Title

The Projection of a Test Genome onto a Reference Population and Applications to Humans and Archaic Hominins

Permalink

https://escholarship.org/uc/item/6pp933nh

Journal

Genetics, 198(4)

ISSN

0016-6731

Authors

Yang, Melinda A
Harris, Kelley
Slatkin, Montgomery

Publication Date

2014-12-01

DOI

10.1534/genetics.112.145359

Peer reviewed

The Projection of a Test Genome onto a Reference Population and Applications to Humans and Archaic Hominins

Melinda A. Yang,* Kelley Harris,† and Montgomery Slatkin*,[1]
*Department of Integrative Biology and †Department of Mathematics, University of California, Berkeley, California 94720

ABSTRACT We introduce a method for comparing a test genome with numerous genomes from a reference population. Sites in the test genome are given a weight, w, that depends on the allele frequency, x, in the reference population. The projection of the test genome onto the reference population is the average weight for each x, $\overline{w}(x)$. The weight is assigned in such a way that, if the test genome is a random sample from the reference population, then $\overline{w}(x) = 1$. Using analytic theory, numerical analysis, and simulations, we show how the projection depends on the time of population splitting, the history of admixture, and changes in past population size. The projection is sensitive to small amounts of past admixture, the direction of admixture, and admixture from a population not sampled (a ghost population). We compute the projections of several human and two archaic genomes onto three reference populations from the 1000 Genomes project—Europeans, Han Chinese, and Yoruba—and discuss the consistency of our analysis with previously published results for European and Yoruba demographic history. Including higher amounts of admixture between Europeans and Yoruba soon after their separation and low amounts of admixture more recently can resolve discrepancies between the projections and demographic inferences from some previous studies.

THE wealth of genomic data now available calls for new methods of analysis. One class of methods estimates parameters of demographic models using samples from multiple populations. Such methods are computationally challenging because they require the simultaneous analysis of genetic drift in several populations under various model assumptions. The demographic models analyzed with these methods are defined in terms of the parameters needed to describe the past growth of each population, their times of divergence from one another, and the history of admixture among them.

Gutenkunst *et al.* (2009) developed an efficient way to numerically solve a set of coupled diffusion equations and then search parameter space for the maximum-likelihood parameter estimates. Their program *dadi* can analyze data from as many as three populations. Harris and Nielsen (2013) use the length distribution of tracts identical by descent within and between populations to estimate model parameters. Their program (unnamed) can handle the same degree of demographic complexity as *dadi*. Excoffier *et al.* (2013) use coalescent simulations to generate the joint frequency spectra under specified demographic assumptions. Their program *fastsimcoal2* approximates the likelihood and then searches for the maximum-likelihood estimates of the model parameters. Using simulations instead of numerical analysis allows *fastsimcoal2* to analyze a much larger range of demographic scenarios than *dadi*. Schiffels and Durbin (2014) recently introduced the multiple sequential Markovian coalescent (MSMC) model, which is a generalization of the pairwise sequential Markovian coalescent model (Li and Durbin 2011). MSMC uses the local heterozygosity of pairs of sequences to infer past effective population sizes and times of divergence.

These and similar methods are especially useful for human populations for which the historical and archaeological records strongly constrain the class of models to be considered. Although human history is much more complicated than tractable models can describe, those models can nonetheless reveal important features of human history that have shaped current patterns of genomic variation.

In this article, we introduce another way to characterize genomic data from two or more populations. Our method is designed to indicate the past relationship between a single

doi: 10.1534/genetics.112.145359
Manuscript received July 28, 2014; accepted for publication October 9, 2014; published Early Online October 15, 2014.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.112.145359/-/DC1.

[1] Corresponding author: Department of Integrative Biology, VLSB 3060, University of California, Berkeley, CA 94720-3140. E-mail: slatkin@berkeley.edu

genome and one or more populations that have already been well studied. Our method is particularly useful for detecting small amounts of admixture between populations and the direction of that admixture, but it can also indicate population size changes. Furthermore, it can also serve as a test of consistency with results obtained from other methods. We first introduce our method and apply it to models of two and three populations, focusing on the effects of gene flow and bottlenecks. Then we present the results of analyzing human and archaic hominin genomes. Some of the patterns in the data are consistent with simple model predictions and others are not. We explore specific examples in some detail to show how our method can be used in conjunction with others. Finally, we use projection analysis to test demographic inferences for European and Yoruba populations obtained from the four previous studies described above.

Analytic Theory

We assume that numerous individuals from a single population, which we call the "reference population," have been sequenced. We also assume that there is an outgroup that allows determination of the derived allele frequency, x, at every segregating site in the reference population. We define the projection of another genome, which we call the "test genome," onto the reference population. For each segregating site in the reference population, a weight, w, is assigned to that site in the test genome as follows. If the site is homozygous ancestral, then $w = 0$; if it is heterozygous, then $w = 1/(2x)$; and if it is homozygous derived, then $w = 1/x$. The projection $\overline{w}(x)$ is the average weight of sites in the test genome at which the frequency of the derived allele in the reference population is x.

With this definition of the projection, $\overline{w}(x) = 1$ independently of x if the test genome is randomly sampled from the reference population. Therefore, deviation of $\overline{w}(x)$ from 1 indicates that the test genome is from another population. To illustrate, assume that the test and reference populations have been of constant size N, that they diverged from each other at a time τ in the past, and that there has been no admixture between them since that time. The results of Chen *et al.* (2007) show that in this model $\overline{w}(x) = e^{-\tau/(2N)}$ independently of x.

Analytic results are not as easily obtained for other models. We used numerical solutions to the coupled diffusion equations when possible and coalescent simulations when necessary to compute the projection under various assumptions about population history. For all models involving two or three populations, numerical solutions for each set of parameter values were obtained from *dadi* (Gutenkunst *et al.* 2009). Models with more than three populations were simulated using *fastsimcoal2* (Excoffier *et al.* 2013).

For all models that we considered, an ancestral effective population size (N_e) of 10,000 with a generation time of 25 years was used. We assumed 150 individuals were sampled from the reference population and one from the test population. In *dadi* and *fastsimcoal2*, the resulting frequency spectrum was transformed into the projection for each frequency category. The parameters used are described in the figure legends.

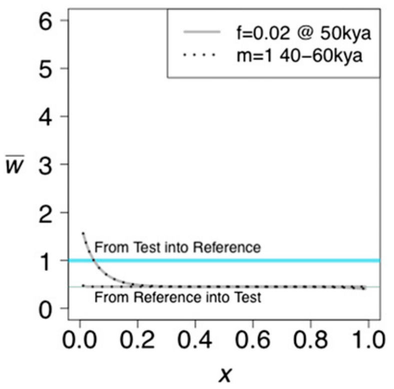


Figure 1 The effect of unidirectional gene flow on the projection of a test genome onto a reference population. Two kinds of gene flow were assumed: either a single pulse of admixture of strength f or a period of immigration at a rate m per generation. Both populations are of constant size n = 10,000. The divergence time is 400 KYA.

Two populations

We first consider two populations of constant size that separated τ generations in the past and experienced gene flow between them after their separation. We allow for two kinds of gene flow: (1) a single pulse of admixture in which a fraction f of one population is replaced by immigrants from the other and (2) a prolonged period of migration during which a fraction m of the individuals in one population are replaced each generation by immigrants from the other. We allow for gene flow in each direction separately. Figure 1 shows typical results. Gene flow from the reference into the test population has no detectable effect while gene flow from the test into the reference population results in the following pattern: $\overline{w}(x)$ decreases monotonically to the value expected in the absence of gene flow. Even very slight gene flow in this direction creates the observed pattern. The projection is not able to distinguish between a single pulse and a prolonged period of gene flow, however. By adjusting the parameters, the projection under the two modes of gene flow can be made the same, as shown.

The intuitive explanation for the effect of gene flow from the test to the reference is that gene flow carries some alleles that were new mutations in the test population. Those alleles will necessarily be in low frequency in the reference population because they arrived by admixture, but they are likely to be in higher frequency in the test population because they were carried by admixture to the reference. Therefore, they will be seen in the test genome more often than expected on the basis of their frequency in the reference population.

The projection deviates from a horizontal line when there is a bottleneck in the reference (Figure 2A, black line) or ancestral population (Figure 2A, blue line), but not when there is a bottleneck in the test population (Figure 2A, red line). The reason for the humped shape of the projection

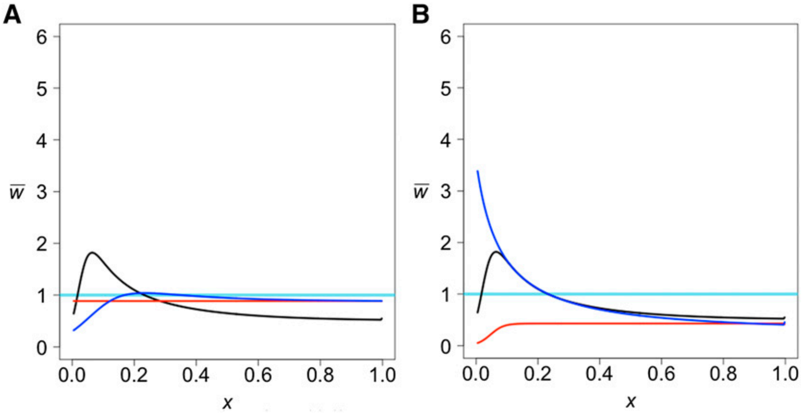


Figure 2 The effect of population size changes in a model with two populations that diverged 60 KYA. (A) A bottleneck occurs in the reference population (black), the test population (red), or the ancestral population (blue). During the bottleneck, the population size is reduced from 10,000 to 1000 from 50 to 20 KYA. (B) For the reference population only, a bottleneck occurs as in A (black), a population expansion from 1000 to 10,000 occurs 20 KYA (red), or the reference population decreases in size from 10,000 to 1000 50 KYA (blue). The test population has the same population size as the ancestral population.

when there is a bottleneck in the reference population is that the bottleneck distorts the site frequency spectrum in that population in such a way that there are more rare and more common alleles than in a population of constant size and fewer alleles with intermediate frequency, and it accelerates the rate of loss of alleles that were previously in low frequency. When the reference population size declines without recovering, the effect is an increase in rare alleles, similar to that of admixture into the reference population (Figure 2B, blue line). When the reference population expands, a slight decrease in rare alleles is observed (Figure 2B, red line).

A bottleneck followed by admixture amplifies the effect of admixture (Figure 3A, black line) while admixture that occurs before or during the bottleneck does not change the shape of the projection as much (Figure 3A, red and blue lines). The effect comes from the increase in population size at the end of the bottleneck, not from the decrease at the beginning (Figure 3B).

Three populations

Three populations lead to a greater variety of effects than can be seen in two. Because samples are analyzed from only two of the populations, the test and the reference, the third population is unsampled. We will follow Beerli (2004) and call the unsampled population a "ghost population." In some situations, all populations may be sampled, but only two at a time are analyzed. In others situations, no samples are available from a population that is known or suspected to have admixed with one or more of the sampled populations. In the latter case, one goal is to determine whether or not there has been admixture from the ghost population.

We first consider the effects of gene flow alone. We will assume a single pulse of admixture of strength f at time t_{GF}. There are three distinct topologies representing the ancestry of the three populations (Figure 4). Gene flow can be from the ghost population into either the test or the reference population. Gene flow from the ghost into the test population has little effect on the projection (Figure 5, A–C), whereas gene flow from the ghost into the reference has an effect that depends on the population tree topology. If the test and ghost populations are sister groups (Figure 4A and Figure 5D), the effect is similar to that of gene flow directly from the test into the reference population (Figure 1). The increase of $\overline{w}(x)$ for small x results from mutations that arose in the ancestral population of the ghost and test populations and then entered the reference population through migration from the ghost population. The magnitude of the ghost gene flow effect thus depends on the length of the internal branch directly ancestral to the ghost and test populations. When there is a longer period of shared ancestry between the test and ghost populations, the admixture has a stronger effect (Figure 5D).

In the second topology (Figure 4B), the reference and ghost populations are sister groups. Here, gene flow from

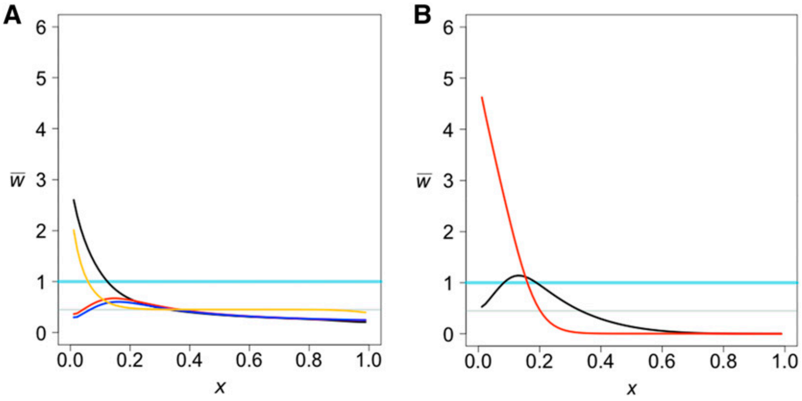


Figure 3 The combined effect of a bottleneck and admixture. The divergence time for both models is 100 KYA. (A) The yellow projection represents no bottleneck but admixture of f = 0.02 at 40 KYA. The other projections include admixture at 40 KYA (black), 80 KYA (red), and 120 KYA (blue) of 0.02 from the test to the reference, where there was a bottleneck from 70 to 90 KYA. The bottleneck reduced the reference population size from 10,000 to 1000, and then the population size increased to 10,000. (B) The reference population size increased from 1000 to 10,000 at 40 KYA only. Admixture of 0.02 from the test to the reference occurred at 30 KYA (red) and 50 KYA (black).

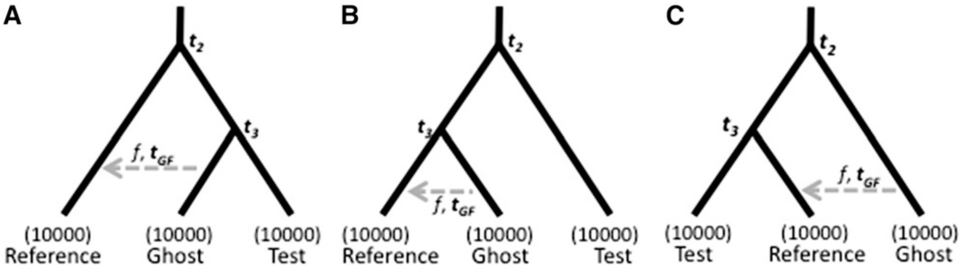


Figure 4 Illustration of three possible population relationships in which there is a pulse of admixture of intensity f at time t_{GF} in the past from the ghost population into the reference population. t_2 and t_3 are the times of population separation. In each topology, either the test and ghost (A), reference and ghost (B), or the test and reference (C) are more closely related to each other than the third population.

the ghost population also increases $\overline{w}(x)$ for small x, but the magnitude of the increase is inversely related to the length of the internal branch ancestral to the ghost and reference populations. The increase of $\overline{w}(x)$ at low frequencies results from alleles that arose in the common ancestral population, drifted to low frequency or loss in the reference population, and by chance drifted to high frequency in both the ghost and test populations. There is little room for this to happen when the reference and ghost populations have diverged very recently and have essentially the same allele frequencies (Figure 5E). When the reference and test populations are sister groups (Figure 4C), and the ghost population is an outgroup, a dip is observed for low frequencies (Figure 5F).

If there is a bottleneck in the reference population after admixture, the effect (Figure 6A) is similar to that seen in the two-population case (Figure 3). The signal of admixture is amplified. In the case where the reference and ghost populations are sister groups (Figure 6B), the characteristic bottleneck effect is observed. As the time of divergence between the reference and ghost population increases, the humped shape due to the bottleneck is reduced in size, presumably due to the increased effect of admixture. When the reference and test populations are sister groups, the humped shape remains, but the effect is reduced as the time of divergence increases (Figure 6C), and the increase in common alleles is still observed.

Ancestral misidentification

Misidentification of the ancestral allele leads to the assumption that an allele is ancestral when it is in fact derived or that an allele is derived when it is in fact ancestral. Hernandez *et al.* (2007) show that ancestral misidentification occurs at levels of ~1–5% in human genome data sets. We use *ms* (Hudson 2002) to simulate two simple demographic

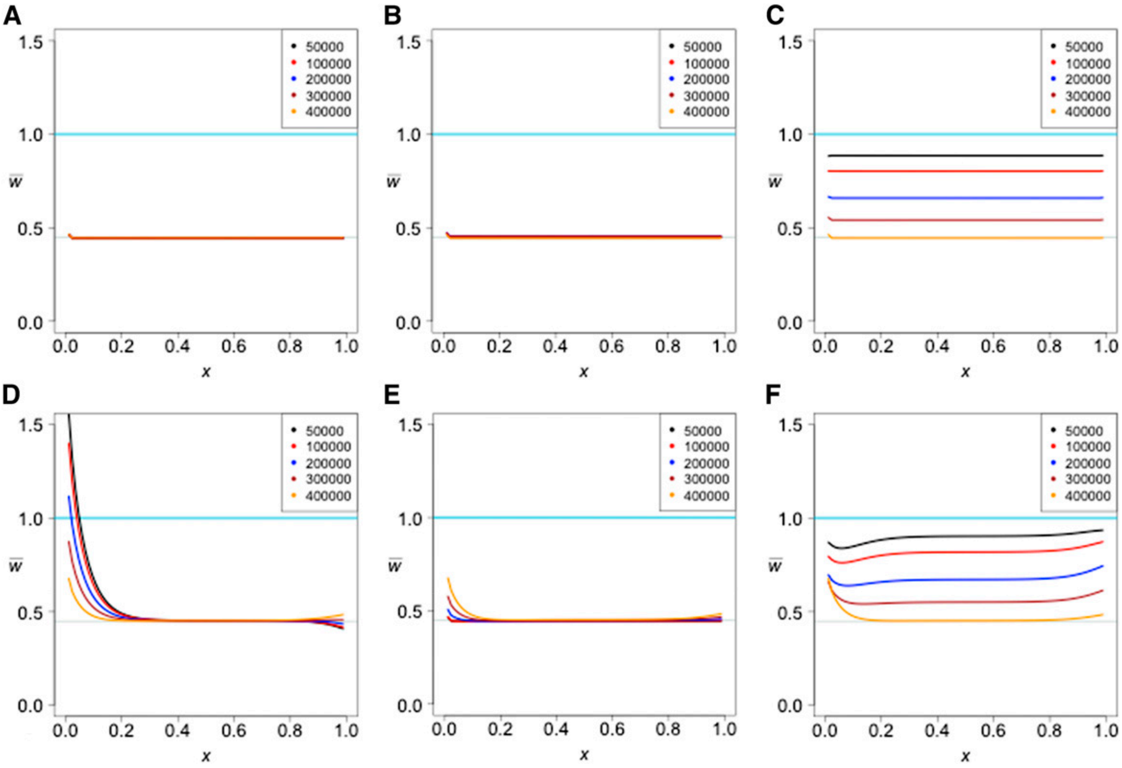


Figure 5 The effect of ghost admixture into the test (A–C) and the reference (D–F). A and D follow the topology in Figure 4A; B and E follow the topology in Figure 4B; and C and F follow the topology in Fig. 4C. t_2 = 400 KYA, f = 0.02, and t_{GF} = 50 KYA. t_3 is varied from 50 to 400. Population sizes remain constant at 10,000.

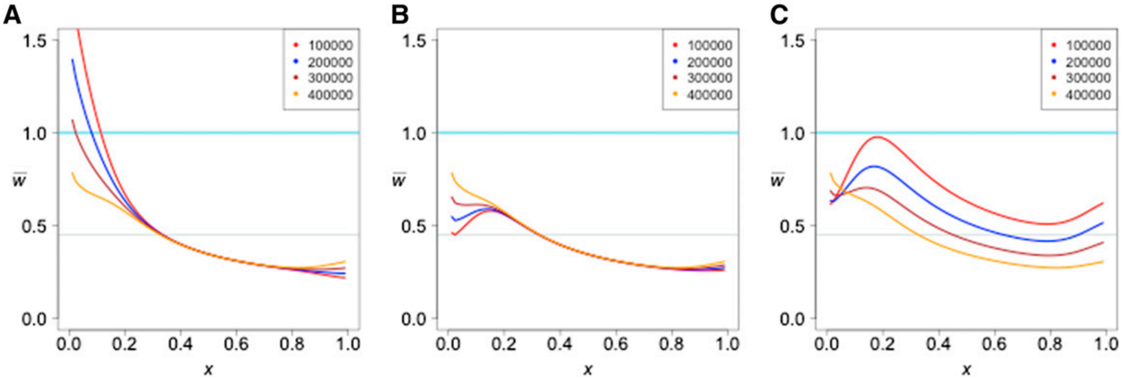


Figure 6 The effects of ghost admixture into the reference with a bottleneck in the reference occurring 70–100 KYA changing the reference population size from 10,000 to 1000 and back to 10,000. t_3 is varied from 100 to 400 KYA. All other parameters are the same as in Figure 5. A follows the topology in Fig. 4A, B follows the topology in Fig. 4B, and C follows the topology in Fig. 4C.

models to determine the effect of ancestral misidentification on the projection: one model has no admixture or population size changes between the reference and test populations and one matches the model with admixture shown in Figure 1. We allowed for 0, 0.1, 1, or 10% of the sites to be misidentified, reversing the ancestral or derived result given by the simulation. Where the frequency spectrum is shown to have an increase for common alleles (Hernandez *et al.* 2007), the projection shows a similar result (Supporting Information, Figure S1).

Application to Humans and Archaic Hominins

We illustrate the use of projection analysis by applying it to genomic data from present-day humans and two archaic hominins (Neanderthal and Denisovan). For the reference populations, we used data from the 1000 Genomes (1000G) project for three populations: Europeans (CEU), Han Chinese (CHB), and Yoruba (YRI) (1000 Genomes Project Consortium 2010). For test genomes, we used the high-coverage Denisovan genome (Meyer *et al.* 2012), the high-coverage Neanderthal genome (Prüfer *et al.* 2014), and some of the high-coverage present-day human genomes sequenced by Meyer *et al.* (2012). We will identify the reference populations by the 1000G abbreviation (CEU, CHB, and YRI) and the test genomes by the labels used by Meyer *et al.* (2012). These labels are provided in a note in Table 1. We used only autosomal biallelic sites with data present in every individual and population sampled. We used the reference chimpanzee genome *PanTro2* to determine the derived and ancestral allele at each site and filtered out all CpG sites.

To show that projections give insight into human demographic history, we developed a 10-population demographic history with realistic parameters taken from the literature and adjusted using different curve-fitting techniques (Table 1 and Figure 7). The initial parameter ranges that we chose were informed by a variety of previous studies, as noted in Table 1. To improve the fit of the simulated model to the projections, we used two techniques. Initially, we focused on two populations at a time. Using *dadi* (Gutenkunst *et al.* 2009) and the Broyden–Fletcher–Goldfarb–Shanno algorithm (Morales and Nocedal 2011), we estimated several demographic parameters simultaneously that gave the best-fitting projection for the two populations. For more than two populations, we used *fastsimcoal2* (Excoffier *et al.* 2013) and Brent's algorithm to vary one parameter at a time, fixing all other parameters. The parameters of interest were cycled through, each varied in turn, until a better-fitting projection could not be found. This technique tended to converge most quickly when we focused on no more than three or four parameters at a time. For both techniques, we used least squares summation (LSS) to determine the best fit.

The demographic scenario displayed in Figure 7 is not meant to be optimal. Instead, it is intended to show that, for a plausible scenario, the predicted projections are similar to ones computed from the data. This model illustrates the sensitivity of projections to major demographic processes that have shaped human history. Here, we note what features of demographic history are necessary to give rise to projections similar to those observed.

Comparison of observed projections to each other

The black curves in Figure 8, Figure 9, Figure 10, and Figure 11 represent the observed projections. The projections were smoothed using a cubic spline and a smoothing parameter of 0.5. This was done to reduce the effect of sampling error in comparisons with the expected projections for the 10-population demographic scenario described in Figure 7, which are represented by the red curves in Figure 9, Figure 10, and Figure 11. Table S1, Table S2, and Table S3 provide the LSS comparing the projections of each test genome onto each reference population, and the diagonal terms provide the LSS for that test genome, relative to the $\overline{w}(x) = 1$ line. The observed projections show that the Neanderthal and Denisovan projections onto CEU, CHB, and YRI look the most different from the $\overline{w}(x) = 1$ line.

Comparison of a test genome with the same population

In Figure 8A, the projection of the French genome onto CEU fits the expectation except for small x. Similar deviations are

Table 1 Description of parameters used in the simulation of the 10-population tree in Figure 7

Description	Parameter	Value	Initial range	Reference	Comments
Effective population size in the present day for each population	N_{DEN}	500	100–5,000	Prüfer *et al.* (2014)	A small effective population size was used for the archaic hominins.
	N_{NEA}	500	100–5,000	Prüfer *et al.* (2014)	
	N_{FRE}	30,000	10,000–40,000	Gravel *et al.* (2011); Schiffels and Durbin (2014)	A large effective population size was used to allow for population expansion.
	N_{HAN}	45,000	10,000–90,000	Gravel *et al.* (2011); Schiffels and Durbin (2014)	
	N_{PAP}	15,000	10,000–40,000		The initial range was set to the same as that for N_{FRE}.
	N_{DIN}	6,000	5,000–40,000		A lower effective population size improved the fit of the Dinka projections.
	N_{YOR}	10,000	10,000–40,000	Gravel *et al.* (2011); Schiffels and Durbin (2014)	The Yoruba population does not have the large population expansion observed in non-Africans.
	N_{MAN}	10,000	NA		The value was set to the same as that for N_{YOR}.
	N_{MBU}	10,000	NA		The value was set to the same as that for N_{YOR}.
	N_{SAN}	10,000	NA		The value was set to the same as that for N_{YOR}.
Population size changes moving backward in time. A value <1 indicates an expansion and a value >1 indicates a decline.	$N_{\mathrm{ANC1}}/N_{\mathrm{FRE}}$	0.2	0.01–1	Gravel *et al.* (2011); Excoffier *et al.* (2013); Harris and Nielsen (2013); Prüfer *et al.* (2014); Schiffels and Durbin (2014)	European population expansion
	$N_{\mathrm{ANC2}}/N_{\mathrm{HAN}}$	0.1	0.01–1	Prüfer *et al.* (2014)	East Asian population expansion
	$N_{\mathrm{ANC3}}/N_{\mathrm{PAP}}$	0.1	0.01–1	Prüfer *et al.* (2014)	Papuan population expansion
	$N_{\mathrm{ANC4}}/N_{\mathrm{YOR}}$	4.5	1.0–10	Excoffier *et al.* (2013); Prüfer *et al.* (2014); Schiffels and Durbin (2014)	A Yoruba population decline improves the fit of the projections onto reference YRI.
	$N_{\mathrm{ANC5}}/N_{\mathrm{ANC1}}$	4	1.0–10	Gravel *et al.* (2011); Harris and Nielsen (2013); Prüfer *et al.* (2014)	Non-African population decline
	$N_{\mathrm{ANC6}}/N_{\mathrm{ANC5}}$	0.9	0.5–1	Gravel *et al.* (2011); Excoffier *et al.* (2013); Harris and Nielsen (2013); Prüfer *et al.* (2014); Schiffels and Durbin (2014)	Ancestral population expansion
Time of Yoruba–Mandenka admixture	T_0	25	NA	Prüfer *et al.* (2014)	The Mandenka and Yoruba populations are closely related, so a recent divergence and admixture time were assumed.
Time of Yoruba–Mandenka divergence	T_1	50	0–1,000		
Time of French–Han–Yoruba admixture	T_2	300	NA		Recent admixture occurred after population expansion.
Time of French, Han, Papuan population size expansion	T_3	350	NA	Schiffels and Durbin (2014)	We assumed that population expansion occurred roughly halfway between the start of expansion and the present.

(continued)

Table 1, *continued*

Description	Parameter	Value	Initial range	Reference	Comments
Time of French–Han divergence	T_4	1,200	600–1,800	Gravel *et al.* (2011)	The value providing the best projections for the French and Han is earlier than the estimated time of divergence in Gravel *et al.* (2011).
Time of Yoruba–Dinka/San/Mbuti/ ancestral admixture, Yoruba population decline	T_5	1,500	NA		Projections onto reference YRI fit best when the time of the Yoruba population decline occurred at this time. Admixture times were also placed here for convenience. Changing the time of admixture did not affect the projection substantially.
Time of Denisovan–Papuan admixture	T_6	1,600	1,200–1,800	Meyer *et al.* (2012)	The time of admixture was placed after the divergence of Papuans from other non-Africans, at a time that could be reasonable for contact between Denisovans and Papuans.
Time of French–Han–Papuan divergence	T_7	1,800		Wollstein *et al.* (2010)	The Papuan divergence time was placed ancestral to the French/ Han divergence because the Papuans had to diverge early enough that admixture with Denisovans was reasonable.
Time of Neanderthal admixture into ancestral non-Africans and the time ancient hominins were sampled	T_8	2,000	NA	Prüfer *et al.* (2014)	The admixture time was set to 50 KYA.
Time of Yoruba admixture with ancestral non-Africans	T_9	2,100	2,000–4000	Gutenkunst *et al.* (2009); Schiffels and Durbin (2014);	The time of higher admixture is earlier than the Neanderthal admixture into non-Africans, to avoid the Yoruba population exhibiting high amounts of admixture from Neanderthals.
Time of Dinka divergence	T_{10}	6,000	NA	Prüfer *et al.* (2014)	The non-African and Dinka divergence time was placed between the Eurasian and Papuan divergence and the Yoruba and non-African divergence.
Time of Yoruba divergence	T_{11}	6,300	1,500–8,000	Gutenkunst *et al.* (2009); Schiffels and Durbin (2014); (1000 Genomes Project Consortium 2010; Gravel *et al.* 2011; Excoffier *et al.* 2013; Harris and Nielsen 2013).	An older divergence time provided a better fit for the Yoruba projections than a younger divergence time.
Time of Mbuti divergence	T_{12}	7,000	NA	Prüfer *et al.* (2014)	The Mbuti and non-African divergence was placed between the Yoruba and non-African divergence, and the San and non-African divergence.

(continued)

Table 1, *continued*

Description	Parameter	Value	Initial range	Reference	Comments
Time of San divergence	T_{13}	8,000	NA	Prüfer *et al.* (2014)	The San and non-African divergence is the earliest human divergence.
Time of Neanderthal–Denisovan admixture	T_{14}	12,000	8000–21,000	Prüfer *et al.* (2014)	An earlier time of admixture and divergence allowed for a better fit of the Denisova projection.
Time of Denisovan Divergence from Neanderthals	T_{15}	21,000	12,000–26,000		
Time of Neanderthal/Denisovan Divergence from Humans	T_{16}	26,000	22,000–30,600	Prüfer *et al.* (2014)	An older divergence allows for a better fit of the Neanderthal projection.
Admixture from the left population to the right population	$f_{\mathrm{MAN–YOR}}$	0.1	0–0.15	Prüfer *et al.* (2014)	With the close relationship between these two populations, admixture was allowed.
	$f_{\mathrm{YOR–MAN}}$	0.1	0–0.15		
	$f_{\mathrm{FRE–HAN}}$	0.03	0–0.15	Gravel *et al.* (2011); Harris and Nielsen (2013)	The increase in rare alleles observed for these populations in several projections can be generated if there is a small amount of admixture between these populations.
	$f_{\mathrm{HAN–FRE}}$	0.01	0–0.15		
	$f_{\mathrm{FRE–YOR}}$	0.001	0-0.15		
	$f_{\mathrm{YOR–FRE}}$	0.005	0–0.15		
	$f_{\mathrm{YOR–HAN}}$	0.003	0–0.15		
	$f_{\mathrm{SAN–YOR}}$	0.05	0–0.15		
	$f_{\mathrm{MBU–YOR}}$	0.05	0–0.15		
	$f_{\mathrm{DIN–YOR}}$	0.01	0–0.15		
	$f_{\mathrm{ANC1–YOR}}$	0.01	0–0.15		
	$f_{\mathrm{ANC1–ANC4}}$	0.4	0–0.5	Gravel *et al.* (2011); Schiffels and Durbin (2014)	The projections of non-African populations onto reference YRI fit better when high levels of ancestral admixture were assumed.
	$f_{\mathrm{ANC4–ANC1}}$	0.2	0–0.5		
	$f_{\mathrm{NEA–DEN}}$	0.01	0–0.05	Prüfer *et al.* (2014)	Low amounts of admixture from archaic hominins were added.
	$f_{\mathrm{DEN–PAP}}$	0.03	0–0.05	Prüfer *et al.* (2014)	
	$f_{\mathrm{NEA–ANC1}}$	0.03	0–0.05	Prüfer *et al.* (2014)	

The initial range is the set of values that was explored for each parameter. "NA" indicates that the parameter was not varied. The initial range choices were based on the articles cited, although the ranges were sometimes expanded to explore the effects of more values. Times are in generations, with 1 generation = 25 years. DEN, Denisovan; DIN, Dinka; FRE, French; MAN, Mandenka; MBU, Mbuti; NEA, Neanderthal; PAP, Papuan; YOR, Yoruba; HAN, Han Chinese; SAN, San. The labels refer to the high coverage individuals from Meyer *et al.* (2012). ANC1-5 refer to the ancestral human populations older than the divergence into modern populations. The corresponding ancestral population can be found in the topology in Figure 7.

seen in Figure 8B in the projection of the Han genome onto CHB and, to a lesser extent, in Figure 8C in the projection of the Yoruba genome onto YRI. This pattern is expected for the smallest frequency classes because the frequency spectrum in the reference populations has more singletons than expected in a population at equilibrium under drift and mutation. See the Appendix for details.

Admixture with Neanderthals and Denisovans

Our simulations show that a bottleneck combined with admixture into the reference population can result in a strong effect on the projection (Figure 3A, black curve). The projections of the Altai Neanderthal onto CEU and CHB show a large excess of rare alleles (Figure 9I and Figure 10I), which requires the combination of a bottleneck in the ancestors of non-Africans and admixture from Neanderthals into non-Africans after that bottleneck. Including both processes in our model, we obtain good fits to the observed projections (Table 2, Figure 9I, and Figure 10I). When admixture is omitted, the result is a decrease in the excess of rare alleles and a worse fit (Table S4 and Figure S2).

Similarly, the projections of the Denisovan genome onto CEU and CHB (Figure 9H and Figure 10H) are consistent with the three-population analysis shown in Figure 4A and Figure 5D. In this case, Neanderthals are the ghost population and Denisovans are the test population. The excess of rare alleles for the Denisova projection is consistent with Neanderthals and Denisovans being sister groups. Some of the new mutations that arose in the shared branch between Neanderthals and Denisovans are carried by admixture to humans and their presence is seen in the projection as an excess of rare alleles (Table 2, Figure 9H, and Figure 10H). The Denisovan projections give a signal of admixture but it is weaker than the signal in the Neanderthal projections.

The projections of the Neanderthal (Figure 11I) and Denisovan (Figure 11H) onto YRI show a signal of admixture even though previous analysis of the Neanderthal genome did not find evidence of direct Neanderthal admixture from the presence of identifiable admixed fragments (Prüfer *et al.* 2014). These projections are consistent with the signal of Neanderthal introgression being carried by recent admixture

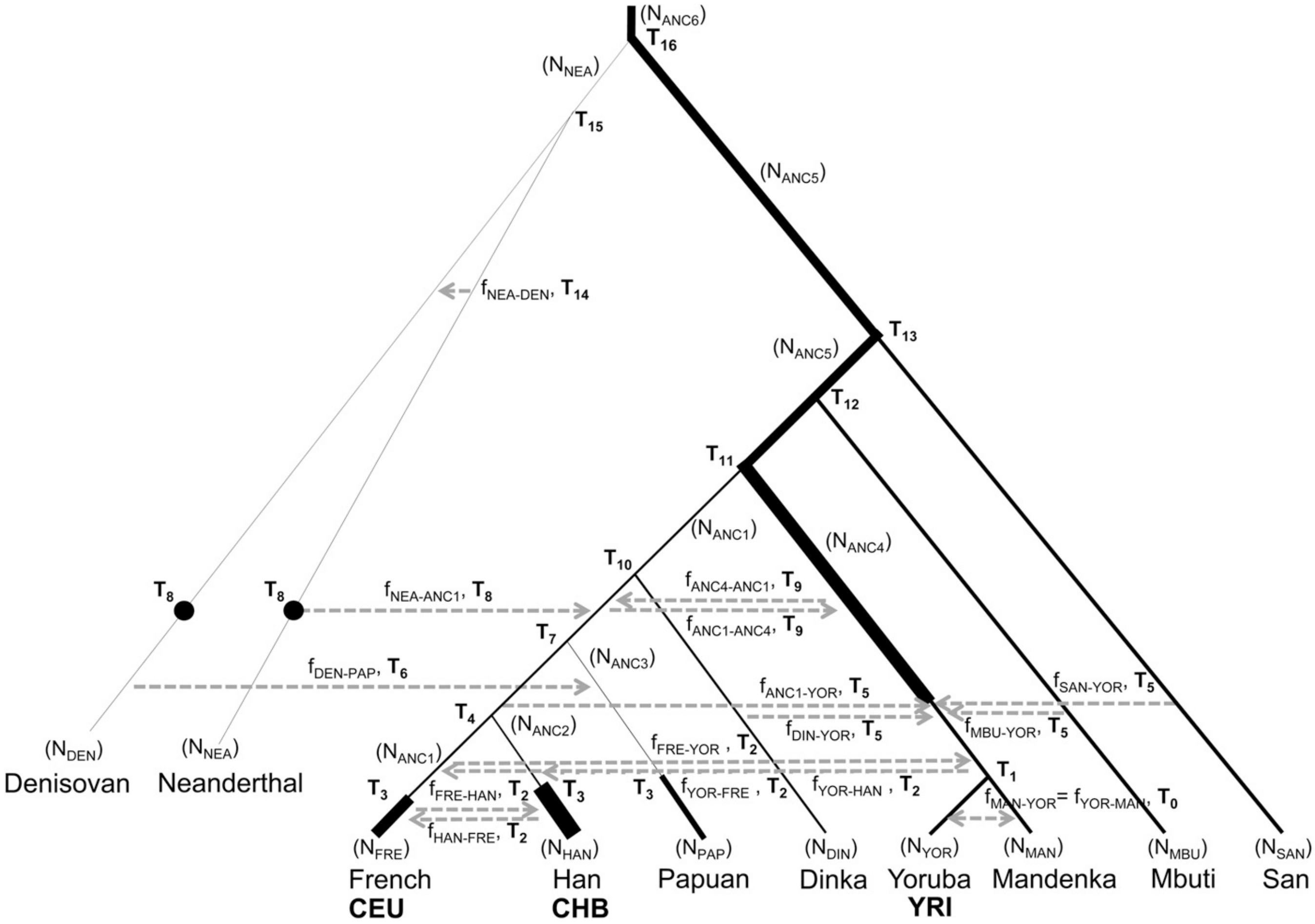


Figure 7 A model of human demographic history for 10 populations that, when simulated, gave projections similar to the observed projections. The populations in boldface type are the reference populations, and the row above them indicates the population origin of each test genome. The values used are found in Table 1. Black dots indicate the time of sampling if not in the present day. Thickness of the branch gives an approximation of the change in effective population size.

from the ancestors of Europeans and East Asians into the ancestors of the Yoruba population. In our model (Figure 7), there is no admixture between an African population and any archaic hominin, but there is gene flow between the ancestors of the Yoruba population and non-Africans. An excess of rare alleles is observed in the simulated projection (Figure 11, H and I). Admixture from non-Africans to Yoruba had to have occurred more recently than the Neanderthal admixture into non-African populations for this signal to be present.

The Altai Neanderthal genome is unusual in that it is marked by long runs of homozygosity, indicating the individual was highly inbred. Prüfer *et al.* (2014) show that the inbreeding coefficient was 1/8. This inbreeding has no effect on the projection, however, because the projection effectively samples a haploid genome from the test individual.

Relationship among non-African populations

The projection of the French genome onto CHB (Figure 10A) differs from the projection of the Han genome onto CEU (Figure 9A). This difference reflects the subtle interplay between admixture and population size changes. A model in which the ancestors of East Asians experienced a bottleneck after their separation from the ancestors of Europeans along with a greater rate of population expansion can explain why the humped shape characteristic of bottlenecks was not swamped out by the signal of admixture. The inclusion of more admixture from Europeans to East Asians can account for the overall increased excess seen in the French projection onto CHB (Figure 10A). When these events are included in our model, the resulting projections are relatively close to the observed projections (Table 2).

The Papuan demographic history modeled here includes divergence from the ancestors of Europeans and East Asians and a bottleneck and population expansion (Figure 7). In this model, we simulated a demographic history in which the Papuans diverged from the population ancestral to Europeans and East Asians, a scenario supported by Wollstein *et al.* (2010), but not by others (Meyer *et al.* 2012; Prüfer *et al.* 2014). We made this assumption because we followed Gravel *et al.* (2011) in assuming that Europeans and East Asians diverged relatively recently. With admixture from Denisovans

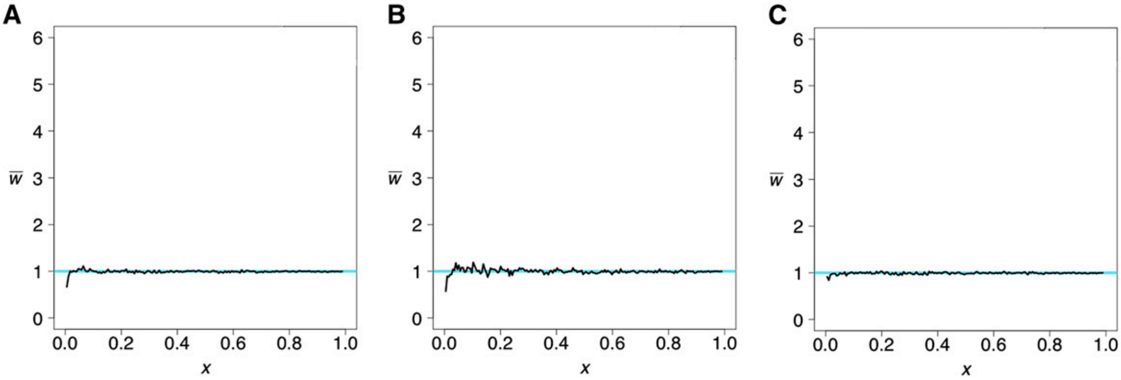


Figure 8 The projections of French onto CEU (A), Han onto CHB (B), and Yoruba onto YRI (C). The sum of LSS scores comparing the observed projection to the line are found in Table S1, Table S2, and Table S3.

to Papuans occurring earlier, assuming the Papuans were the outgroup to Europeans and East Asians was more appropriate. Using this model, the projections fit relatively well (Table 2, Figure 9B, and Figure 10B).

Relationship between non-Africans and YRI

The projections of the Papuan, French, and Han genomes onto YRI (Figure 11, A, B, and D) are similar despite the difference between the Han and Papuan projections onto CEU (Figure 9, A and B). These observations can be accounted for if there were high levels of admixture between the ancestors of non-Africans and the ancestors of the Yoruba population as well as a large ancestral Yoruba population that had declined in the recent past. These two processes together explain the dip observed and the increase to $\overline{w}(x) = 1$ for larger x, and they lead to a good fit to the observed projections (Table 2 and Figure 11, A, B, and D). Varying these two parameters in our model shows their effect on the projection for rare alleles and that higher values for both of these parameters give the best-fitting simulated projections (Table S5 and Figure S3).

African projections onto CEU and YRI

The projections of all five African genomes—San, Yoruba, Mandenka, Dinka, and Mbuti—onto CEU (Figure 9, C–G) are similar to one another and similar to their projections onto CHB (Figure 10, C–G). All these projections are consistent with low levels of admixture from the African populations into the ancestors of Europeans and East Asians. Previous analyses (Lachance *et al.* 2012; Meyer *et al.* 2012; Pickrell *et al.* 2012;

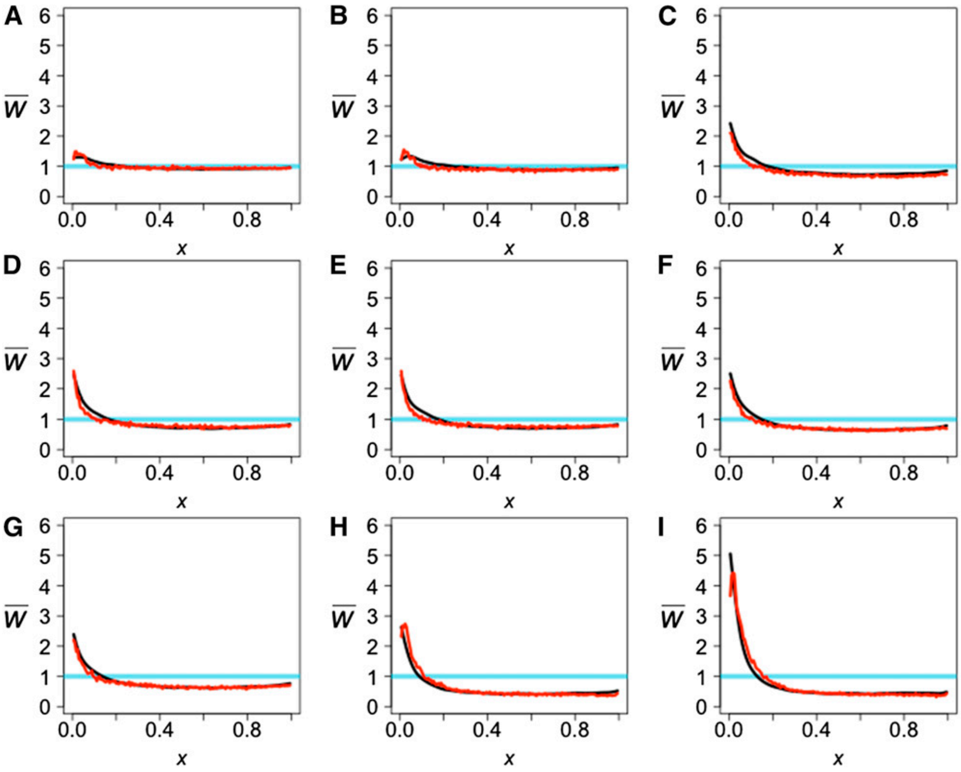


Figure 9 The observed projection (black line) and simulated projection from our model (red line) for the CEU reference population. The test genomes are Han (A), Papuan (B), Dinka (C), Yoruba (D), Mandenka (E), Mbuti (F), San (G), Denisovan (H), and Neanderthal (I). The LSS scores comparing the observed projections to each other and the expectation can be found in Table S1, and the LSS scores comparing the observed and simulated projections can be found in Table 2.

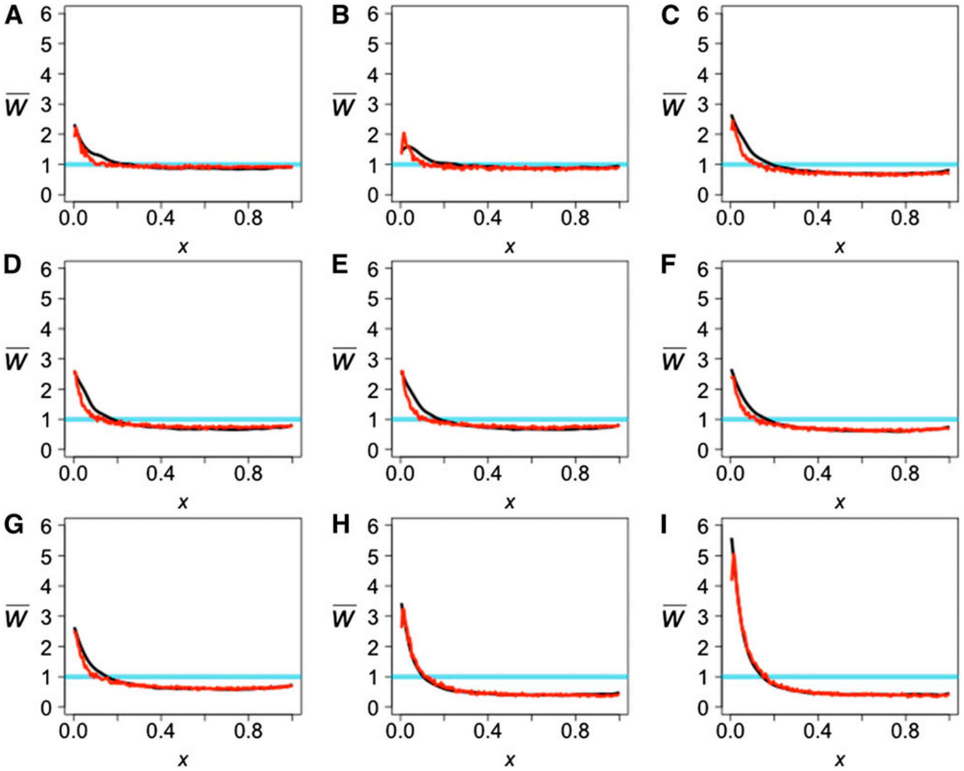


Figure 10 The observed projection (black line) and simulated projection from our model (red line) for the CHB reference population. The test genomes are French (A), Papuan (B), Dinka (C), Yoruba (D), Mandenka (E), Mbuti (F), San (G), Denisovan (H), and Neanderthal (I). The LSS scores comparing the observed projections to each other and the expectation can be found in Table S2, and the LSS scores comparing the observed and simulated projections can be found in Table 2.

Prüfer *et al.* 2014) showed that the San population diverged from other African populations before the other African populations diverged from one another and before the ancestors of Europeans and East Asians diverged from each other. The separate history of the San is not reflected in the projection of the San genome onto CEU and CHB. Because the demographic history in the reference populations has a strong effect on the projections, the bottleneck in Europeans combined with low amounts of admixture between the Yoruba and San and between the Yoruba and non-Africans are enough to give

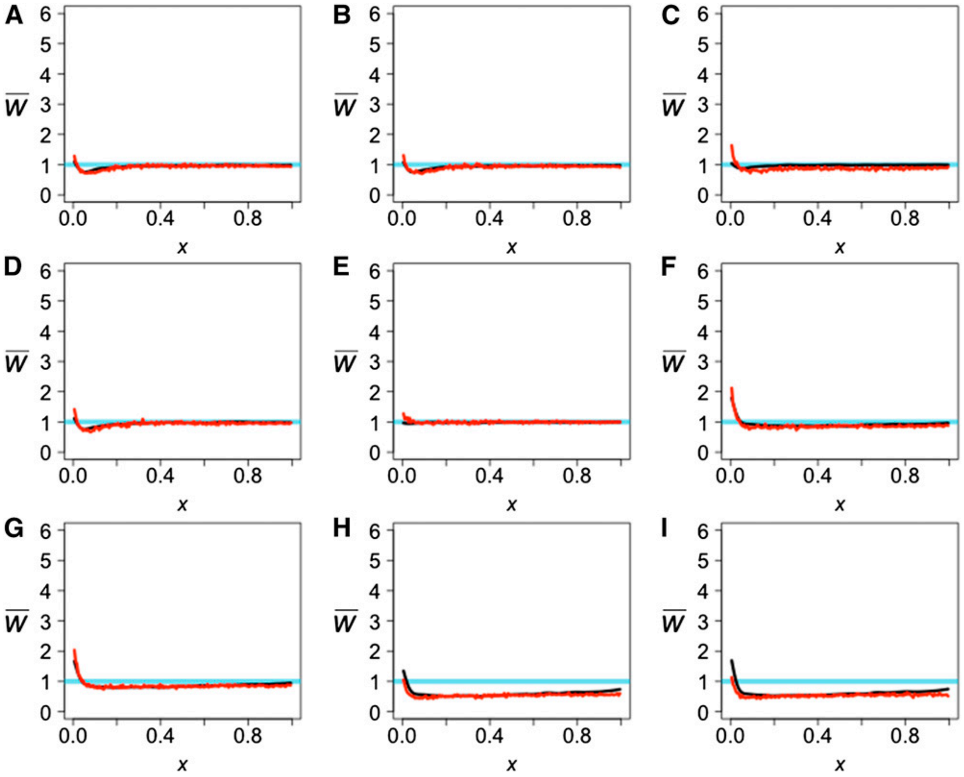


Figure 11 The observed projection (black line) and simulated projection from our model (red line) for the YRI reference population. The test genomes are Han (A), Papuan (B), Dinka (C), French (D), Mandenka (E), Mbuti (F), San (G), Denisovan (H), and Neanderthal (I). The LSS scores comparing the observed projections to each other and the expectation can be found in Table S3, and the LSS scores comparing the observed and simulated projections can be found in Table 2.

Table 2 LSS comparing the simulated projection from our model (Figure 7) to the observed projections (Figure 9, Figure 10, and Figure 11)

	Reference		
Test	CEU	CHB	YRI
French	*	2.12	0.34
Han	0.54	*	0.37
Papuan	1.00	2.31	2.91
Dinka	2.03	4.18	0.45
Yoruba	1.50	4.12	*
Mandenka	1.59	4.32	0.36
Mbuti	1.42	2.67	0.73
San	0.92	1.98	0.48
Denisovan	3.15	1.31	1.33
Neanderthal	4.98	2.68	2.30

*No simulated projection to compare to for LSS

results similar to the observed projections (Table 2 and Figure 9, C–G). A closer look at the middle of the projection for reference CEU shows that the San projection is slightly lower than the Yoruba projection (Figure 9, D and G), which suggests that the difference in divergence time is weakly reflected in the projection.

The projections of different African genomes (Dinka, Mandenka, Mbuti, San) onto YRI (Figure 11, C and E–G) illuminate the relationship between these four African populations and the Yoruba. Other studies (Tishkoff *et al.* 2009; Meyer *et al.* 2012; Prüfer *et al.* 2014) have shown that, while the San and Mbuti are the most diverged from all other populations sampled, the Mandenka and Yoruba populations have only recently separated, and the Dinka population shares some ancestry with non-African populations. The San and Mbuti projections onto YRI show a slight excess of rare alleles, suggesting some admixture from their ancestors into the ancestors of YRI. The Mbuti is closer to the $\overline{w}(x) = 1$ line, which suggests that it is less diverged from YRI than is the San, agreeing with the model proposed in other studies (Tishkoff *et al.* 2009; Meyer *et al.* 2012; Prüfer *et al.* 2014). The Mandenka projection falls nearly on the $\overline{w}(x) = 1$ line,

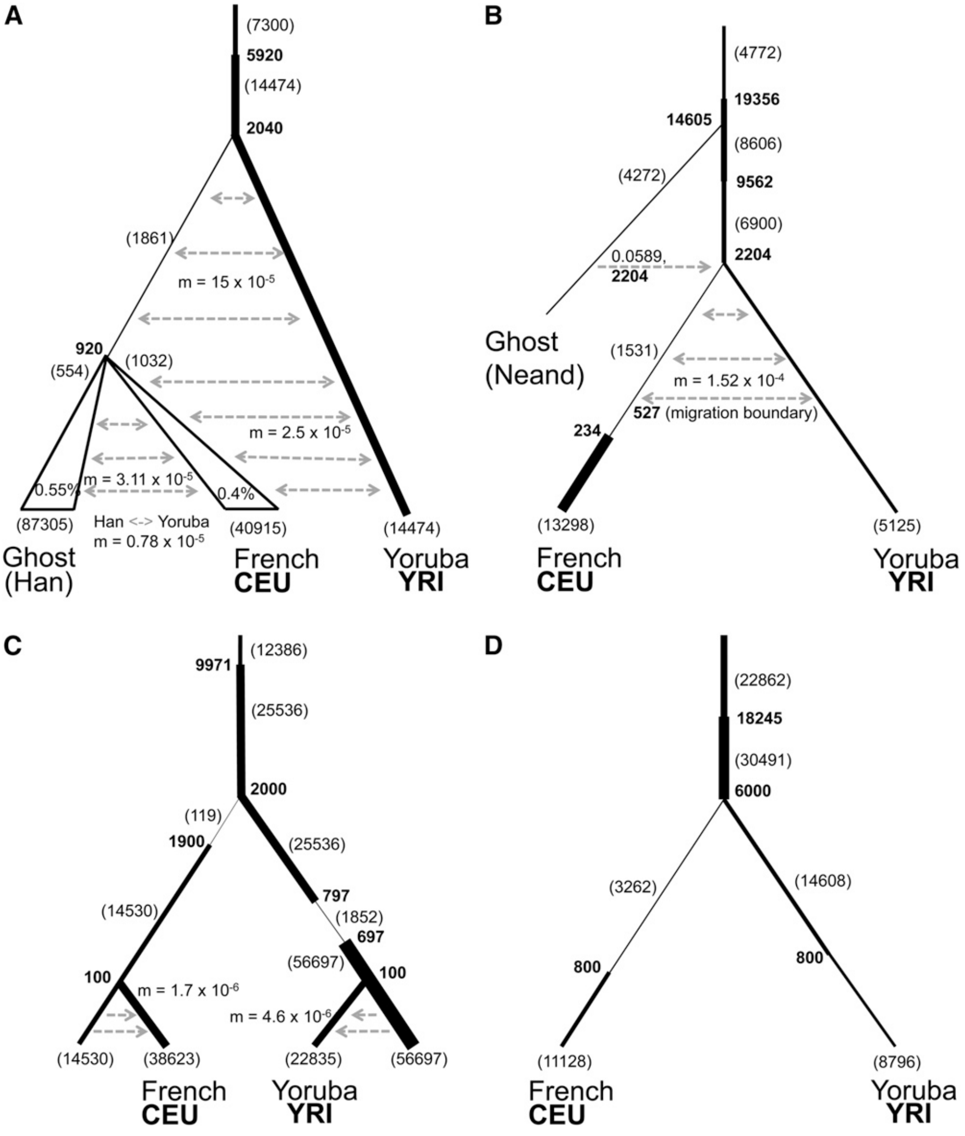


Figure 12 The demographic models from each of the four previous studies: Gravel *et al.* (2011) (model A), Harris and Nielsen (2013) (model B), Excoffier *et al.* (2013) (model C), and Schiffels and Durbin (2014) (model D). Shading and symbols have the same meaning as in Figure 7, and the triangle indicates growth at the given percentage.

Table 3 LSS comparing the simulated projections for the best estimates from four previous studies (models A–D) and the modified estimates from four previous studies (models A*–D*) to the observed projections

	Test/Reference	
Model	Yoruba/CEU	French/YRI
A	1.26	0.23
B	5.55	5.88
C	15.45	0.74
D	13.91	7.32
A*	0.64	0.24
B*	0.93	0.14
C*	2.24	0.68
D*	3.17	1.20

suggesting that it is indistinguishable from a random Yoruba individual. Finally, the Dinka projection onto YRI exhibits a dip that is similar, although of reduced magnitude, to those observed in all the non-African projections, perhaps due to greater admixture between the ancestors of the Dinka and Yoruba in Africa. Including these events in the model (Figure 7) gives a close fit to the observed projections (Table 2 and Figure 11, C and E–G).

Test of Published Models

We used observed projections to test for consistency with inferred demographic parameters from four studies (Gravel *et al.* 2011; Excoffier *et al.* 2013; Harris and Nielsen 2013; Schiffels and Durbin 2014) for European and Yoruba populations. All four studies applied their methods to these two populations.

We obtained projections by using *fastsimcoal2* (Excoffier *et al.* 2013) to simulate 1 million SNPs with the estimated demographic parameters from each of these four models. The demographic parameters used are shown in Figure 12. We compare the simulated projections to the observed projections of a Yoruba genome projected onto CEU and of a French genome projected onto YRI. The visual differences highlight aspects of each model that agree or disagree with the observed projections.

The four models overlap but differ in the estimates of a number of parameters. All models assume a population decrease in ancestral Europeans, presumably during dispersal out of Africa. The severity of the population size change ranges from 0.0047 (model C) to 0.22 (model B) and occurs at the time when the ancestors of the Yoruba and European populations diverged. Models A, B, and D assume a subsequent population expansion, while model C, which has the most extreme reduction, recovers 100 generations after the population decrease. In model A, the Yoruba population is assumed to be of constant size while the size declines in models B and D. In model C, the ancestral Yoruba population underwent a bottleneck 797 generations ago. In all four models, the population ancestral to Europeans and Yoruba increases in size before the two populations separated. In models A–C, the time of divergence of Europeans and Yoruba is ~50 KYA. In model D, the separation time is at least 150 KYA.

Model A assumes higher rates of migration soon after the European and Yoruba divergence and a lower rate more recently. Model B allows for migration between these two populations, and it also includes a parameter for ghost admixture from an archaic hominin that diverged 14,605 generations ago. Model C uses a continent-island model, in which Europeans and Yoruba diverged from continental European and African populations recently, receiving migrants from those populations until the present. However, neither they nor their ancestral populations admix with each other. Model D does not allow for migration between the two populations, although Schiffels and Durbin (2014) say that such migration probably occurred.

The simulated projections show that model A gives the best fit to the observed projections (Table 3 and Figure 13). For model A, increasing the rate of recent migration from

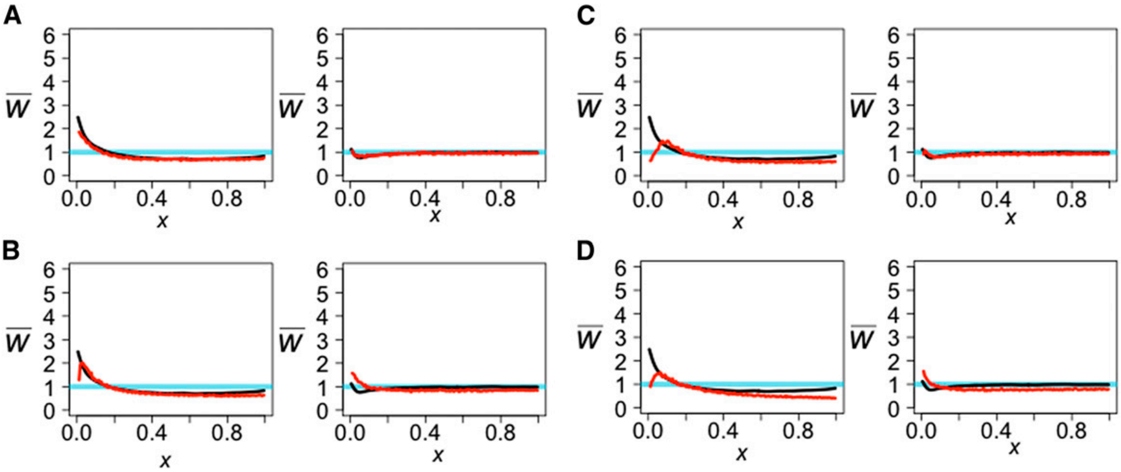


Figure 13 The observed projections (black line) and simulated projections from demographic models inferred from other studies (red line). For each model A–D in Figure 12, the left projection is the Yoruba genome projected onto CEU and the right projection is the French genome projected onto YRI. LSS scores are in Table 3.

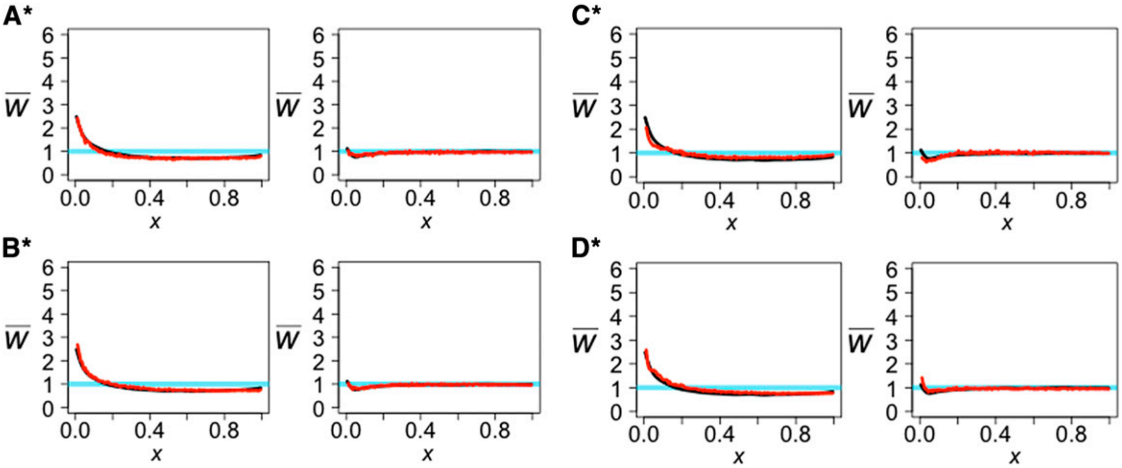


Figure 14 Projections for previous studies (models A–D) where the parameters for migration or admixture between Europeans and Yorubans have been added or modified for a better fit. For each model A*–D*, the left projection is the Yoruba genome projected onto CEU and the right projection is the French genome projected onto YRI. LSS scores are in Table 3.

Yoruba to Europeans from 0.000025 migrants/generation to 0.00005 migrants/generation led to a slightly better fit (Table 3 and Figure 14). In model B, increasing the migration rate from Europeans to Yoruba to 0.00083 migrants/generation and adding admixture 150 generations ago at a rate of 0.02 from Europeans to Yoruba and a reverse rate of 0.015 resulted in a better fit. In model C, adding admixture at two different times led to a better fit. We first added recent admixture at a rate of 0.07, 150 generations ago from Europeans to Yoruba with a reverse rate of 0.1. Then, we added ancestral admixture at a rate of 0.37 from Europeans to Yoruba and a reverse rate of 0.2, 1710 generations ago. In model D, adding symmetric admixture of 0.01, 150 generations ago between Yoruba and Europeans, and allowing for migration beginning at 1662 generations ago of 0.0007 migrants/generation from Europeans to Yoruba and 0.0003 migrants/generation from Yoruba to Europeans results in a better fit (models A*–D*, Table 3; Figure 14).

Our projection analysis supports the hypothesis that there was significant gene flow between the ancestors of Europeans and Yoruba after there was introgression from Neanderthals into Europeans. Adding or modifying gene flow in models A–D substantially improved the fits to the observed projections.

Discussion and Conclusions

We have introduced projection analysis as a visual way of comparing a single genomic sequence with one or more reference populations. The projection summarizes information from the joint site-frequency spectrum of two populations. We have shown that projections are affected by various demographic events, particularly population size changes in the reference population and admixture into the reference population. The time since two populations had a common ancestor also affects the projection, as does the interaction with unsampled populations.

Projection analysis is primarily a visual tool and is not intended to replace methods that estimate model parameters such as those developed by Gutenkunst *et al.* (2009), Harris and Nielsen (2013), Excoffier *et al.* (2013), and Schiffels and Durbin (2014). Projection analysis uses less information than these methods. Instead, projection analysis is intended to be a method of exploratory data analysis. It provides a way to compare a single genomic sequence, perhaps of unknown provenance, with several reference populations, and it provides a way to test the consistency of hypotheses generated by other means.

Our applications of projection analysis to human and archaic hominin populations largely confirmed conclusions from previous studies. In particular, we support the hypothesis that Neanderthals admixed with the ancestors of Europeans and Han Chinese and the hypothesis that Neanderthals and Denisovans are sister groups.

By analyzing present-day human populations, we provide strong support for the conclusion of Gutenkunst *et al.* (2009) and Gravel *et al.* (2011) that there was continuing gene flow between the ancestors of Yoruba and the ancestors of Europeans long after their initial separation. The fit of other models improves when such gene flow is included.

Harris and Nielsen (2013) incorporate migration in their model, but they assume a small amount from the time of separation until a few thousand years ago. The Excoffier *et al.* (2013) model does provide a good fit for the French projection onto YRI, perhaps because of the large bottleneck that they infer in the ancestral Yoruba, but the Yoruba projection onto CEU requires some admixture for a better fit. The Schiffels and Durbin (2014) model does not allow for estimation of migration parameters. However, they argue that there was probably an initial divergence with subsequent migration before a full separation. Our conclusion is consistent with theirs. There was likely substantial gene flow between the ancestors of Europeans and Yoruba after their initial separation but before movement out of Africa. Then, stronger geographic barriers led to lower rates of gene flow and effectively complete isolation.

Throughout we have assumed that population history can be represented by a phylogenetic tree. Although that assumption is convenient and is made in most other studies as well, we recognize that a population tree may not be a good representation of the actual history. For example, the inferred period of gene flow between Europeans and Yoruba may actually reflect a complex pattern of isolation by distance combined with the appearance and disappearance of geographic barriers to gene flow. At this point, introducing a more complex model with more parameters will not help because there is insufficient power to estimate those parameters or to distinguish among several plausible historical scenarios.

The effect of ancestral misidentification on projection analysis was also a concern. We show that low levels of ancestral misidentification lead to an increase in common alleles. Thus, we expect and do see a slight increase of $\overline{w}(x)$ in common alleles in most observed projections.

Projection analysis is designed for analyzing whole-genome sequences, but it can be applied to other data sets including partial genomic sequences, dense sets of SNPs, and whole-exome sequences. However, ascertainment of SNPs could create a problem by reducing the sample sizes of low- and high-frequency alleles. Of course, the smaller the number of segregating sites in the reference genome, the larger will be the sampling error in the projection. The number of samples from the reference population also affects the utility of the projection. As we have shown, an important feature of many projections is the dependence of $\overline{w}(x)$ on small x. Relatively large samples from the reference population (≥50 individuals) are needed to see that dependence clearly. When sufficiently large samples are available, projection analysis provides a convenient way to summarize the joint site-frequency spectra of multiple populations and to compare observations with expectations from various models of population history.

Acknowledgments

We thank N. Patterson, B. Peter, F. Racimo, D. R. Reich, and J. G. Schraiber for helpful discussions of this topic and for comments on previous versions of this paper. M.A.Y. was supported in part by a National Institutes of Health (NIH) National Research Service Award Traineeship (T32 HG 00047) and a National Science Foundation Graduate Research Fellowship. K.H. was supported in part by an NIH grant (IR01GM109454-01 to Rasmus Nielsen, Yun Song, and Steve Evans) and a National Science Foundation Graduate Research Fellowship. M.S. was supported in part by NIH grant R01-GM40282.

Literature Cited

Beerli, P., 2004 Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. Mol. Ecol. 13: 827–836.

Chen, H., R. E. Green, S. Pääbo, and M. Slatkin, 2007 The joint allele-frequency spectrum in closely related species. Genetics 177: 387–398.

Excoffier, L., I. Dupanloup, E. Huerta-Sanchez, V. C. Sousa, and M. Foll, 2013 Robust demographic inference from genomic and SNP data. PLoS Genet. 9: e1003905.

1000 Genomes Project Consortium, G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks *et al.*, 2010 A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.

Gravel, S., B. M. Henn, R. N. Gutenkunst, A. R. Indap, and G. T. Marth *et al.*, 2011 Demographic history and rare allele sharing among human populations. Proc. Natl. Acad. Sci. USA 108: 11983–11988.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, 2009 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5: e1000695.

Harris, K., and R. Nielsen, 2013 Inferring demographic history from a spectrum of shared haplotype lengths. PLoS Genet. 9: e1003521.

Hernandez, R. D., S. H. Williamson, and C. D. Bustamante, 2007 Context dependence, ancestral misidentification, and spurious signatures of natural selection. Mol. Biol. Evol. 24: 1782–1800.

Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.

Lachance, J., B. Vernot, C. C. Elbers, B. Ferwerda, A. Froment *et al.*, 2012 Evolutionary history and adaptation from high-coverage whole genome sequences of diverse African hunter-gatherers. Cell 150: 457–469.

Li, H., and R. Durbin, 2011 Inference of human population history from individual whole-genome sequences. Nature 475: 493–496.

Meyer, M., M. Kircher, M.-T. Gansauge, H. Li, F. Racimo *et al.*, 2012 A high-coverage genome sequence from an archaic Denisovan individual. Science 338: 222–226.

Morales, J. L., and J. Nocedal, 2011 L-BFGS-B: remark on Algorithm 778: L-BFGS-B, FORTRAN routines for large scale bound constrained optimization. ACM Trans. Math. Softw. 38: 1.

Pickrell, J. K., N. Patterson, C. Barbieri, F. Berthold, L. Gerlach *et al.*, 2012 The genetic prehistory of southern Africa. Nat. Commun. 3: 1143.

Prüfer, K., F. Racimo, N. Patterson, F. Jay, S. Sankararaman *et al.*, 2014 The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505: 43–49.

Schiffels, S., and R. Durbin, 2014 Inferring human population size and separation history from multiple genome sequences. Nat. Genet. 46: 919–925.

Tishkoff, S. A., F. A. Reed, F. R. Friedlander, C. Ehret, A. Ranciaro *et al.*, 2009 The genetic structure and history of Africans and African Americans. Science 324: 1035–1044.

Wollstein, A., O. Lao, C. Becker, S. Brauer, R. J. Trent *et al.*, 2010 Demographic history of Oceania inferred from genome-wide data. Curr. Biol. 20: 1983–1992.

Communicating editor: M. W. Hahn

Appendix

The Projection of a Test Genome Onto a Reference Population and Applications to Humans and Archaic Hominids

The aim of this appendix is to present a theoretical justification for the "dip" at low frequencies observed in Figure 8, which shows the French genome projected onto the reference CEU population, the Han Chinese genome projected onto the reference CHB population, and the Yoruba genome projected onto the reference YRI population. In each case, the test genome appears to carry fewer of the reference population's derived singletons and doubletons than expected given the close relationship between the test and reference genomes. We argue that this is a consequence of finite reference population size in a species that has inflated counts of low frequency alleles due to recent population growth.

Each comparison in Figure 8 is akin to the scenario of starting with a reference population of $N + 1$ genomes, picking one genome uniformly at random, and projecting this "test" genome onto the remaining N-genome panel. If we fix a frequency x and let N go to infinity, it is trivial to see that the projection $\overline{w}(x)$ should approach x. However, this does not imply that $\overline{\varpi}(k/N)/(k/N)$ should approach 1 as N goes to infinity with k fixed.

We can compute the expected value of $\overline{w}(k/N)$ in terms of the frequency spectrum $(x_1, x_2, \ldots, x_{N+1})$ of the entire population sample, where x_1 is the frequency of singletons, x_2 is the frequency of doubletons, and so on. In terms of these frequencies, $\overline{\varpi}(k/N)$ has the following expected value:

$$\mathbb{E}(\overline{w}(k/N)) = \frac{1}{k/N} \cdot \frac{x_{k+1} \cdot \frac{k+1}{N+1}}{x_{k+1} \cdot \frac{k+1}{N+1} + x_k \cdot \frac{N+1-k}{N+1}} = \frac{x_{k+1} \cdot (k+1)}{x_k \cdot k} + O(k/N)$$

Here, the factor $(k + 1)/(N + 1)$ is the probability that the test individual has the derived allele given that $k + 1$ out of the $N + 1$ members of the reference population have the derived allele. Likewise, $(N + 1 - k)/(N + 1)$ is the probability that the test individual has the ancestral allele given that k out of $N + 1$ members of the reference population have the derived allele. This implies $\mathbb{E}(\overline{\varpi}(k/N)) = k/N$ if and only if

$$x_{k+1}/x_k = k/(k+1). \tag{1}$$

In a panmictic population that has reached effective population size equilibrium, coalescent theory does predict that $x_{k+1}/x_k = k/(k + 1)$. However, the site frequency spectrum is so sensitive to past changes in effective population size that equation (1) does not often hold for real datasets, and in general, low frequency variants show the most deviation from (1). In addition, some 1000 Genomes reference population "singletons" may be sequencing errors that have a very low probability of being observed in a test genome because they are not true segregating genetic variants. Somatic cell line mutations are similarly unlikely to be shared. Cryptic population structure may be another source of deviation from the $\overline{\varpi} = 1$ expectation at low allele frequencies.

Let W_k denote the quantity $x_{k+1} \cdot (k + 1)/(x_k \cdot k)$. Table 1 lists values of W_k for the CEU, CHB, and YRI reference populations from the 1000 Genomes Project. letting k range from 1 to 9. Assuming that the panel contains no sequencing errors, W_k is the expected value of $\overline{\varpi}(k)$ for the projection when the test genome is a member of the reference population. Both the CEU and CHB reference populations have W_k values that are less than 1 for $k < 5$ as a result of recent population growth, explaining the pronounced dip we see in these projections. In contrast, the YRI panel does not contain excess low frequency variants, suggesting that the smaller dip at $k = 1$ seen in the Yoruba projection may result from other causes such as sequencing error or structure in the reference population.

Table 1 Expected projection values $W_k = x_{k+1} \cdot (k + 1)/(x_k \cdot k)$ for small values of k in the three 1000 Genomes reference populations

Panel	k = 1	2	3	4	5	6	7	8	9
CEU	0.743	0.902	0.964	0.982	0.986	1.004	1.014	1.015	1.006
CHB	0.64	0.805	0.942	0.968	1.009	0.988	1.005	0.999	1.028
YRI	1.219	0.996	0.995	0.991	0.991	0.983	0.985	0.982	1.001

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.112.145359/-/DC1

The Projection of a Test Genome onto a Reference Population and Applications to Humans and Archaic Hominins

Melinda A. Yang, Kelley Harris, and Montgomery Slatkin

DOI: 10.1534/genetics.112.145359

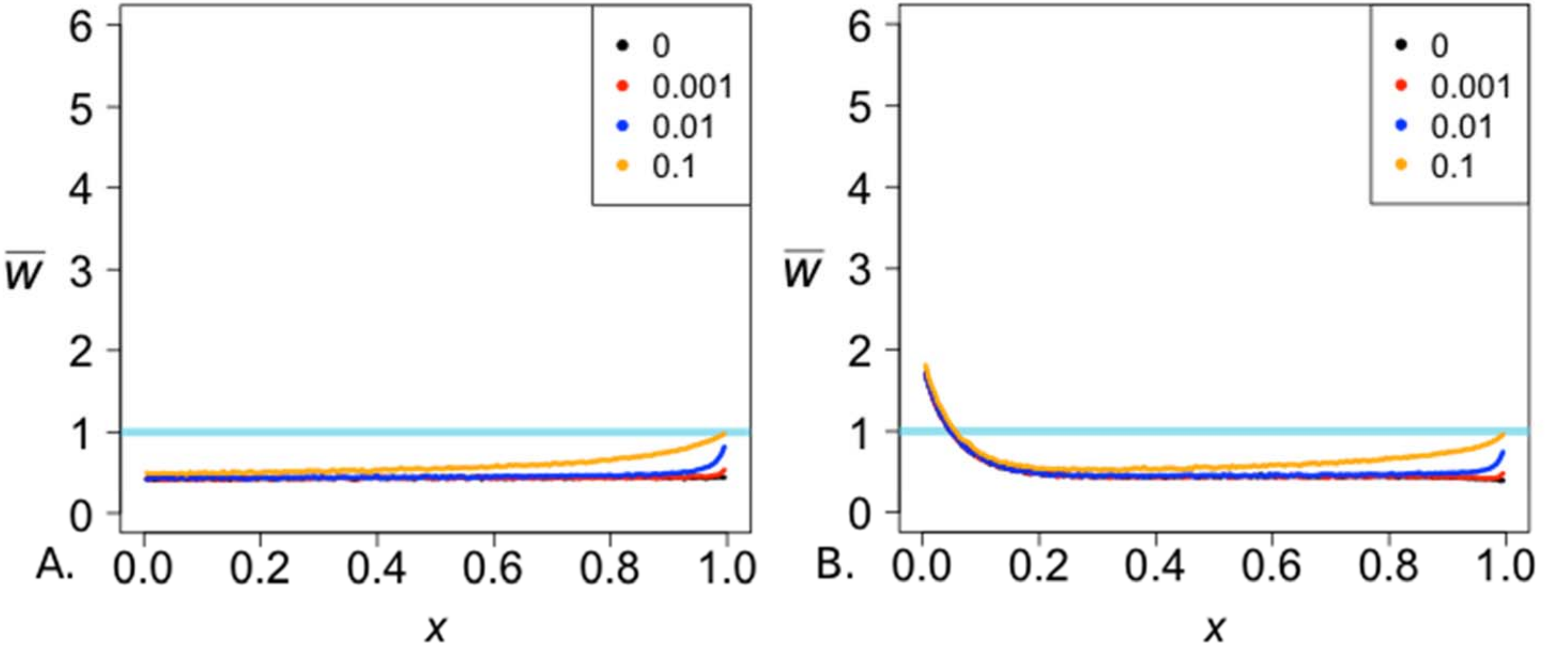


Figure S1 Simulated projections with ancestral misidentification. The proportion of sites misidentified is given in the legend. For both models, the population is a constant size of 10,000 and the time of divergence is 400 kya. In part B, there is an additional admixture event of 0.02 from the test to the reference population at 50 kya.

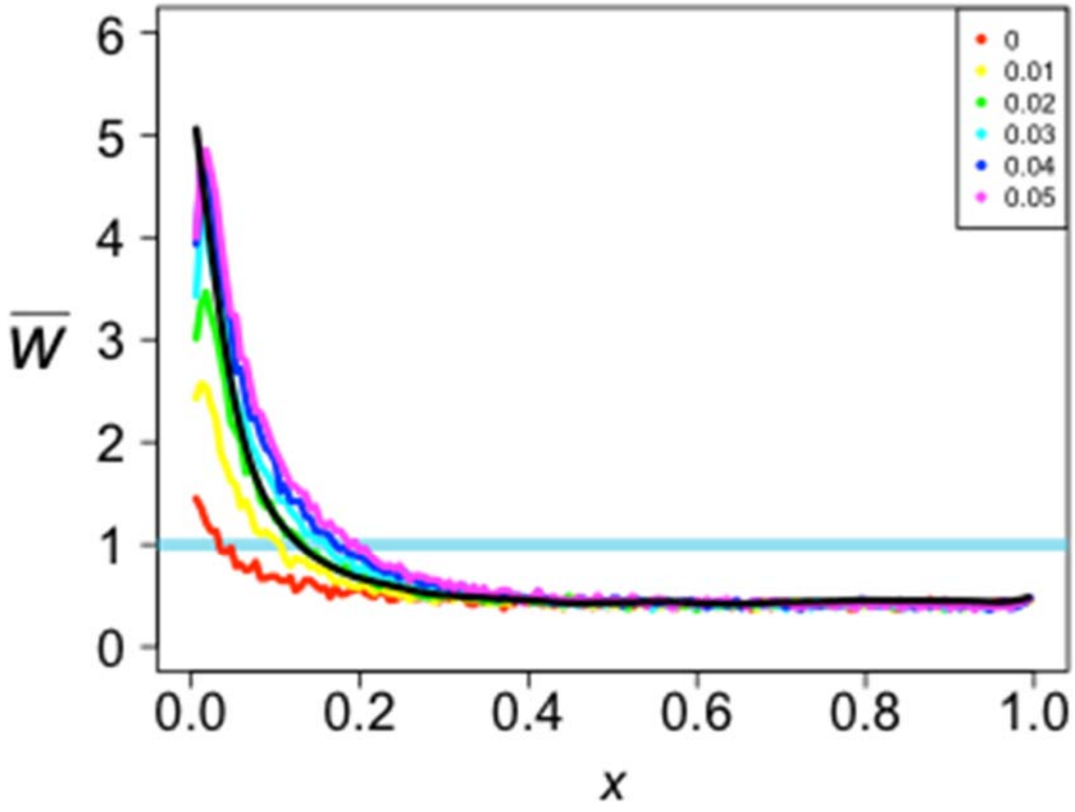


Figure S2 The simulated projections for reference CEU and test Neanderthal for our model when altering the amount of admixture ($f_{NEA\text{-}ANC1}$). The black line is the observed projection.

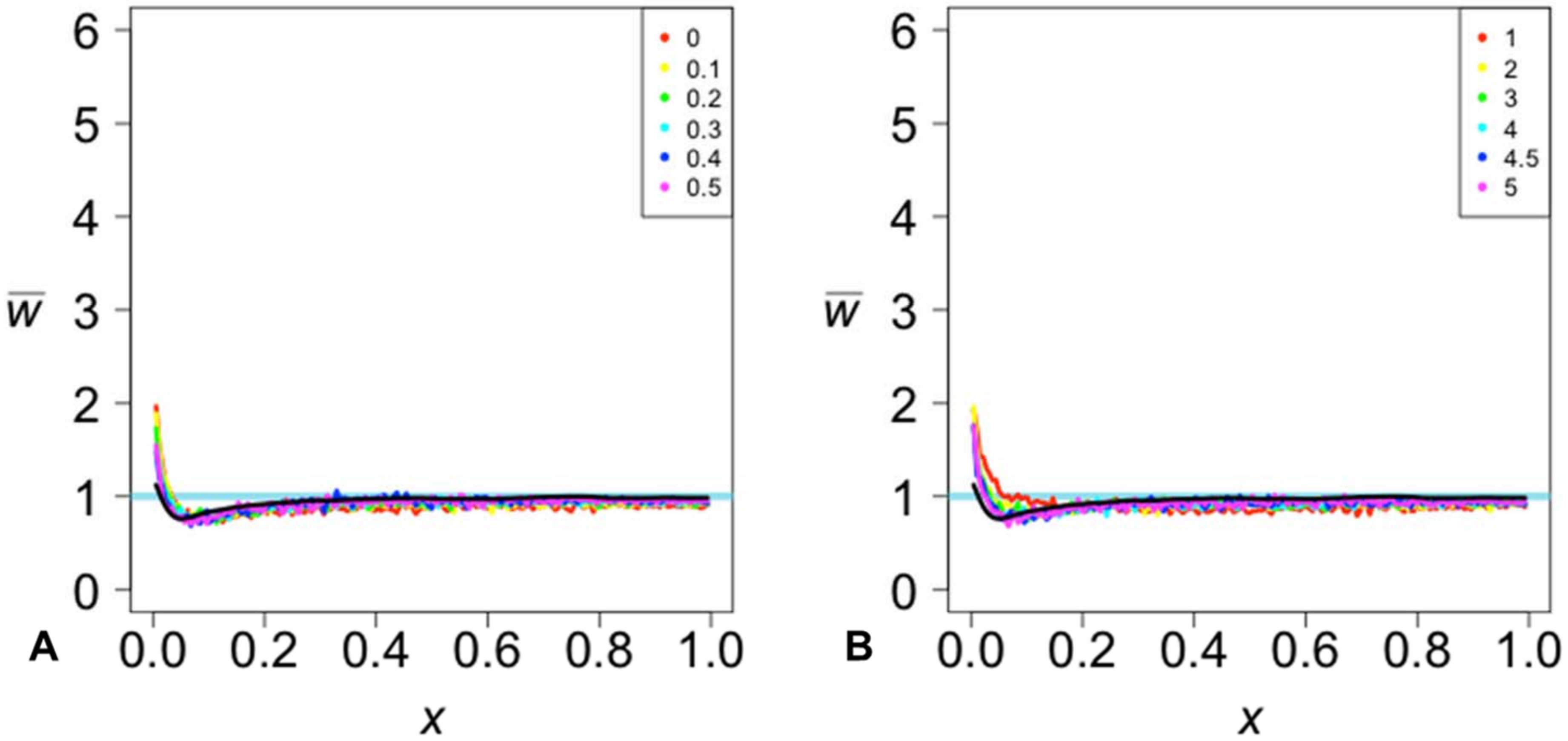
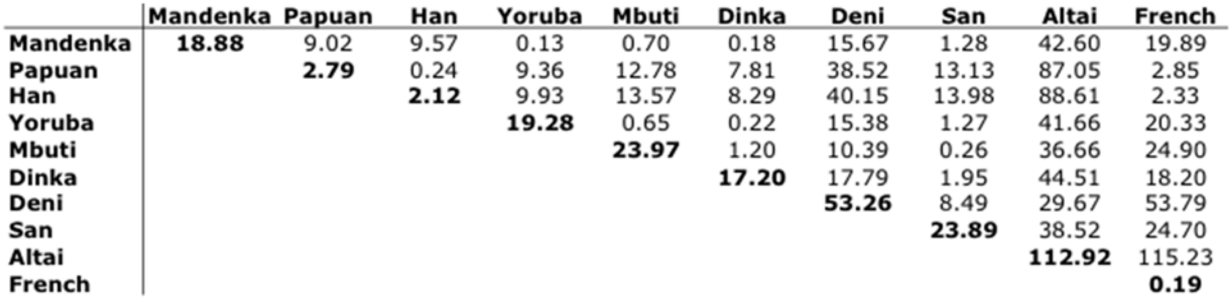


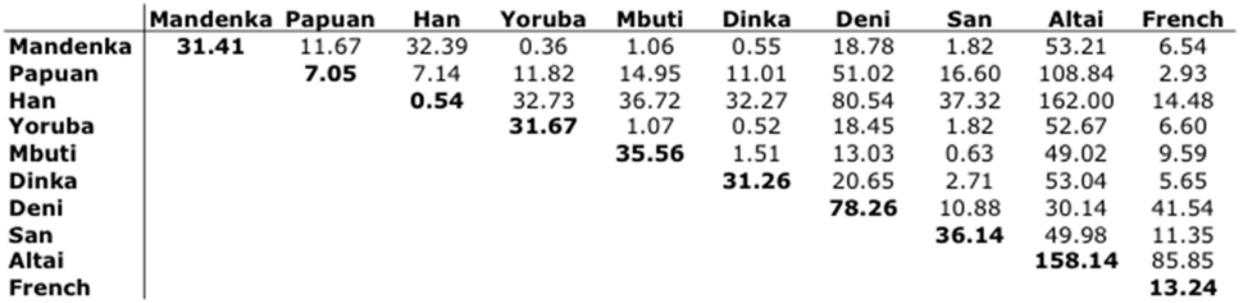
Figure S3 The simulated projections for reference YRI and test French for our model when altering (A) the population increase in the Yoruba population backwards in time (N_{ANC4}/N_{YOR}) and (B) the amount of admixture from Europeans to Yorubans ($f_{ANC1\text{-}ANC4}$). The black line is the observed projection.

Table S1 LSS comparing the observed projections for reference CEU to each other. The diagonals compare that test genome to the $\overline{w}(x)=1$ line.



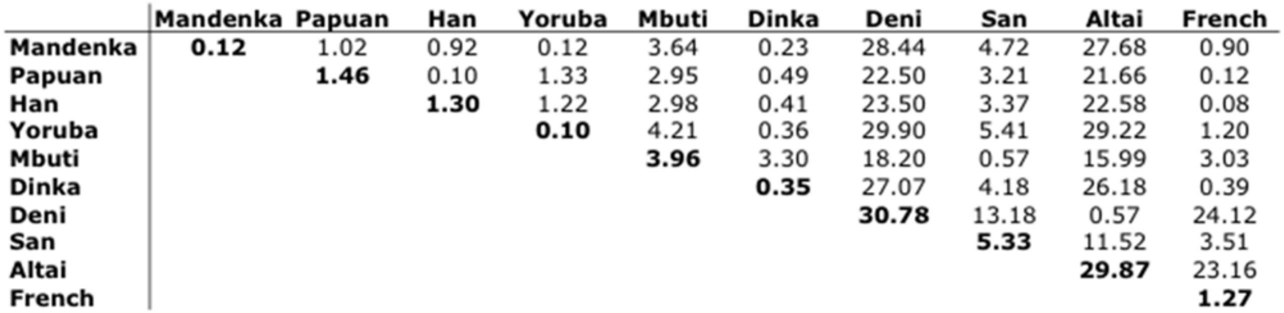
	Mandenka	Papuan	Han	Yoruba	Mbuti	Dinka	Deni	San	Altai	French
Mandenka	**18.88**	9.02	9.57	0.13	0.70	0.18	15.67	1.28	42.60	19.89
Papuan		**2.79**	0.24	9.36	12.78	7.81	38.52	13.13	87.05	2.85
Han			**2.12**	9.93	13.57	8.29	40.15	13.98	88.61	2.33
Yoruba				**19.28**	0.65	0.22	15.38	1.27	41.66	20.33
Mbuti					**23.97**	1.20	10.39	0.26	36.66	24.90
Dinka						**17.20**	17.79	1.95	44.51	18.20
Deni							**53.26**	8.49	29.67	53.79
San								**23.89**	38.52	24.70
Altai									**112.92**	115.23
French										**0.19**

Table S2 LSS comparing the observed projections for reference CHB to each other. The diagonals compare that test genome to the $\overline{w}(x)=1$ line.



	Mandenka	Papuan	Han	Yoruba	Mbuti	Dinka	Deni	San	Altai	French
Mandenka	**31.41**	11.67	32.39	0.36	1.06	0.55	18.78	1.82	53.21	6.54
Papuan		**7.05**	7.14	11.82	14.95	11.01	51.02	16.60	108.84	2.93
Han			**0.54**	32.73	36.72	32.27	80.54	37.32	162.00	14.48
Yoruba				**31.67**	1.07	0.52	18.45	1.82	52.67	6.60
Mbuti					**35.56**	1.51	13.03	0.63	49.02	9.59
Dinka						**31.26**	20.65	2.71	53.04	5.65
Deni							**78.26**	10.88	30.14	41.54
San								**36.14**	49.98	11.35
Altai									**158.14**	85.85
French										**13.24**

Table S3 LSS comparing the observed projections for reference YRI to each other. The diagonals compare that test genome to the $\bar{w}(x)=1$ line.



	Mandenka	Papuan	Han	Yoruba	Mbuti	Dinka	Deni	San	Altai	French
Mandenka	**0.12**	1.02	0.92	0.12	3.64	0.23	28.44	4.72	27.68	0.90
Papuan		**1.46**	0.10	1.33	2.95	0.49	22.50	3.21	21.66	0.12
Han			**1.30**	1.22	2.98	0.41	23.50	3.37	22.58	0.08
Yoruba				**0.10**	4.21	0.36	29.90	5.41	29.22	1.20
Mbuti					**3.96**	3.30	18.20	0.57	15.99	3.03
Dinka						**0.35**	27.07	4.18	26.18	0.39
Deni							**30.78**	13.18	0.57	24.12
San								**5.33**	11.52	3.51
Altai									**29.87**	23.16
French										**1.27**

Table S4 LSS (Least Sum of Squares) comparing our model to the observed projections, altering the amount of admixture from Neanderthals to non-Africans ($f_{NEA\text{-}ANC1}$).

Test Neanderthal			
$f_{NEA\text{-}ANC1}$	**CEU**	**CHB**	**YRI**
0	69.51	109.74	3.32
0.01	25.02	41.76	2.51
0.02	7.88	11.34	2.11
0.03	**4.33**	**2.74**	1.34
0.04	7.03	3.91	1.19
0.05	13.34	8.79	**0.95**

Table S5 LSS comparing our model to the observed projections, altering the population increase in the Yoruba population backwards in time (*N_{ANC4}/N_{YOR}*) and the amount of admixture (*$f_{ANC1\text{-}ANC4}$*) from Europeans to Yorubans.

Test French			
N_{ANC4}/N_{YOR}	**YRI**	*$f_{ANC1\text{-}ANC4}$*	**YRI**
1	4.44	0	2.72
2	2.23	0.1	1.89
3	1.51	0.2	1.16
4	1.31	0.3	0.75
4.5	**1.21**	0.4	**0.56**
5	1.32	0.5	0.59