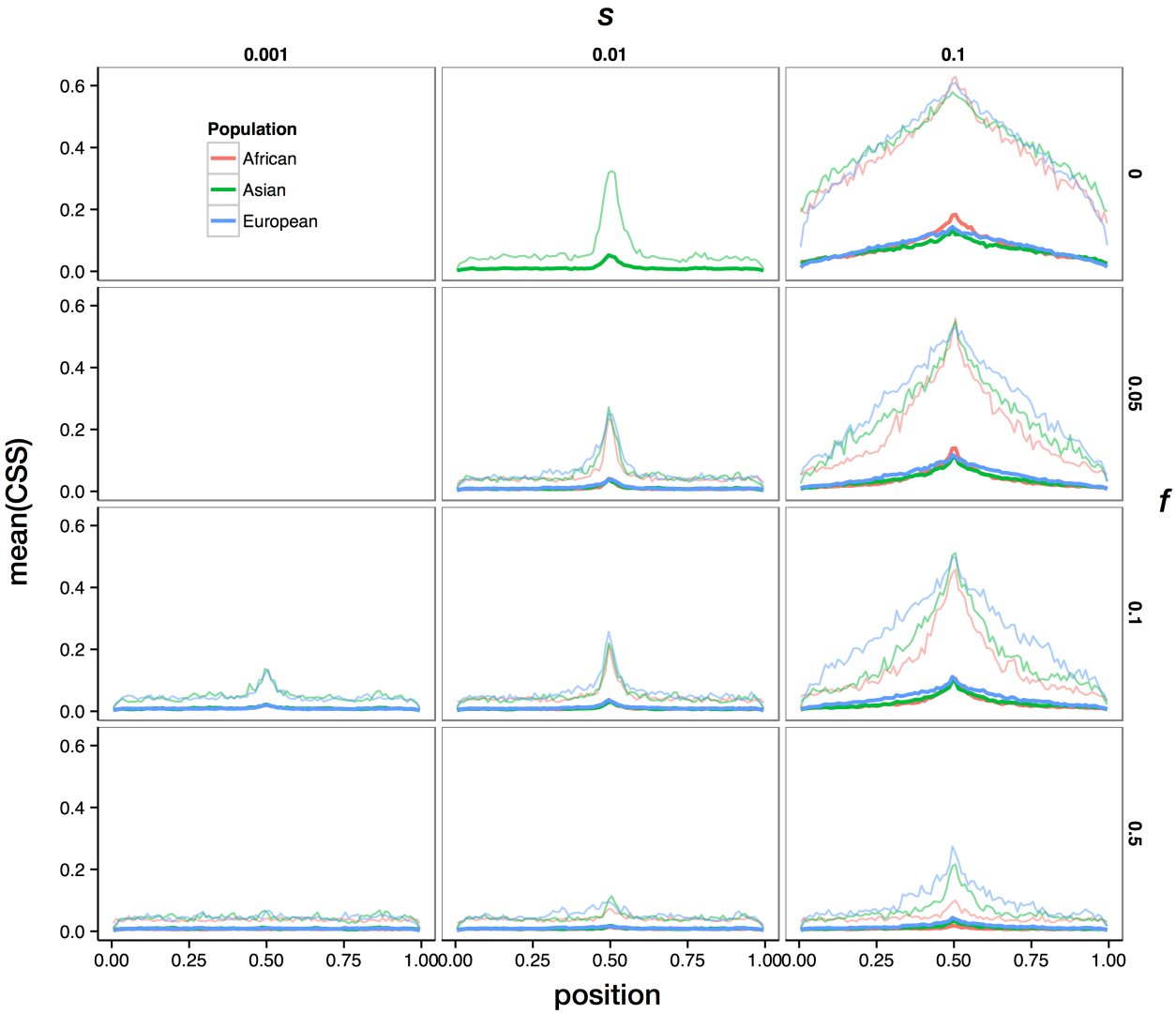


Figure C.12 Binned mean CSS_1 displays an isolated peak at the selected site in simulations including selection

For each combination of *s* and *f*, the average CSS_1 binned by position is plotted in thick, darker lines. Thin, lighter lines indicate the binned 95% confidence interval. Simulations including selection in each of three populations are shown in each panel where available. Populations are colored according to the legend at top left.

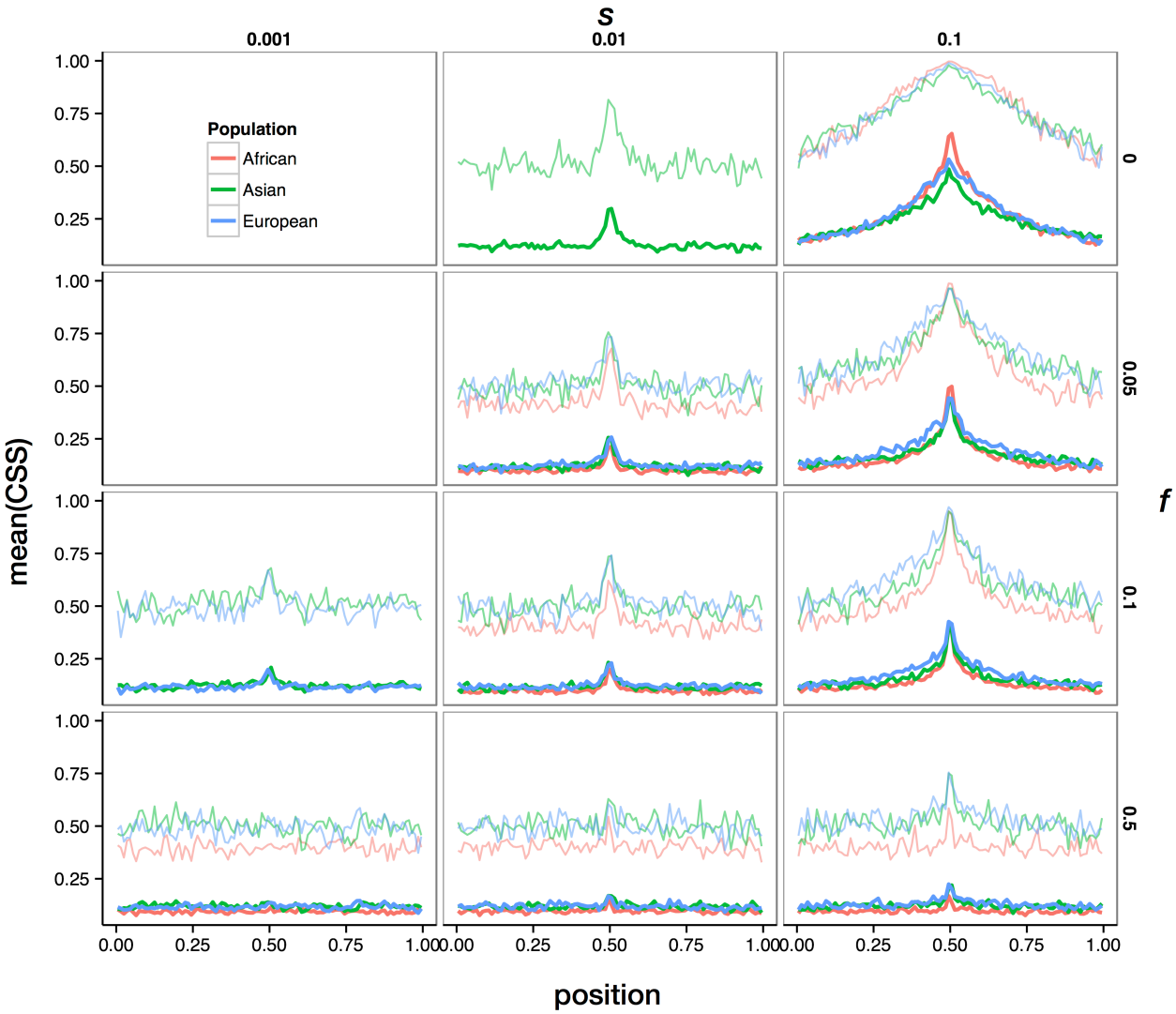


Figure C.13 Binned mean CSS_5 displays a marked peak at the selected site in simulations including selection

For each combination of *s* and *f*, the average CSS_5 binned by position is plotted in thick, darker lines. Thin, lighter lines indicate the binned 95% confidence interval. Simulations including selection in each of three populations are shown in each panel where available. Populations are colored according to the legend at top left.

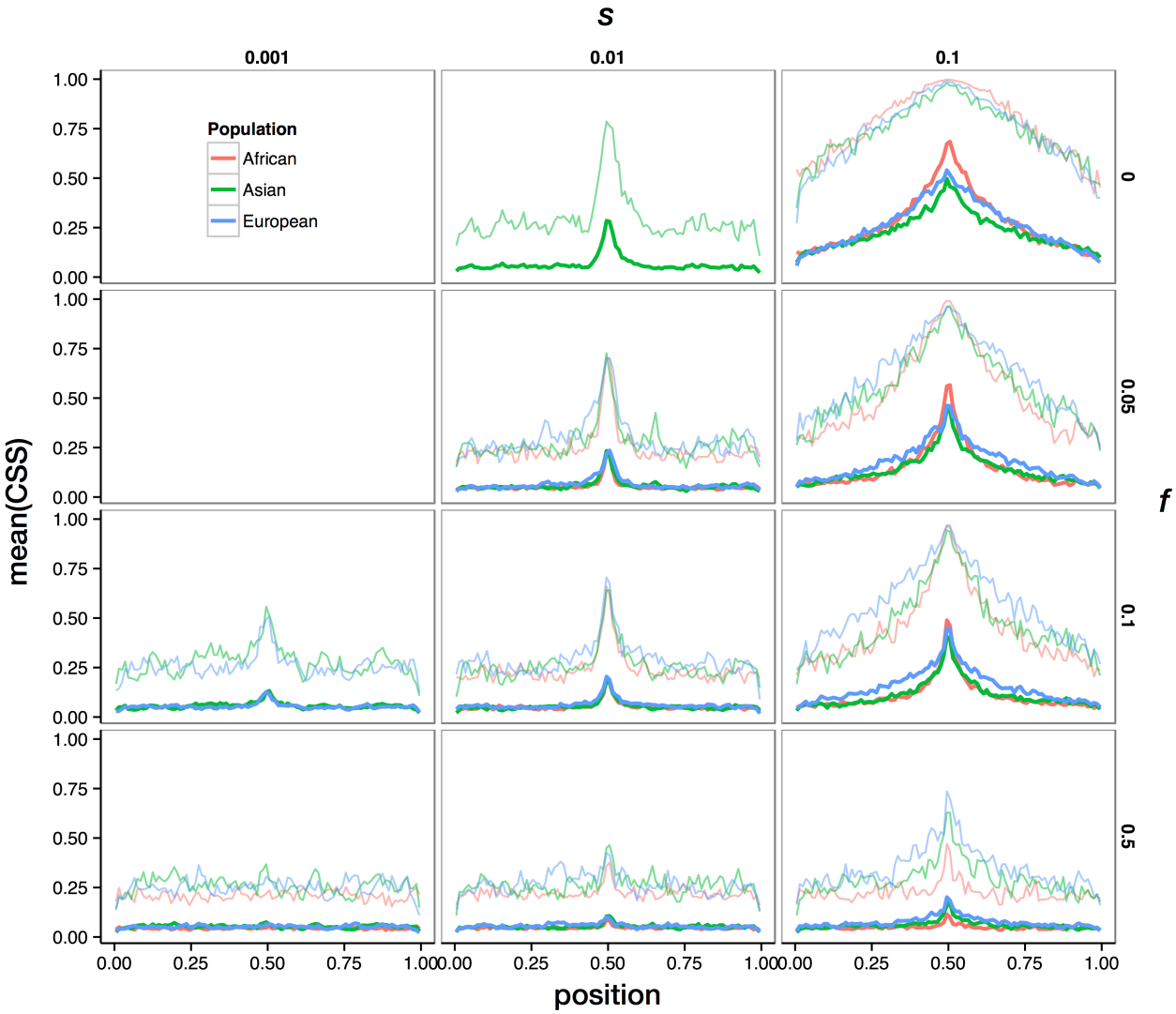


Figure C.14 Binned mean CSS_6 displays a marked peak at the selected site in simulations including selection

For each combination of *s* and *f*, the average CSS_6 binned by position is plotted in thick, darker lines. Thin, lighter lines indicate the binned 95% confidence interval. Simulations including selection in each of three populations are shown in each panel where available. Populations are colored according to the legend at top left.

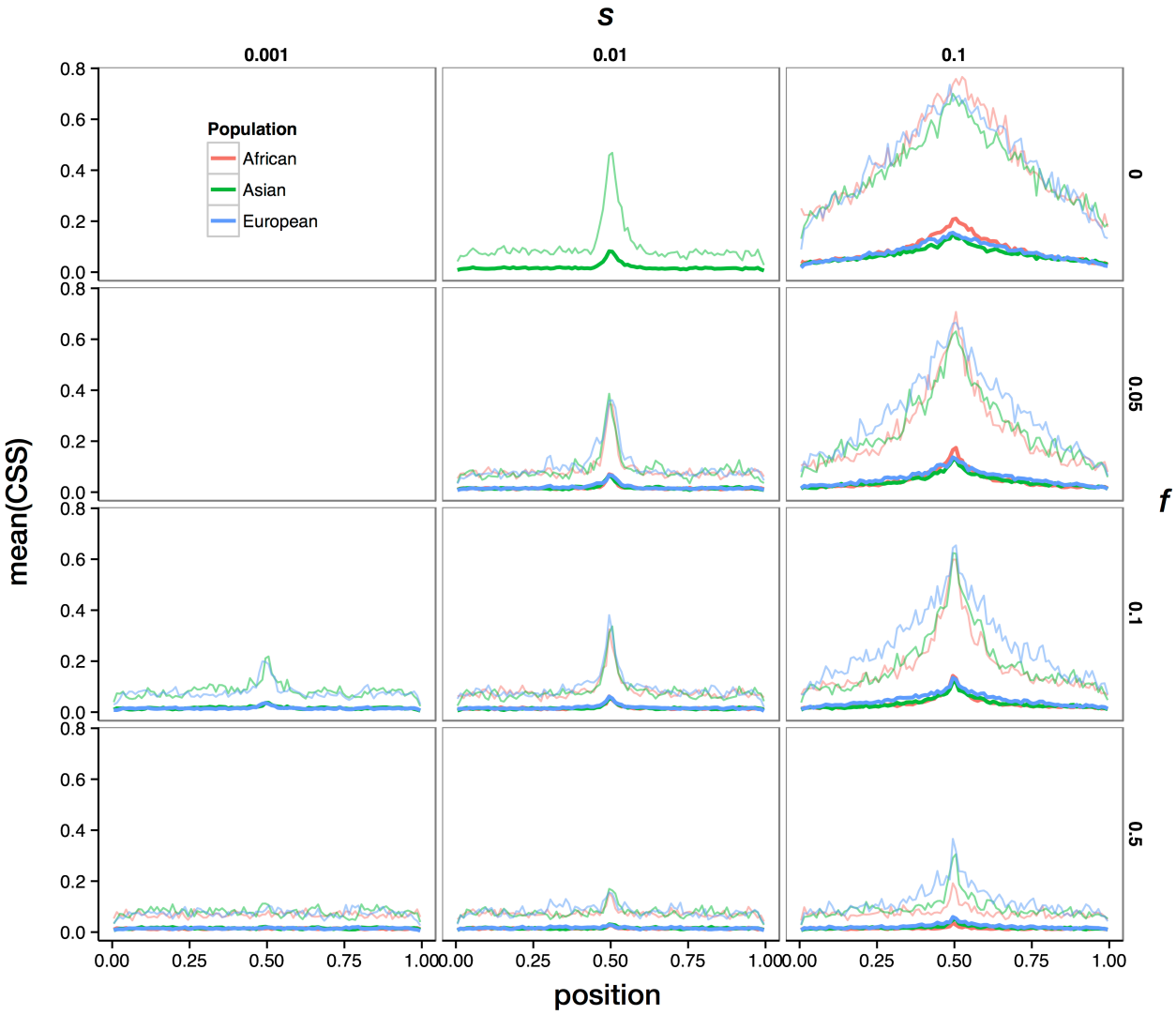


Figure C.15 Binned mean CSS_3 displays an isolated peak at the selected site in simulations including selection

For each combination of *s* and *f*, the average CSS_3 binned by position is plotted in thick, darker lines. Thin, lighter lines indicate the binned 95% confidence interval. Simulations including selection in each of three populations are shown in each panel where available. Populations are colored according to the legend at top left.

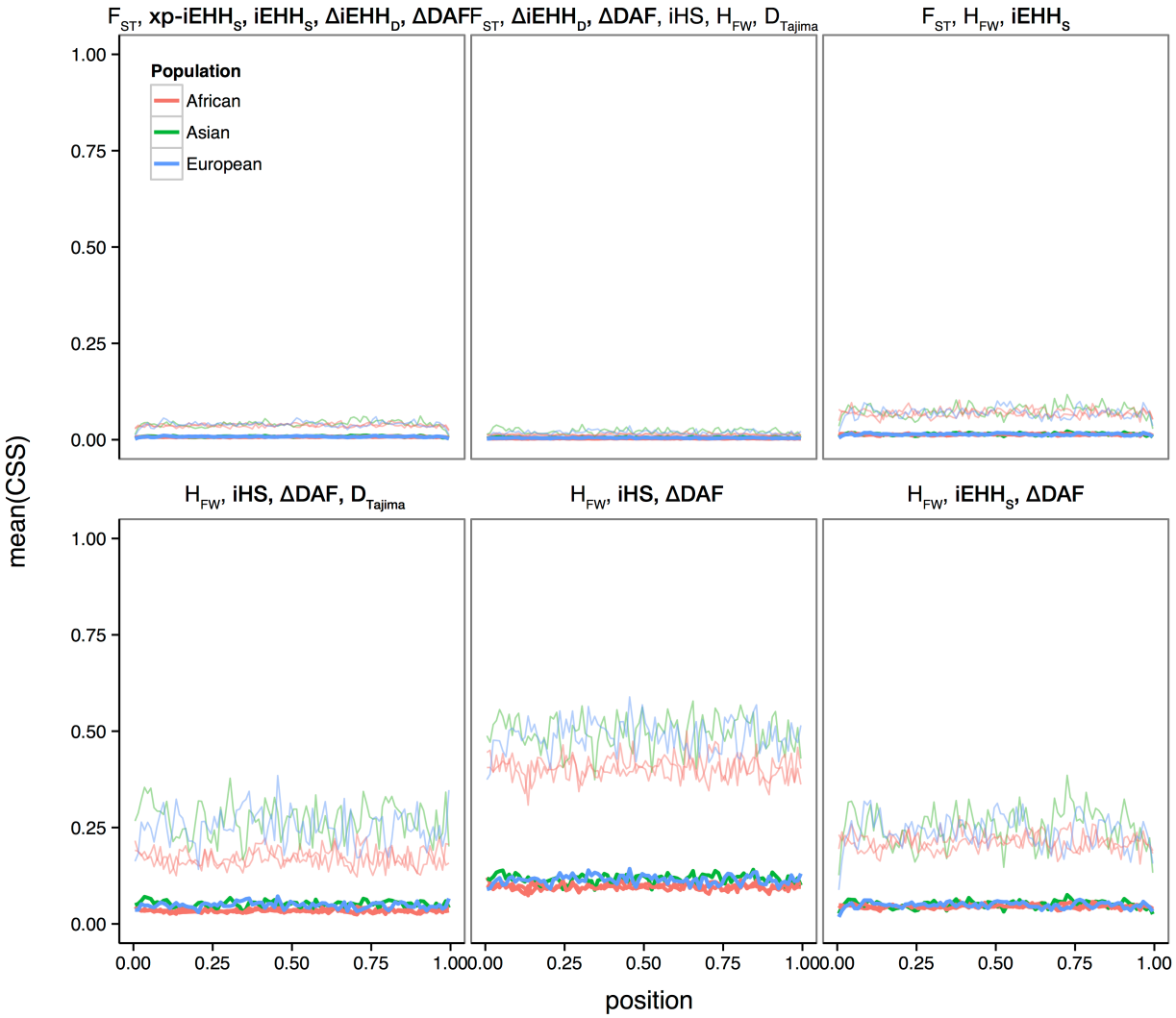


Figure C.16 Binned mean CSS shows no notable peaks in neutral simulations
For each combination of *s* and *f*, the average CSS binned by position is plotted in thick, darker lines. Thin, lighter lines indicate the binned 95% confidence interval. Simulations including selection in each of three populations are shown in each panel where available. Populations are colored according to the legend at top left. From the right to left and bottom to top, panels show CSS_1, CSS_2, CSS_3, CSS_4, CSS_5, and CSS_6.

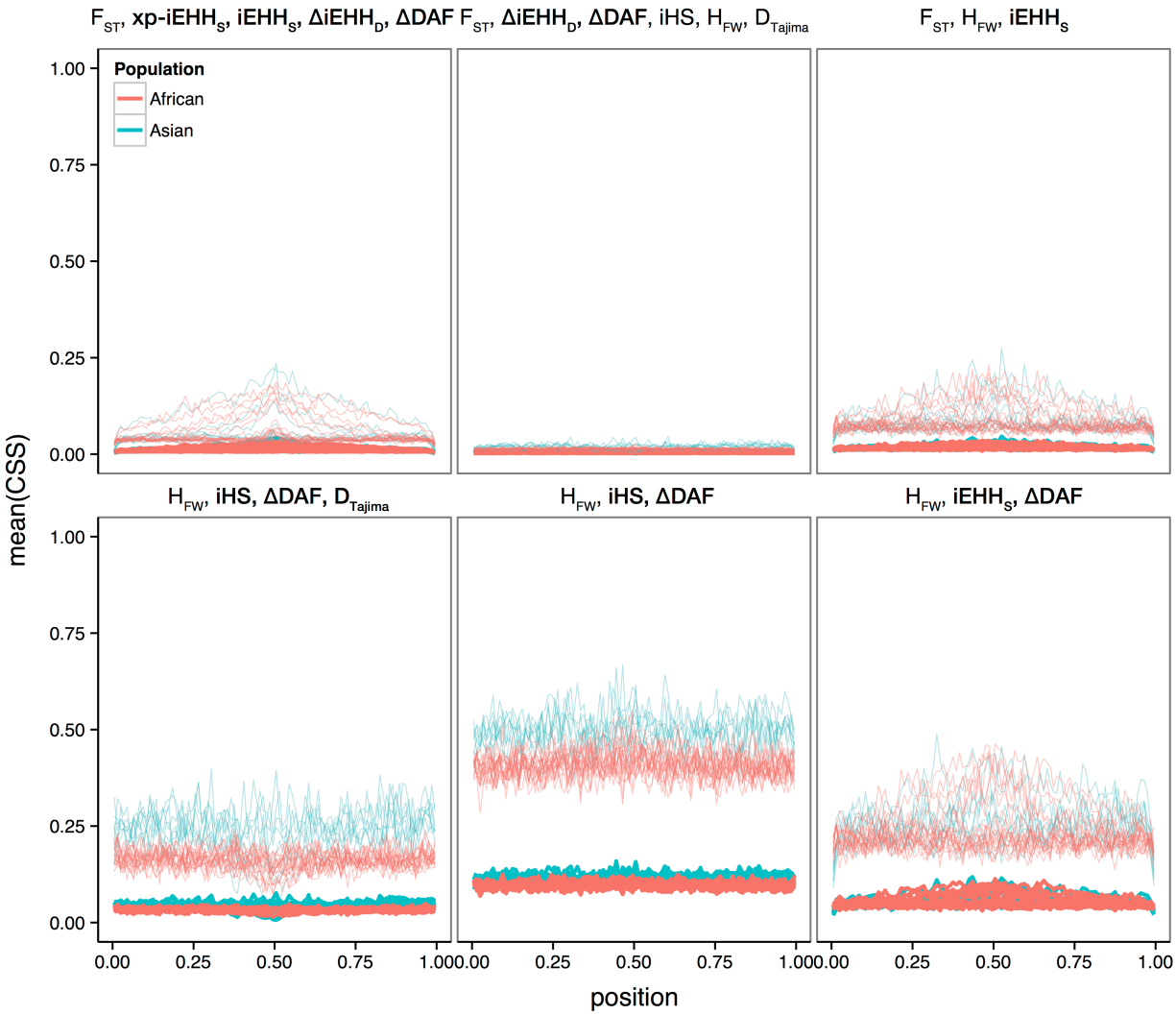


Figure C.17 Binned mean CSS shows no notable peaks in non-selected populations from simulations with selection.

For each combination of *s* and *f*, the average CSS binned by position is plotted in thick, darker lines. Thin, lighter lines indicate the binned 95% confidence interval. Simulations including selection in each of three populations are shown in each panel where available. Populations are colored according to the legend at top left. From the right to left and bottom to top, panels show CSS_1, CSS_2, CSS_3, CSS_4, CSS_5, and CSS_6.

C.2 Supplementary Tables

Table C.1 `ms` **and** `msms` **command lines for coalescent simulations**

Population experiencing selection	*f*	*s*	Simulation command line
African	0.05	0.01	java -Xmx3G -Xms3G -jar msms.jar 200 50 -t 6000.000000 -r 8959.991040 1000000 -I 2 100 100 -ej 0.00875 2 1 -en 0.0005 1 0.24 -en 0.001 2 0.077 -en 0.00475 2 0.00373 -en 0.004875 2 0.077 -en 0.0075 1 0.03125 -en 0.007625 1 0.24 -en 0.0085 2 0.00294 -en 0.008625 2 0.077 -en 0.0425 1 0.12 -Smark -N 100000 -Sp 0.500000 -oOC -Smu 0.006000 -SI 0.008400 2 0.050000 0.050000 -Sc 0 1 2000.000000 1000.000000 0.000000 -SFC -oFP 0.0000000
African	0.05	0.1	java -Xmx3G -Xms3G -jar msms.jar 200 50 -t 6000.000000 -r 8959.991040 1000000 -I 2 100 100 -ej 0.00875 2 1 -en 0.0005 1 0.24 -en 0.001 2 0.077 -en 0.00475 2 0.00373 -en 0.004875 2 0.077 -en 0.0075 1 0.03125 -en 0.007625 1 0.24 -en 0.0085 2 0.00294 -en 0.008625 2 0.077 -en 0.0425 1 0.12 -Smark -N 100000 -Sp 0.500000 -oOC -Smu 0.006000 -SI 0.008400 2 0.050000 0.050000 -Sc 0 1 20000.000000 10000.000000 0.000000 -SFC -oFP 0.0000000
African	0.0	0.1	java -Xmx3G -Xms3G -jar msms.jar 200 50 -t 6000.000000 -r 8959.991040 1000000 -I 2 100 100 -ej 0.00875 2 1 -en 0.0005 1 0.24 -en 0.001 2 0.077 -en 0.00475 2 0.00373 -en 0.004875 2 0.077 -en 0.0075 1 0.03125 -en 0.007625 1 0.24 -en 0.0085 2 0.00294 -en 0.008625 2 0.077 -en 0.0425 1 0.12 -Smark -N 100000 -Sp 0.500000 -oOC -Smu 0.006000 -SI 0.008400 2 0.000000 0.000000 -Sc 0 1 20000.000000 10000.000000 0.000000 -SFC -oFP 0.0000000
African	0.1	0.01	java -Xmx3G -Xms3G -jar msms.jar 200 50 -t 6000.000000 -r 8959.991040 1000000 -I 2 100 100 -ej 0.00875 2 1 -en 0.0005 1 0.24 -en 0.001 2 0.077 -en 0.00475 2 0.00373 -en 0.004875 2 0.077 -en 0.0075 1 0.03125 -en 0.007625 1 0.24 -en 0.0085 2 0.00294 -en 0.008625 2 0.077 -en 0.0425 1 0.12 -Smark -N 100000 -Sp 0.500000 -oOC -Smu 0.006000 -SI 0.008400 2 0.100000 0.100000 -Sc 0 1 2000.000000 1000.000000 0.000000 -SFC -oFP 0.0000000
African	0.1	0.1	java -Xmx3G -Xms3G -jar msms.jar 200 50 -t 6000.000000 -r 8959.991040 1000000 -I 2 100 100 -ej 0.00875 2 1 -en 0.0005 1 0.24 -en 0.001 2 0.077 -en 0.00475 2 0.00373 -en 0.004875 2 0.077 -en 0.0075 1 0.03125 -en 0.007625 1 0.24 -en 0.0085 2 0.00294 -en 0.008625 2 0.077 -en 0.0425 1 0.12 -Smark -N 100000 -Sp 0.500000 -oOC -Smu 0.006000 -SI 0.008400 2 0.100000 0.100000 -Sc 0 1 20000.000000 10000.000000 0.000000 -SFC -oFP 0.0000000
African	0.5	0.001	java -Xmx3G -Xms3G -jar msms.jar 200 50 -t 6000.000000 -r 8959.991040 1000000 -I 2 100 100 -ej 0.00875 2 1 -en 0.0005 1 0.24 -en 0.001 2 0.077 -en 0.00475 2 0.00373 -en 0.004875 2 0.077 -en 0.0075 1 0.03125 -en 0.007625 1 0.24 -en 0.0085 2 0.00294 -en 0.008625 2 0.077 -en 0.0425 1 0.12 -Smark -N 100000 -Sp 0.500000 -oOC -Smu 0.006000 -SI 0.008400 2 0.500000 0.500000 -Sc 0 1 200.000000 100.000000 0.000000 -SFC -oFP 0.0000000
African	0.5	0.01	java -Xmx3G -Xms3G -jar msms.jar 200 50 -t 6000.000000 -r

			8959.991040 1000000 -l 2 100 100 -ej 0.00875 2 1 -en 0.0005 1 0.24 -en 0.001 2 0.077 -en 0.00475 2 0.00373 -en 0.004875 2 0.077 -en 0.0075 1 0.03125 -en 0.007625 1 0.24 -en 0.0085 2 0.00294 -en 0.008625 2 0.077 -en 0.0425 1 0.12 -Smark -N 100000 -Sp 0.500000 -oOC -Smu 0.006000 -Sl 0.008400 2 0.500000 0.500000 -Sc 0 1 2000.000000 1000.000000 0.000000 -SFC -oFP 0.0000000
African	0.5	0.1	java -Xmx3G -Xms3G -jar msms.jar 200 50 -t 6000.000000 -r 8959.991040 1000000 -l 2 100 100 -ej 0.00875 2 1 -en 0.0005 1 0.24 -en 0.001 2 0.077 -en 0.00475 2 0.00373 -en 0.004875 2 0.077 -en 0.0075 1 0.03125 -en 0.007625 1 0.24 -en 0.0085 2 0.00294 -en 0.008625 2 0.077 -en 0.0425 1 0.12 -Smark -N 100000 -Sp 0.500000 -oOC -Smu 0.006000 -Sl 0.008400 2 0.500000 0.500000 -Sc 0 1 20000.000000 10000.000000 0.000000 -SFC -oFP 0.0000000
Asian	0.05	0.01	java -Xmx3G -Xms3G -jar msms.jar 200 50 -t 6000.000000 -r 8959.991040 1000000 -l 2 100 100 -ej 0.00875 2 1 -en 0.0005 1 0.24 -en 0.001 2 0.077 -en 0.00475 2 0.00373 -en 0.004875 2 0.077 -en 0.0075 1 0.03125 -en 0.007625 1 0.24 -en 0.0085 2 0.00294 -en 0.008625 2 0.077 -en 0.0425 1 0.12 -Smark -N 100000 -Sp 0.500000 -oOC -Smu 0.006000 -Sl 0.008400 2 0.050000 0.050000 -Sc 0 2 2000.000000 1000.000000 0.000000 -SFC -oFP 0.0000000
Asian	0.05	0.1	java -Xmx3G -Xms3G -jar msms.jar 200 50 -t 6000.000000 -r 8959.991040 1000000 -l 2 100 100 -ej 0.00875 2 1 -en 0.0005 1 0.24 -en 0.001 2 0.077 -en 0.00475 2 0.00373 -en 0.004875 2 0.077 -en 0.0075 1 0.03125 -en 0.007625 1 0.24 -en 0.0085 2 0.00294 -en 0.008625 2 0.077 -en 0.0425 1 0.12 -Smark -N 100000 -Sp 0.500000 -oOC -Smu 0.006000 -Sl 0.008400 2 0.050000 0.050000 -Sc 0 2 20000.000000 10000.000000 0.000000 -SFC -oFP 0.0000000
Asian	0.0	0.01	java -Xmx3G -Xms3G -jar msms.jar 200 50 -t 6000.000000 -r 8959.991040 1000000 -l 2 100 100 -ej 0.00875 2 1 -en 0.0005 1 0.24 -en 0.001 2 0.077 -en 0.00475 2 0.00373 -en 0.004875 2 0.077 -en 0.0075 1 0.03125 -en 0.007625 1 0.24 -en 0.0085 2 0.00294 -en 0.008625 2 0.077 -en 0.0425 1 0.12 -Smark -N 100000 -Sp 0.500000 -oOC -Smu 0.006000 -Sl 0.008400 2 0.000000 0.000000 -Sc 0 2 2000.000000 1000.000000 0.000000 -SFC -oFP 0.0000000
Asian	0.0	0.1	java -Xmx3G -Xms3G -jar msms.jar 200 50 -t 6000.000000 -r 8959.991040 1000000 -l 2 100 100 -ej 0.00875 2 1 -en 0.0005 1 0.24 -en 0.001 2 0.077 -en 0.00475 2 0.00373 -en 0.004875 2 0.077 -en 0.0075 1 0.03125 -en 0.007625 1 0.24 -en 0.0085 2 0.00294 -en 0.008625 2 0.077 -en 0.0425 1 0.12 -Smark -N 100000 -Sp 0.500000 -oOC -Smu 0.006000 -Sl 0.008400 2 0.000000 0.000000 -Sc 0 2 20000.000000 10000.000000 0.000000 -SFC -oFP 0.0000000
Asian	0.1	0.001	java -Xmx3G -Xms3G -jar msms.jar 200 50 -t 6000.000000 -r 8959.991040 1000000 -l 2 100 100 -ej 0.00875 2 1 -en 0.0005 1 0.24 -en 0.001 2 0.077 -en 0.00475 2 0.00373 -en 0.004875 2 0.077 -en 0.0075 1 0.03125 -en 0.007625 1 0.24 -en 0.0085 2 0.00294 -en 0.008625 2 0.077 -en 0.0425 1 0.12 -Smark -N 100000 -Sp 0.500000 -oOC -Smu 0.006000 -Sl 0.008400 2 0.100000 0.100000 -Sc 0 2 200.000000 100.000000 0.000000 -SFC -oFP 0.0000000
Asian	0.1	0.01	java -Xmx3G -Xms3G -jar msms.jar 200 50 -t 6000.000000 -r 8959.991040 1000000 -l 2 100 100 -ej 0.00875 2 1 -en 0.0005 1 0.24 -en 0.001 2 0.077 -en 0.00475 2 0.00373 -en 0.004875 2 0.077 -en 0.0075 1 0.03125 -en 0.007625 1 0.24 -en 0.0085 2 0.00294 -en 0.008625 2 0.077 -en 0.0425 1 0.12 -Smark -N 100000 -Sp 0.500000 -

			oOC -Smu 0.006000 -SI 0.008400 2 0.100000 0.100000 -Sc 0 2 2000.000000 1000.000000 0.000000 -SFC -oFP 0.0000000
Asian	0.1	0.1	java -Xmx3G -Xms3G -jar msms.jar 200 50 -t 6000.000000 -r 8959.991040 1000000 -I 2 100 100 -ej 0.00875 2 1 -en 0.0005 1 0.24 -en 0.001 2 0.077 -en 0.00475 2 0.00373 -en 0.004875 2 0.077 -en 0.0075 1 0.03125 -en 0.007625 1 0.24 -en 0.0085 2 0.00294 -en 0.008625 2 0.077 -en 0.0425 1 0.12 -Smark -N 100000 -Sp 0.500000 -oOC -Smu 0.006000 -SI 0.008400 2 0.100000 0.100000 -Sc 0 2 20000.000000 10000.000000 0.000000 -SFC -oFP 0.0000000
Asian	0.5	0.001	java -Xmx3G -Xms3G -jar msms.jar 200 50 -t 6000.000000 -r 8959.991040 1000000 -I 2 100 100 -ej 0.00875 2 1 -en 0.0005 1 0.24 -en 0.001 2 0.077 -en 0.00475 2 0.00373 -en 0.004875 2 0.077 -en 0.0075 1 0.03125 -en 0.007625 1 0.24 -en 0.0085 2 0.00294 -en 0.008625 2 0.077 -en 0.0425 1 0.12 -Smark -N 100000 -Sp 0.500000 -oOC -Smu 0.006000 -SI 0.008400 2 0.500000 0.500000 -Sc 0 2 200.000000 100.000000 0.000000 -SFC -oFP 0.0000000
Asian	0.5	0.01	java -Xmx3G -Xms3G -jar msms.jar 200 50 -t 6000.000000 -r 8959.991040 1000000 -I 2 100 100 -ej 0.00875 2 1 -en 0.0005 1 0.24 -en 0.001 2 0.077 -en 0.00475 2 0.00373 -en 0.004875 2 0.077 -en 0.0075 1 0.03125 -en 0.007625 1 0.24 -en 0.0085 2 0.00294 -en 0.008625 2 0.077 -en 0.0425 1 0.12 -Smark -N 100000 -Sp 0.500000 -oOC -Smu 0.006000 -SI 0.008400 2 0.500000 0.500000 -Sc 0 2 2000.000000 1000.000000 0.000000 -SFC -oFP 0.0000000
Asian	0.5	0.1	java -Xmx3G -Xms3G -jar msms.jar 200 50 -t 6000.000000 -r 8959.991040 1000000 -I 2 100 100 -ej 0.00875 2 1 -en 0.0005 1 0.24 -en 0.001 2 0.077 -en 0.00475 2 0.00373 -en 0.004875 2 0.077 -en 0.0075 1 0.03125 -en 0.007625 1 0.24 -en 0.0085 2 0.00294 -en 0.008625 2 0.077 -en 0.0425 1 0.12 -Smark -N 100000 -Sp 0.500000 -oOC -Smu 0.006000 -SI 0.008400 2 0.500000 0.500000 -Sc 0 2 20000.000000 10000.000000 0.000000 -SFC -oFP 0.0000000
European	0.05	0.01	java -Xmx3G -Xms3G -jar msms.jar 200 50 -t 6000.000000 -r 8959.991040 1000000 -I 2 100 100 -ej 0.00875 2 1 -en 0.0005 1 0.24 -en 0.000875 2 0.077 -en 0.001 2 0.0125 -en 0.001125 2 0.077 -en 0.0075 1 0.03125 -en 0.007625 1 0.24 -en 0.0085 2 0.00294 -en 0.008625 2 0.077 -en 0.0425 1 0.12 -Smark -N 100000 -Sp 0.500000 -oOC -Smu 0.006000 -SI 0.008400 2 0.050000 0.050000 -Sc 0 2 2000.000000 1000.000000 0.000000 -SFC -oFP 0.0000000
European	0.05	0.1	java -Xmx3G -Xms3G -jar msms.jar 200 50 -t 6000.000000 -r 8959.991040 1000000 -I 2 100 100 -ej 0.00875 2 1 -en 0.0005 1 0.24 -en 0.000875 2 0.077 -en 0.001 2 0.0125 -en 0.001125 2 0.077 -en 0.0075 1 0.03125 -en 0.007625 1 0.24 -en 0.0085 2 0.00294 -en 0.008625 2 0.077 -en 0.0425 1 0.12 -Smark -N 100000 -Sp 0.500000 -oOC -Smu 0.006000 -SI 0.008400 2 0.050000 0.050000 -Sc 0 2 20000.000000 10000.000000 0.000000 -SFC -oFP 0.0000000
European	0.0	0.1	java -Xmx3G -Xms3G -jar msms.jar 200 50 -t 6000.000000 -r 8959.991040 1000000 -I 2 100 100 -ej 0.00875 2 1 -en 0.0005 1 0.24 -en 0.000875 2 0.077 -en 0.001 2 0.0125 -en 0.001125 2 0.077 -en 0.0075 1 0.03125 -en 0.007625 1 0.24 -en 0.0085 2 0.00294 -en 0.008625 2 0.077 -en 0.0425 1 0.12 -Smark -N 100000 -Sp 0.500000 -oOC -Smu 0.006000 -SI 0.008400 2 0.000000 0.000000 -Sc 0 2 20000.000000 10000.000000 0.000000 -SFC -oFP 0.0000000
European	0.1	0.001	java -Xmx3G -Xms3G -jar msms.jar 200 50 -t 6000.000000 -r 8959.991040 1000000 -I 2 100 100 -ej 0.00875 2 1 -en 0.0005 1 0.24 -

			en 0.000875 2 0.077 -en 0.001 2 0.0125 -en 0.001125 2 0.077 -en 0.0075 1 0.03125 -en 0.007625 1 0.24 -en 0.0085 2 0.00294 -en 0.008625 2 0.077 -en 0.0425 1 0.12 -Smark -N 100000 -Sp 0.500000 -oOC -Smu 0.006000 -Sl 0.008400 2 0.100000 0.100000 -Sc 0 2 200.000000 100.000000 0.000000 -SFC -oFP 0.0000000
European	0.1	0.01	java -Xmx3G -Xms3G -jar msms.jar 200 50 -t 6000.000000 -r 8959.991040 1000000 -l 2 100 100 -ej 0.00875 2 1 -en 0.0005 1 0.24 -en 0.000875 2 0.077 -en 0.001 2 0.0125 -en 0.001125 2 0.077 -en 0.0075 1 0.03125 -en 0.007625 1 0.24 -en 0.0085 2 0.00294 -en 0.008625 2 0.077 -en 0.0425 1 0.12 -Smark -N 100000 -Sp 0.500000 -oOC -Smu 0.006000 -Sl 0.008400 2 0.100000 0.100000 -Sc 0 2 2000.000000 1000.000000 0.000000 -SFC -oFP 0.0000000
European	0.1	0.1	java -Xmx3G -Xms3G -jar msms.jar 200 50 -t 6000.000000 -r 8959.991040 1000000 -l 2 100 100 -ej 0.00875 2 1 -en 0.0005 1 0.24 -en 0.000875 2 0.077 -en 0.001 2 0.0125 -en 0.001125 2 0.077 -en 0.0075 1 0.03125 -en 0.007625 1 0.24 -en 0.0085 2 0.00294 -en 0.008625 2 0.077 -en 0.0425 1 0.12 -Smark -N 100000 -Sp 0.500000 -oOC -Smu 0.006000 -Sl 0.008400 2 0.100000 0.100000 -Sc 0 2 20000.000000 10000.000000 0.000000 -SFC -oFP 0.0000000
European	0.5	0.001	java -Xmx3G -Xms3G -jar msms.jar 200 50 -t 6000.000000 -r 8959.991040 1000000 -l 2 100 100 -ej 0.00875 2 1 -en 0.0005 1 0.24 -en 0.000875 2 0.077 -en 0.001 2 0.0125 -en 0.001125 2 0.077 -en 0.0075 1 0.03125 -en 0.007625 1 0.24 -en 0.0085 2 0.00294 -en 0.008625 2 0.077 -en 0.0425 1 0.12 -Smark -N 100000 -Sp 0.500000 -oOC -Smu 0.006000 -Sl 0.008400 2 0.500000 0.500000 -Sc 0 2 200.000000 100.000000 0.000000 -SFC -oFP 0.0000000
European	0.5	0.01	java -Xmx3G -Xms3G -jar msms.jar 200 50 -t 6000.000000 -r 8959.991040 1000000 -l 2 100 100 -ej 0.00875 2 1 -en 0.0005 1 0.24 -en 0.000875 2 0.077 -en 0.001 2 0.0125 -en 0.001125 2 0.077 -en 0.0075 1 0.03125 -en 0.007625 1 0.24 -en 0.0085 2 0.00294 -en 0.008625 2 0.077 -en 0.0425 1 0.12 -Smark -N 100000 -Sp 0.500000 -oOC -Smu 0.006000 -Sl 0.008400 2 0.500000 0.500000 -Sc 0 2 2000.000000 1000.000000 0.000000 -SFC -oFP 0.0000000
European	0.5	0.1	java -Xmx3G -Xms3G -jar msms.jar 200 50 -t 6000.000000 -r 8959.991040 1000000 -l 2 100 100 -ej 0.00875 2 1 -en 0.0005 1 0.24 -en 0.000875 2 0.077 -en 0.001 2 0.0125 -en 0.001125 2 0.077 -en 0.0075 1 0.03125 -en 0.007625 1 0.24 -en 0.0085 2 0.00294 -en 0.008625 2 0.077 -en 0.0425 1 0.12 -Smark -N 100000 -Sp 0.500000 -oOC -Smu 0.006000 -Sl 0.008400 2 0.500000 0.500000 -Sc 0 2 20000.000000 10000.000000 0.000000 -SFC -oFP 0.0000000
Neutral (Asian + African)			ms 200 100 -t 6000.000000 -r 8959.991040 1000000 -l 2 100 100 -ej 0.00875 2 1 -en 0.0005 1 0.24 -en 0.001 2 0.077 -en 0.00475 2 0.00373 -en 0.004875 2 0.077 -en 0.0075 1 0.03125 -en 0.007625 1 0.24 -en 0.0085 2 0.00294 -en 0.008625 2 0.077 -en 0.0425 1 0.12
Neutral (European + African)			ms 200 100 -t 6000.000000 -r 8959.991040 1000000 -l 2 100 100 -ej 0.00875 2 1 -en 0.0005 1 0.24 -en 0.000875 2 0.077 -en 0.001 2 0.0125 -en 0.001125 2 0.077 -en 0.0075 1 0.03125 -en 0.007625 1 0.24 -en 0.0085 2 0.00294 -en 0.008625 2 0.077 -en 0.0425 1 0.12