The projection of a test genome onto a reference population and applications to humans and archaic hominins

Melinda A. Yang and Montgomery Slatkin

Corresponding author
Montgomery Slatkin
Department of Integrative Biology
University of California
Berkeley, CA 94720-3140

slatkin@berkeley.edu

Key words: Frequency spectrum, archaic admixture, human population genetics, human demography

Running head: Projection analysis

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, $\bar{w}(x)$. The weight is assigned in such a way that if the test genome is a random sample from the reference population, $\bar{w}(x) = 1$. 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 projection of several human and two archaic genomes onto three reference populations from the 1000 Genomes project, Europeans (CEU), Han Chinese (CHB) and Yoruba (YRI) 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.

Introduction

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 (PSMC) 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 record 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. The applications of these methods have led to a relatively clear picture of population splitting, past bottlenecks in population size and episodes of admixture after separation.

In this paper, we introduce another way to characterize genomic data. Our method is designed to indicate the past relationship between a single 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 the occurrence of bottlenecks in population size. 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 in order 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 there is an outgroup that allows determination of the derived allele frequency, x, at every segregating site in the reference population. We next 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 fixed ancestral, w=0; if it is heterozygous, w=1/(2x); and if it is fixed derived w=1/x. The projection is the average weight of sites in the test genome at which the frequency of the derived allele in the reference population is x, $\overline{w}(x)$.

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 so 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, the numerical solutions for each set of parameter values were obtained from *dadi* (GUTENKUNST *et al*. 2009). All

models with more than three populations were simulated using *fastsimcoal2* (EXCOFFIER *et al*. 2013).

For all models, an ancestral effective population size (N_e) of 10000 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 captions.

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, a single pulse of admixture in which a fraction *f* of one population is replaced by immigrants from the other, and a prolonged period of gene flow during which a fraction *m* of the individuals in one population are replaced each generation by immigrants from the other. We allowed 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 results in the pattern shown: $\bar{w}(x)$ decreases monotonically to the value expected in the absence of gene flow. Even very slight amounts of gene flow in this direction create 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 a small amount of gene flow will carry some alleles that were new mutations in the

test population. Those alleles will necessarily be in low frequency in the reference population 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.

A bottleneck in the reference (Fig. 2B) or ancestral population (Fig. 2C) affects the projection but a bottleneck in only the test population does not (Fig. 2A). The reason for the humped shape of the projection 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.

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 the decrease at the beginning (Fig. 3B).

Three populations

Three populations lead to a greater variety of effects than can be seen in two. Because only two populations are sampled, 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 test for the presence of the ghost population.

We first consider the effects of gene flow alone. We will assume a single pulse of admixture of strength f at a time t_{GF}. There are three distinct topologies representing the ancestry of the three populations (Fig. 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 in all three topologies (Fig. 5A-C). How ghost gene flow into the reference population affects the projection depends on the population relationships. If the test and ghost populations are sister groups (Fig. 4A, 5D), then the effect is similar to that of gene flow directly from the test into the reference population (Fig. 1). When gene flow is directly from the test population, the higher values of $\overline{w}(x)$ for small x reflect new mutations that arose in the test population after separation from the reference. When gene flow is from the ghost population, only mutations that arose in the internal branch contribute. Thus, when there is a longer period of shared ancestry between the test and ghost populations, there is a stronger effect due to admixture (Fig. 5D).

In the second topology (Fig. 4B), the reference and ghost populations are sister groups. Here, gene flow from the ghost population has the opposite effect on the projection. It reduces $\overline{w}(x)$ for small x. The intuitive reason is that some alleles carried from the ghost to the reference population arose as new mutations in the ghost population and hence cannot be in the test population. Therefore, there are fewer low frequency alleles in the test genome than expected. For a given time and magnitude of admixture, the effect increases as the time of separation of the ghost and reference populations increases (Fig. 5E). When the reference and test populations are sister groups (Fig. 4C),

and the ghost population is an outgroup, a dip is observed for low frequencies and a slight increase is observed for common alleles (Fig. 5F).

If there is a bottleneck in the reference population after admixture, the effect is similar to that seen in the two-population case (Fig. 6A). The signal of admixture is amplified. In the case where the reference and ghost populations are sister groups (Fig. 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 increasing 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 (Fig. 6C), and the increase in common alleles is still observed.

Application to humans and archaic hominins

Projection analysis is intended to be applicable to a wide variety of situations. We will illustrate its use by applying it to genomic data from present-day humans and two archaic hominins (a Neanderthal and a Denisovan). For the reference populations, we used data from the 1000 Genomes (1000G) project for three reference populations, Europeans (CEU), Han Chinese (CHB) and Yoruba (YRI) (1000 GENOMES 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 names used by Meyer et al. (2012). 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 constructed a realistic demographic history using *dadi*, *fastsimcoal2*, and various curve fitting algorithms (Fig. 7). For two-population models we obtained parameters estimates using the limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm and then calculated the projections using *dadi*. Additional populations were added, and two to three variables at a time were optimized using Brent's algorithm and simulated projections in *fastsimcoal2*. The parameters were chosen by combining results from previous studies of human demographic history (Gutenkunst et al. 2009, Harris and Nielsen 2013, and Excoffier et al. 2013). All projections from this model were obtained using *fastsimcoal2*. We calculated the sum of least squares (LSS) between the simulated observed projections to allow quantitative comparison of the fit to different sets of parameters.

We arrived at a realistic set of demographic parameters (Fig. 7) that produce projections similar to those found from analyzing the human genomic data (Fig. 9-11, black curve = observed projection, red curve = simulated projection). This model is not meant to be optimal; instead it is intended to show that, for a plausible demographic scenario, the projections are close to projections computed from the data. This model illustrates the sensitivity of projections to important 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 Fig. 9-11 represent the observed projections. Tables 1-3 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 $\bar{w}(x) = 1$ line. The observed projections show that the Neanderthal and Denisovan projections onto CEU, CHB, and YRI look the most different from the $\bar{w}(x) = 1$ line. The test genome belonging to each reference population has the lowest LSS score for that reference. The African genomes show similar projections onto CEU and CHB. For projections onto YRI, Mandenka is similar to Yoruba, the San and Mbuti are similar to each other, and the Dinka projections more similar to those of non-Africans.

Comparison of a test genome with the same population

In Fig. 8A, the projection of the French genome onto CEU fits the expectation except for a small x. Similar patterns are seen in Fig. 8B in the projection of the Han genome onto CHB and in Fig. 8C in the projection of the Yoruba genome onto YRI. This pattern is expected for the smallest frequency classes because of the finite sample size of reference populations (Appendix).

Admixture with Neanderthals and Denisovans

Our simulations show that a bottleneck combined with admixture into the reference population could result in drastic changes to the projection (Fig. 3A, black curve). The projections of the Neanderthal genome onto CEU and CHB show a large excess of rare alleles (Fig. 9I and 10I), which requires the combination of a bottleneck in the ancestors of non-Africans and admixture from the ancestors of Neanderthals into the ancestors of non-Africans after that bottleneck. In our model (Fig. 7), including these two processes, we obtain good fits to the observed projections (Table 4, Fig. 9I and 10I).

When admixture or the bottleneck is removed, the result is a decrease in the excess of rare alleles and a worse fit to the projections (Supplementary Table 1, Supplementary Fig. 1).

Similarly, the projections of the Denisovan genome onto CEU and CHB (Fig. 9H and 10H) are consistent with the three-population analysis shown in Fig. 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 the Neanderthal and Denisovan populations 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. When the Denisova and Neanderthal populations are sister groups, there is an excess of rare alleles that is predicted by our model (Table 4, Fig. 9H and 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 (Fig. 11I) and Denisovan (Fig. 11H) genomes 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 from the ancestors of CEU and CHB into the ancestors of YRI. In our model (Fig. 7), there is no admixture between YRI and any archaic hominin, but there is gene flow between the ancestors of YRI and non-Africans. An excess of rare alleles is observed in the simulated projection (Fig. 11H and 11I). Admixture from non-Africans to Yoruba must have

occurred more recently than the Neanderthal admixture into non-African populations for this signal to be present.

Relationship among Non-African Populations

The projections of the French genome onto CHB (Fig. 10A) differs from the projection of the Han genome onto CEU (Fig. 9A). This difference reflects the subtle interplay between admixture and population size changes. A model in which the ancestors of CHB experienced a bottleneck after their separation from the ancestors of CEU 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 CEU to CHB than the reverse can account for the overall increased excess seen in the French projection onto CHB (Fig. 10A). When these events are included in our model, the resulting projections are relatively close to the observed projections (Table 4). Including a more extreme bottleneck with less recent population expansion in the Papuan demographic history (Fig. 7) also leads to relatively similar projections (Table 4, Fig. 9B and 10B).

Relationship Between Non-Africans and YRI

The projections of the Papuan, French and Han genomes onto YRI (Fig. 11A, 11B, and 11D) are quite similar despite the difference between the Han and Papuan projection onto CEU (Fig. 9A and 9B). 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 declined in the recent past. These two processes together explain the dip observed and the increase to $\bar{w}(x)=1$ for larger x, and they lead to a good fit to the observed projections (Table 4, Fig.

11A, 11B, 11D). 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 (Supplementary Table 2, Supplementary Fig. 2).

African projections onto CEU and YRI

The projections of all five African genomes—San, Yoruba, Mandenka, Dinka, and Mbuti—onto CEU (Fig. 9C-G) are similar to one another and similar to their projections onto CHB (Fig. 10C-G). All these projections are consistent with low levels of admixture from the African populations into the ancestors of CEU and CHB. This conclusion is surprising. Previous analyses (Meyer et al 2012, Prüfer et al 2014, Lachance et al 2012, Pickrell et al, 2012) showed that the San population diverged from other African populations before the other African populations diverged from one another and before the ancestors of CEU and CHB diverged from each other. The separate history of the San is not reflected in their 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 similar results to the observed projections (Table 4, Fig. 9C-G). A closer look at the middle of the projection for reference CEU shows that the San projection is lower than the Yoruba projection (Fig. 9D and 9G), which suggests that the difference in divergence time has a slight effect on the projection.

The projections of different African genomes (Dinka, Mandenka, Mbuti, San) onto YRI (Fig. 11C, 11E-G) illuminate the relationship between these four African populations and the Yoruba. Other studies (e.g. Tishkoff et al 2009) 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 $\bar{w}(x) = 1$ line, which suggests it is less diverged from YRI than is the San, agreeing with the model proposed in other studies (e. g. Tishkoff et al 2009). The Mandenka projection falls nearly on the $\bar{w}(x) = 1$ line, suggesting it is indistinguishable from a random YRI individual. Finally, the Dinka projection onto YRI exhibits a dip that is similar, though 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 (Fig. 7) gives a close fit to the observed projections (Table 4, Fig. 11C, 11E-G).

Test of Published Models

We used observed projections to test for consistency with inferred demographic parameters from four studies (GUTENKUNST *et al*. 2009; EXCOFFIER *et al*. 2013; HARRIS AND NIELSEN 2013; SCHIFFELS AND DURBIN 2014) for European and Yoruba populations. All four of these studies applied their methods to CEU and YRI.

We obtained projections by using *fastsimcoal2* (EXCOFFIER *et al*. 2013) to simulate one 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 individual 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 YRI and CEU 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, YRI is assumed to be of constant size while the size declines in Models B and D. In Model C, the ancestral YRI population undergoes a bottleneck 797 generations ago. In all four models, the population ancestral to CEU and YRI increased in size before the two populations separated. In Models B and C, the time of divergence of CEU and YRI was about 50 kya, while in Model A it is 140 kya. In Model D the separation time is at least 150 kya.

Model A assumes higher rates of migration soon after the CEU-YRI divergence and a lower rate more recently. Model B allows for migration between CEU and YRI and it also includes a parameter for ghost admixture from an archaic hominin that diverged 14605 generations ago (365 kya). Model C uses a continent-island model, in which CEU and YRI diverged from continental European and African populations recently, receiving migrants from those populations until the present. However, neither they nor their ancestral populations admixed with each other. Model D does not allow for migration between CEU and YRI, though 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 5, Figure 13). The other three models do not give projections that fit as well. Modifying the amount of admixture and/or migration in each of Models B-D resulted in a substantially better fit (Table 5, Figure 14). In Model B, increasing the migration rate from YRI to CEU and adding admixture 150 generations ago at a rate of 0.02 from CEU to YRI 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 (Table 5, Models B*-D*, Fig. 14).

We show that projection analysis strongly supports the hypothesis that there was significant gene flow between the ancestors of CEU and YRI after there was introgression from Neanderthals into CEU. Gutenkunst et al. (2009) (Model A) reached this conclusion by allowing for such gene flow in their analysis. Adding such gene flow to Models B-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 such as those developed by Gutenkunst et al. (2009), Harris and Nielsen (2013), Excoffier et al. (2013), and Schiffels and Durbin (2014) that estimate model parameters. 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) 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 model a low amount from the separation until a few thousand years ago. Excoffier et al (2013) do provide a good fit for the French projection onto YRI, perhaps because of the large

bottleneck they infer in the ancestral Yoruba, but the Yoruba projection onto CEU requires some admixture for a better fit. Schiffels and Durbin's (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 for simplicity that 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.

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 does affect 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 or more 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.

Acknowledgements

This research was supported in part by a National Institutes of Health NRSA Traineeship (T32 HG 00047) and a National Science Foundation Graduate Research Fellowship Program Division of Graduate Education (1106400) to M.A.Y, and in part by a National Institutes of Health grant (R01-GM40282) to M.S. We thank K. Harris, 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.

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.

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 Genetics 5.

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

Harris, K., and R. Nielsen, 2014 Error-prone polymerase activity causes multinucleotide mutations in humans. Genome Research, in press.

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-gatheres. 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.

Pickrell J.K., N. Patterson, C. Barbieri, F. Berthold, L. Gerlach *et al*., 2012 The genetic prehistory of southern Africa. Nature Communications 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 advance online publication.

The 1000 Genomes Project Consortium, 2010 A map of human genome variation from population-scale sequencing. Nature 467**:** 1061-1073.

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.

Appendix: The projection of a test genome onto a reference population and applications to humans and archaic hominids

Kelley Harris

Given a reference population of infinitely many individuals and a SNP of derived allele frequency x, the probability that a randomly sampled individual will carry the derived allele is x. Although this result is a trivial consequence of the definition of an allele frequency, it breaks down when the reference population is finite of size N and x is close to $1/N$. We can see this by using a simple binomial sampling argument, and this argument explains the dip observed at low frequencies in projection plots such as Figure 8 where the test individual is sampled from the reference population.

If the reference panel consists of N individuals, then addition of a test individual from the same population creates a panel of $N+1$ individuals. We will let $f_{N+1}(x)$ denote the site frequency spectrum of the augmented panel containing N reference individuals plus the test individual. In terms of these frequency spectra, the expected projection $\overline{w}(x)$ can be expressed as follows:

$$\overline{w}\left(k/N\right) = \frac{f_{N+1}(\frac{k+1}{N+1}) \cdot \frac{k+1}{N+1}}{f_{N+1}(\frac{k+1}{N+1}) \cdot \frac{k+1}{N+1} + f_{N+1}(\frac{k}{N+1}) \cdot \frac{N+1-k}{N+1}}$$

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 augmented panel 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 the $N+1$ augmented panel members have the derived allele.

If the population has uniform mate choice and has reached effective population size equilibrium, coalescent theory predicts that $f_{N+1}(k/N) \propto 1/k$. In this simple case, finite reference panel size does not affect projections much. Instead of $\overline{w}(k/N) = k/N$, we compute that $\overline{w}(k/N) = k/(N+1)$:

$$\overline{w}\left(k/N\right) = \frac{\frac{1}{k+1} \cdot \frac{k+1}{N+1}}{\frac{1}{k+1} \cdot \frac{k+1}{N+1} + \frac{1}{k} \cdot \frac{N+1-k}{N+1}} = \frac{k}{N+1}$$

However, in cases where the demography is not so trivial and projection analysis will be more interesting, the ratio of singletons to doubletons often departs from the ratio of 2.0 expected under neutrality. Recent human population growth has produced an excess of singletons; when we calculated the site frequency spectrum for the 1000 Genomes panel that was utilized in this paper, we found about 2.5 times as many singletons as doubletons, meaning that

$$\overline{w}(1/N) = \frac{\frac{2}{N+1}}{\frac{2}{N+1} + 2.5 \cdot \frac{N}{N+1}} \approx 0.8 \cdot \frac{1}{N}.$$

In Figure 8, we see that the CEU and CHB populations have lower values of $\overline{w}(x)$ for small x than does the YRI population. This makes sense given that non-African populations have larger singleton-to-doubleton ratios as a result of stronger departures from population size equilibrium.

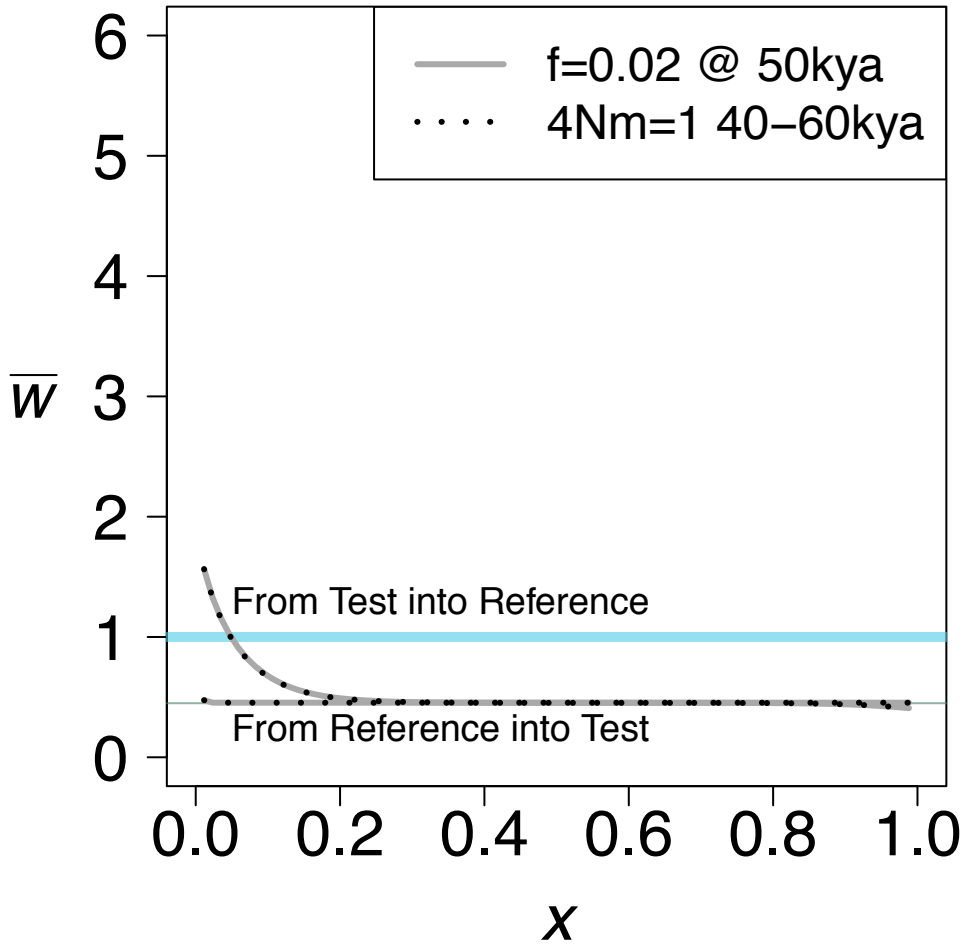


Figure 1: The effect of unidirectional gene flow on the projection of a test population onto a reference population. Two kinds of gene flow were assumed, either a single pulse of admixture of strength *f*, or a period of immigration at a rate *m* per generation. Both populations are of constant size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

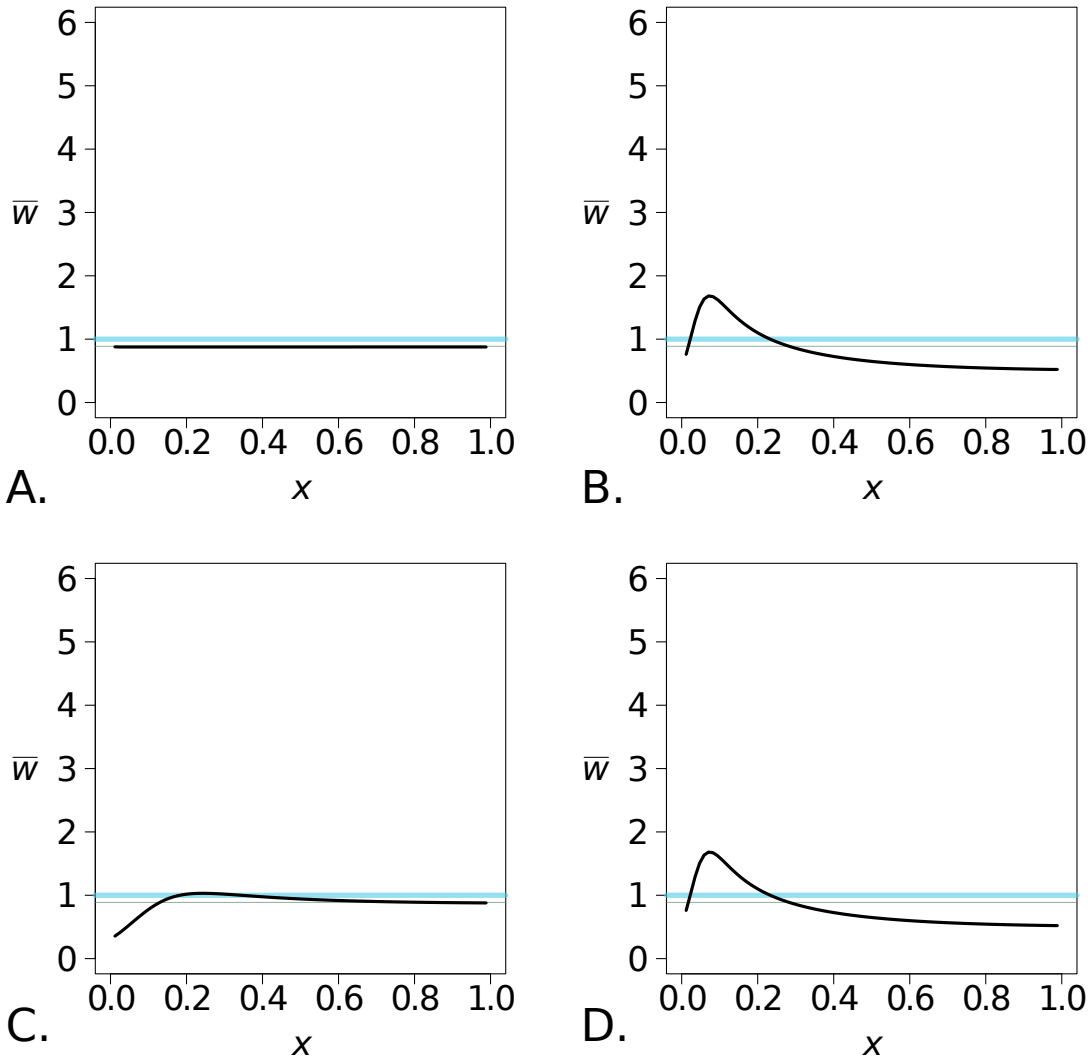


Figure 2: The effect of bottlenecks in population size in a model of two populations diverged at time τ=60,000 years. The population size was N=10,000 except during the bottleneck when N was reduced to 1,000. The bottleneck in the test population (A), the reference population (B), or both (D) was from 20,000 to 50,000 years in the past, and the bottleneck in the ancestor was (C) from 70,000 to 100,000 years in the past.

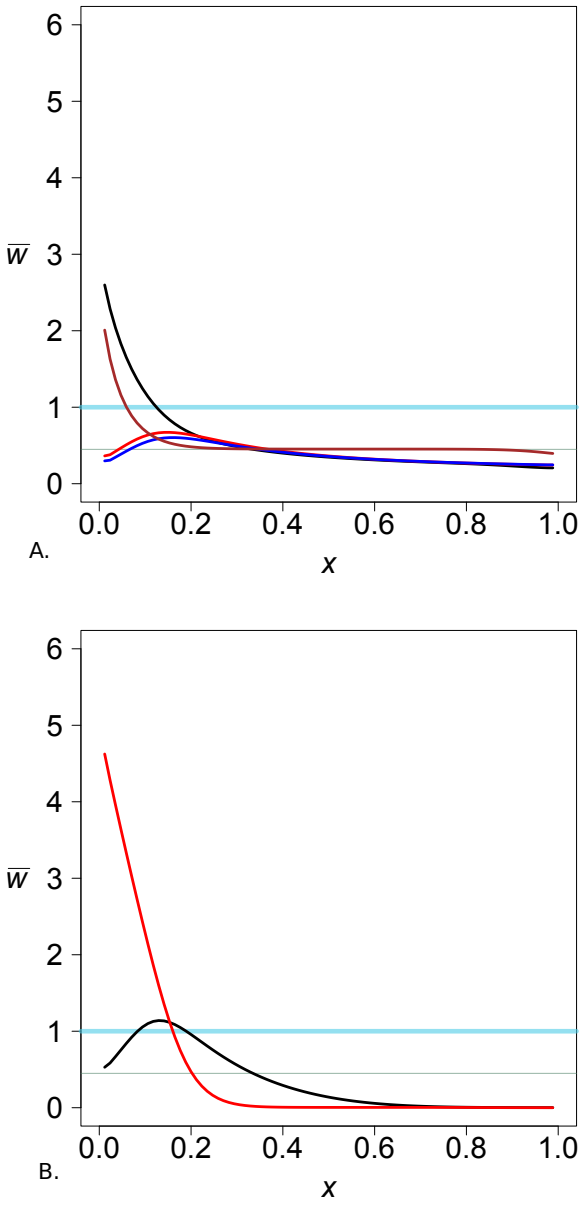


Figure 3: The combined effect of a bottleneck and admixture. The divergence time was τ=100,000 years. For A, the brown projection represents no bottleneck but admixture of f=0.02 at 40 kya. The other projections include admixture at 40 kya (black), 80 kya (red), and 120 kya (blue) of 0.02 from the test to the reference, where there was a bottleneck from 70 – 90 kya. The bottleneck reduced the reference population size from 10,000 to 1,000, and it increased to 10,000. For B, the reference population size increased from 1,000 to 10,000 at 40 kya only. The divergence time was τ=100,000 years. Admixture of 0.02 from the test to the reference occurred at 30 kya (red) and 50 kya (black).

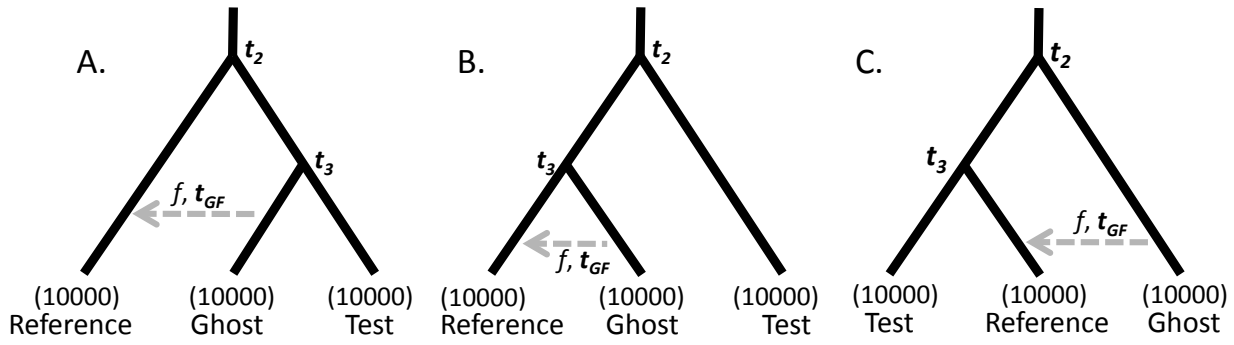


Figure 4: Illustration of three possible population relationships in which there is a pulse of admixture of intensity f at time t_{GF} in the past from the ghost population into the reference population. t_2 and t_3 are the times of population separation.

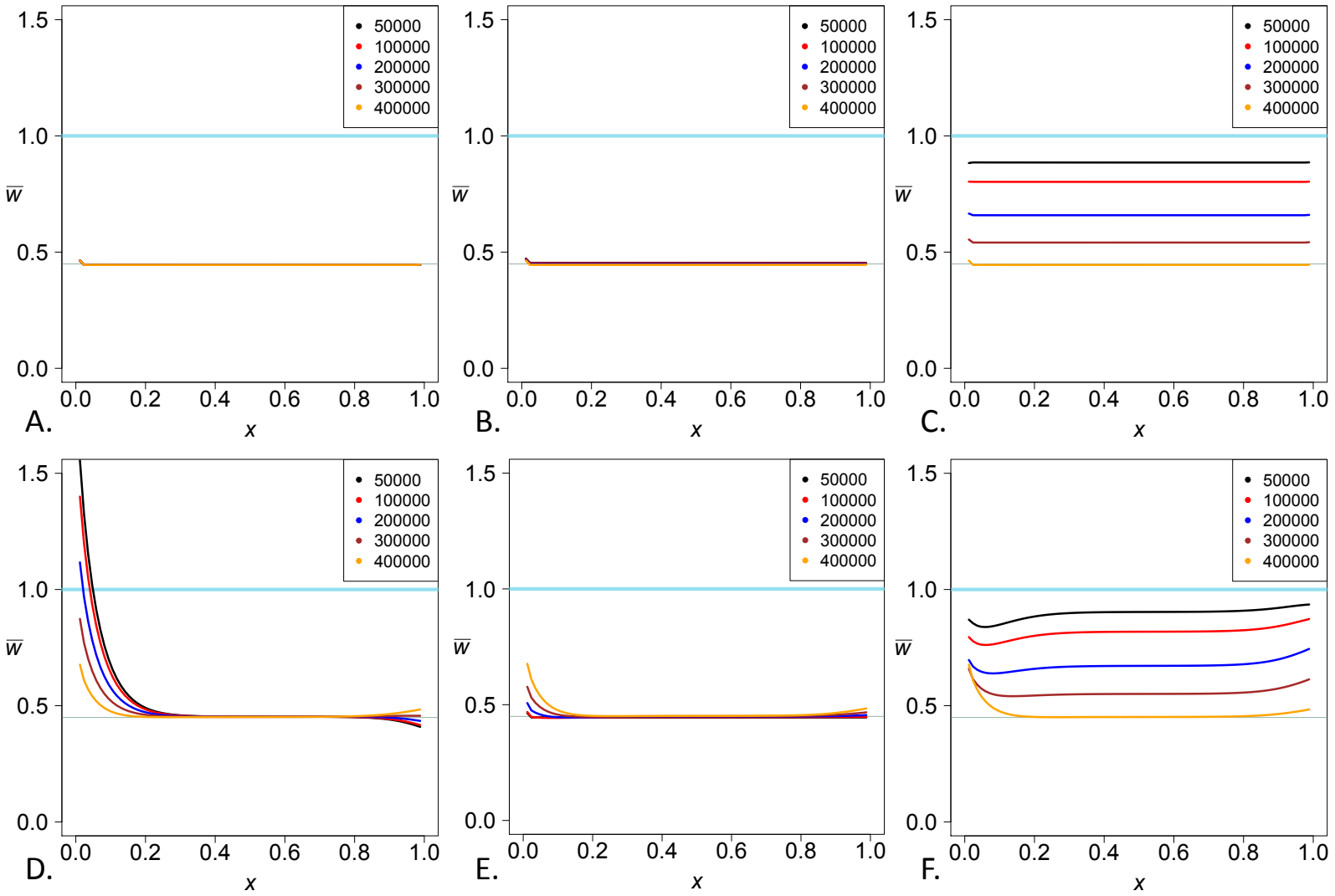


Figure 5: The effect of ghost admixture into the test (A-C) and the reference (D-F). A and D follow the topology in Fig. 4A, B and E follow Fig. 4B, and C and F follow Fig. 4C. t_2 = 400 kya, f = 0.02, and t_{GF} = 50 kya. t_3 is varied from 50 kya to 400 kya, according to the legend. Population sizes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

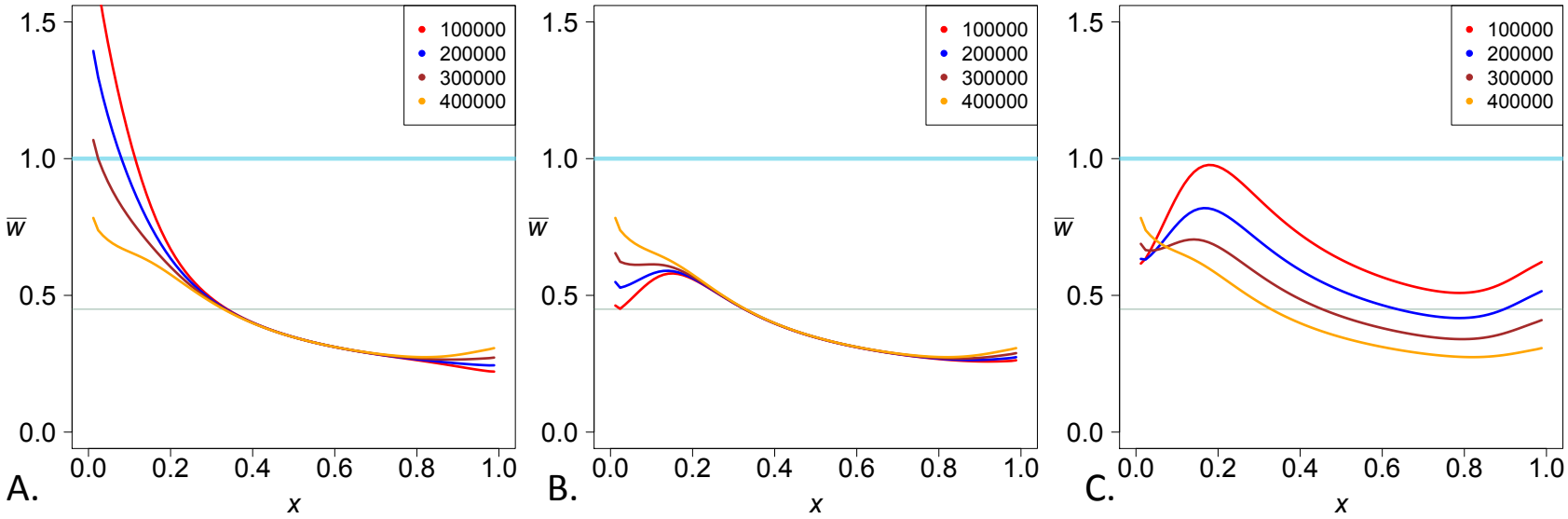


Figure 6: The effects of ghost admixture into the reference with a bottleneck in the reference occurring 70-100 kya changing the reference population size from 10,000 to 1,000 and back to 10,000. t_3 is varied from 100 kya to 400 kya. All other parameters are the same as in Fig. 5.

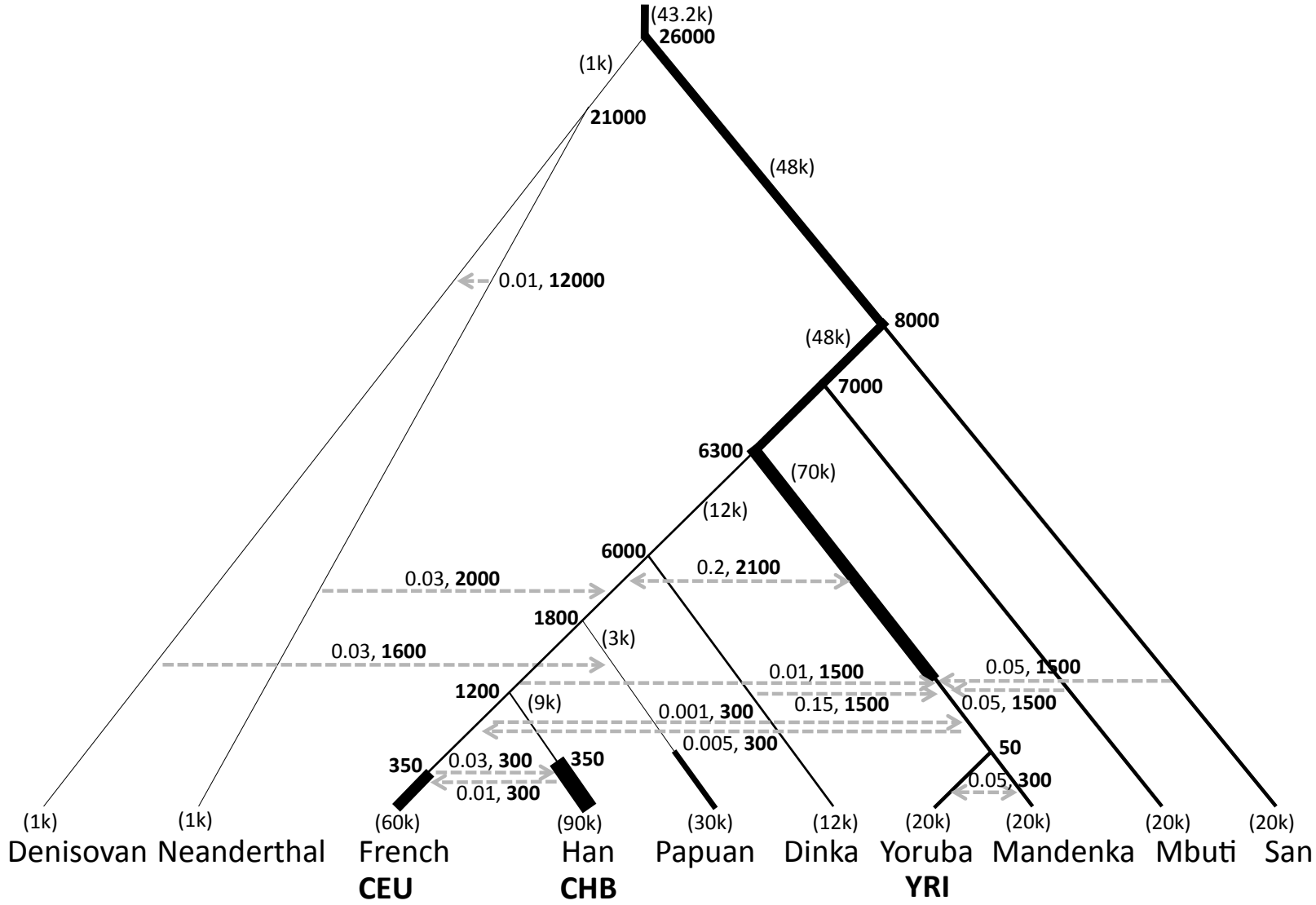


Figure 7: A model of human demographic history for ten populations that gives simulated projections similar to the observed projections. The bottom row refers to the reference populations, and the row above indicates the population origin of each test genome. Numbers in bold indicate time in the past an event occurred (in generations). The effective population sizes are indicated in parentheses.

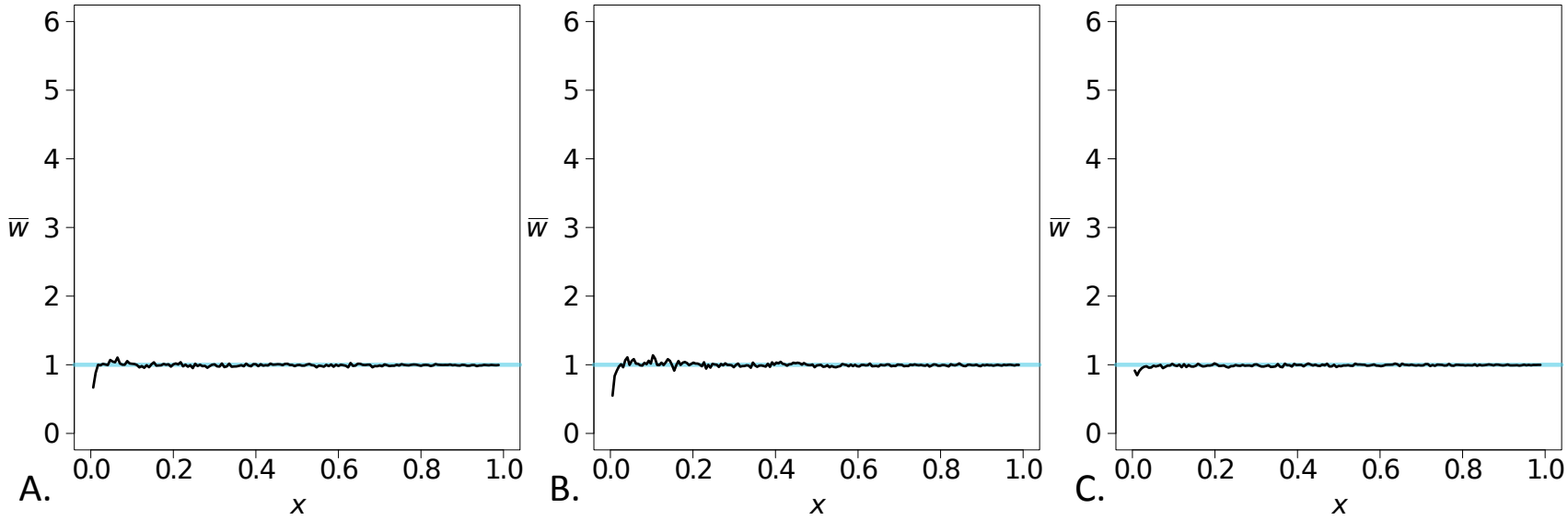


Figure 8: The projections of French onto CEU (A), Han onto CHB (B), and Yoruba onto YRI (C).

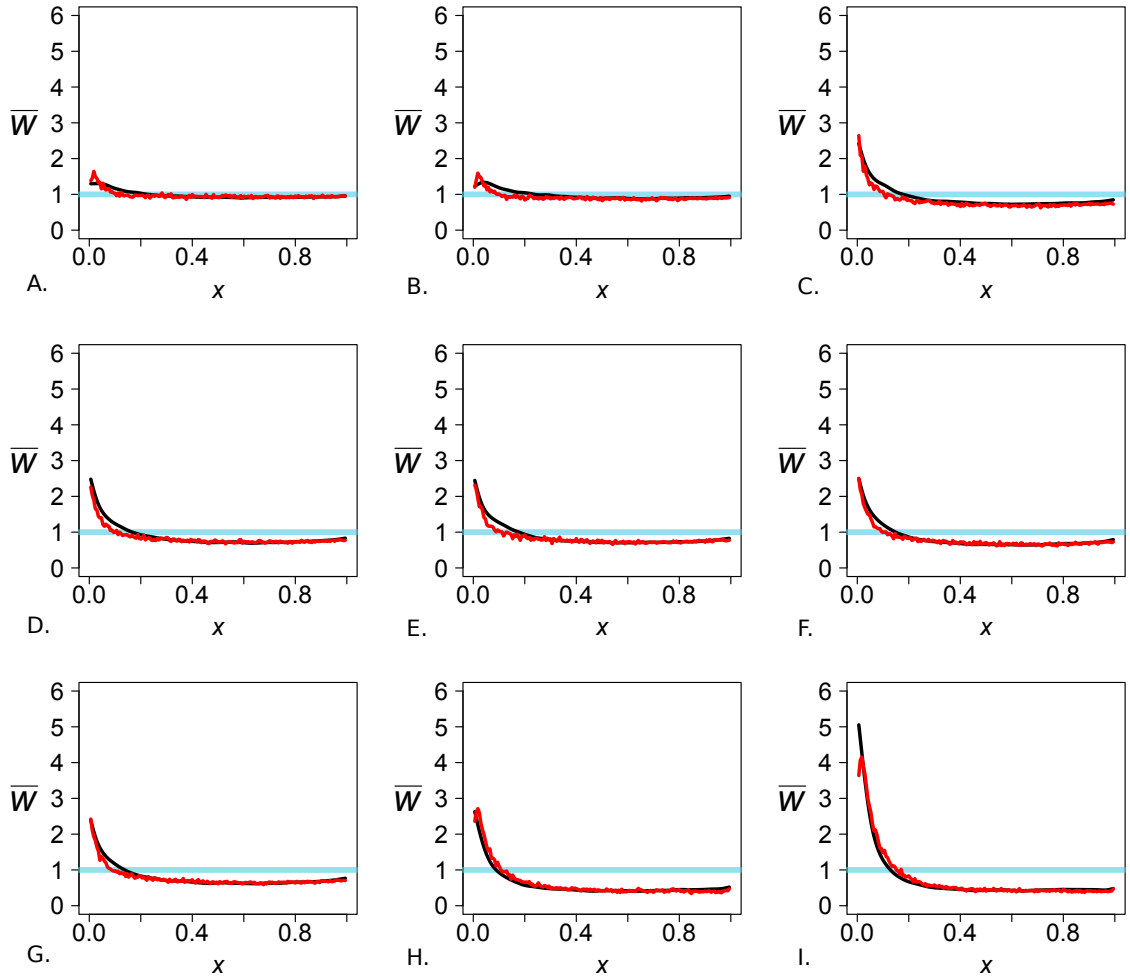


Figure 9: The observed projection (black) and simulated projection from our model (red) for the CEU reference population. The test genomes are Han (A), Papuan (B), Dinka (C), Yoruba (D), Mandenka (E), Mbuti (F), San (G), Denisovan (H), and Neanderthal (I).

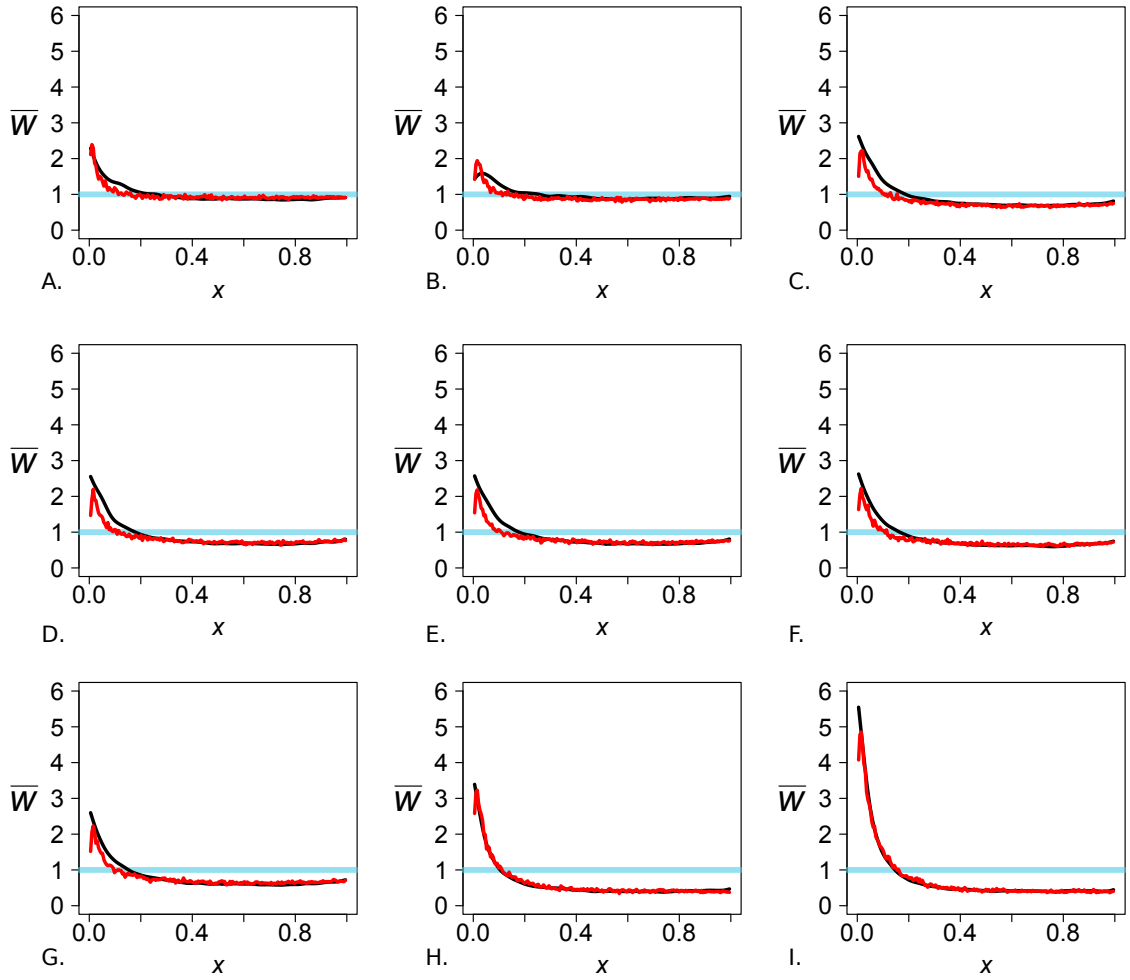


Figure 10: The observed projection (black) and simulated projection from our model (red) for the CHB reference population. The test genomes are French (A), Papuan (B), Dinka (C), Yoruba (D), Mandenka (E), Mbuti (F), San (G), Denisovan (H), and Neanderthal (I).

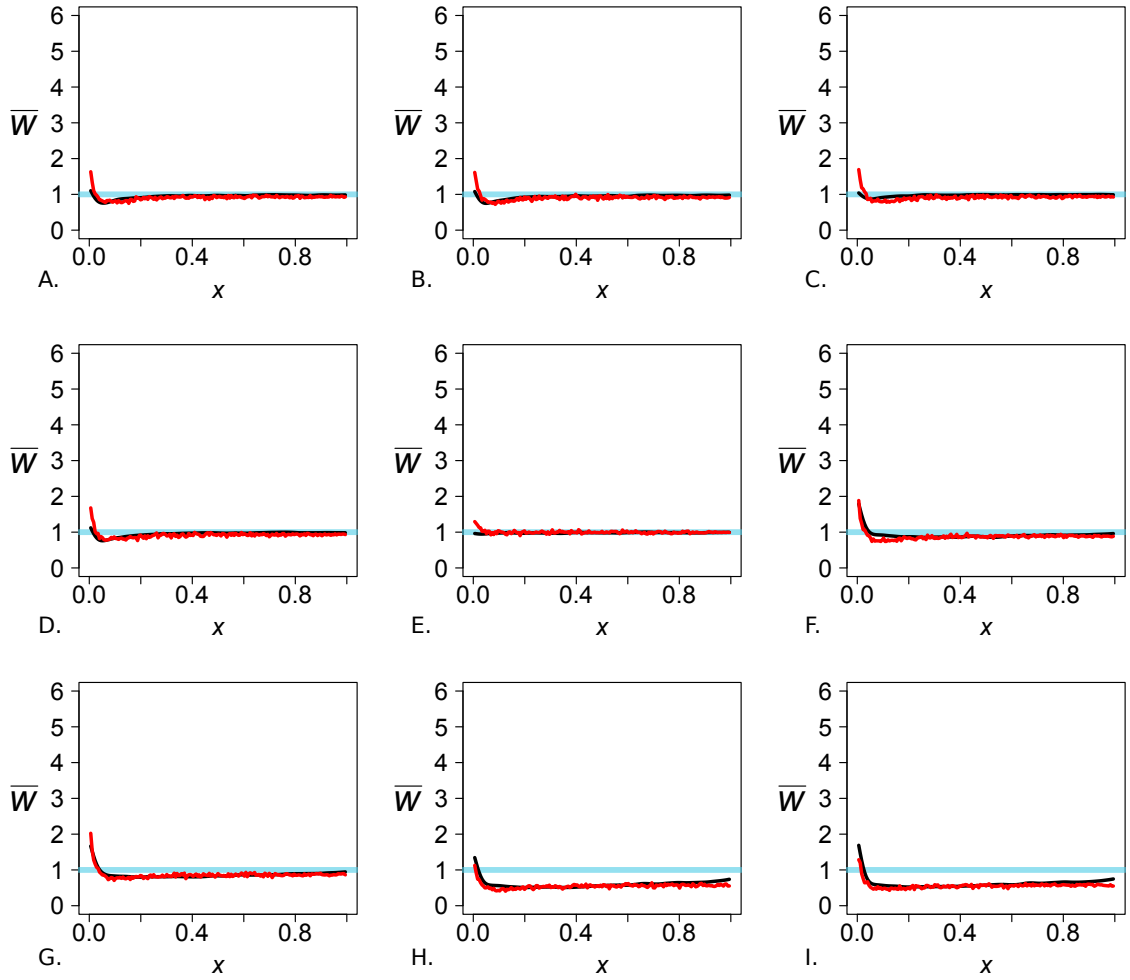


Figure 11: The observed projection (black) and simulated projection from our model (red) for the YRI reference population. The test genomes are Han (A), Papuan (B), Dinka (C), French (D), Mandenka (E), Mbuti (F), San (G), Denisovan (H), and Neanderthal (I).

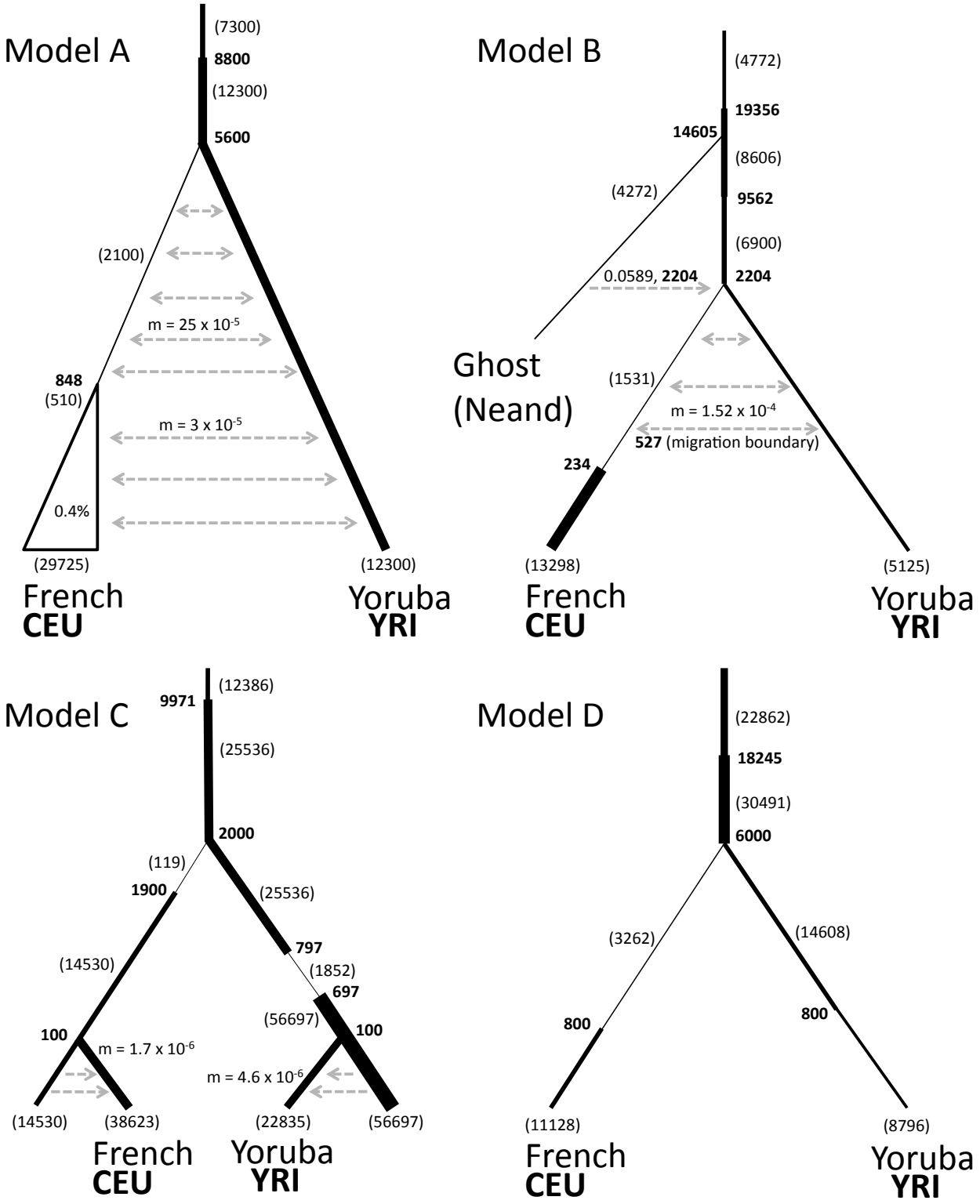


Figure 12: The demographic models from each of the four previous studies, Gutenkunst et al (2009, Model A), Harris and Nielsen (2011, Model B), Excoffier et al (2013, Model C), and Schiffels and Durbin (2014, Model D). Shading and symbols have the same meaning as in Figure 7, and the triangle indicates growth at the given percentage.

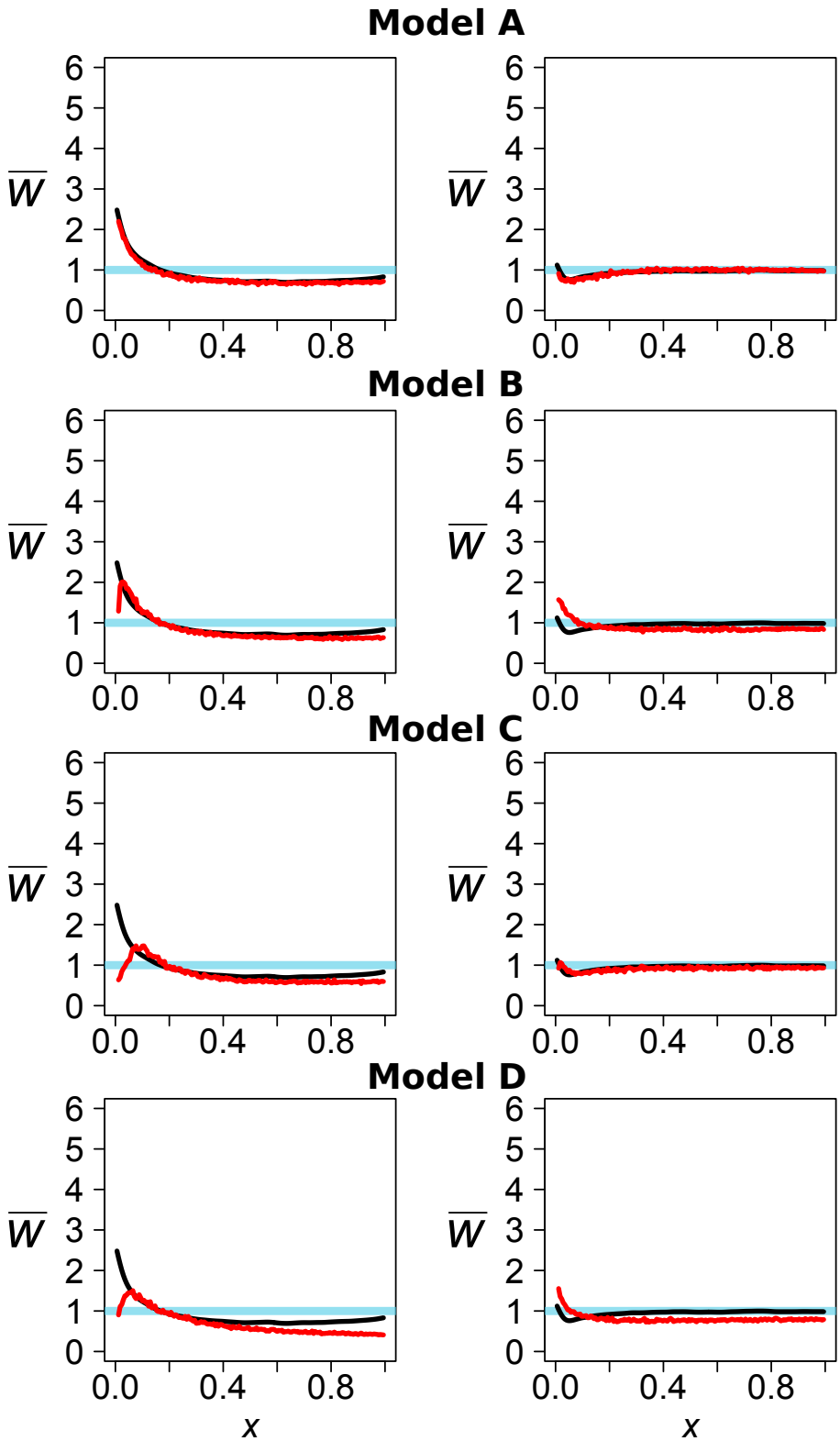


Figure 13: The observed projections (black) and simulated projections from demographic models inferred from other studies (red). The left column is the Yoruba genome projected on CEU and the right column is the French genome projected on YRI. The rows represent the estimates from Models A-D. LSS scores are in Table 5.

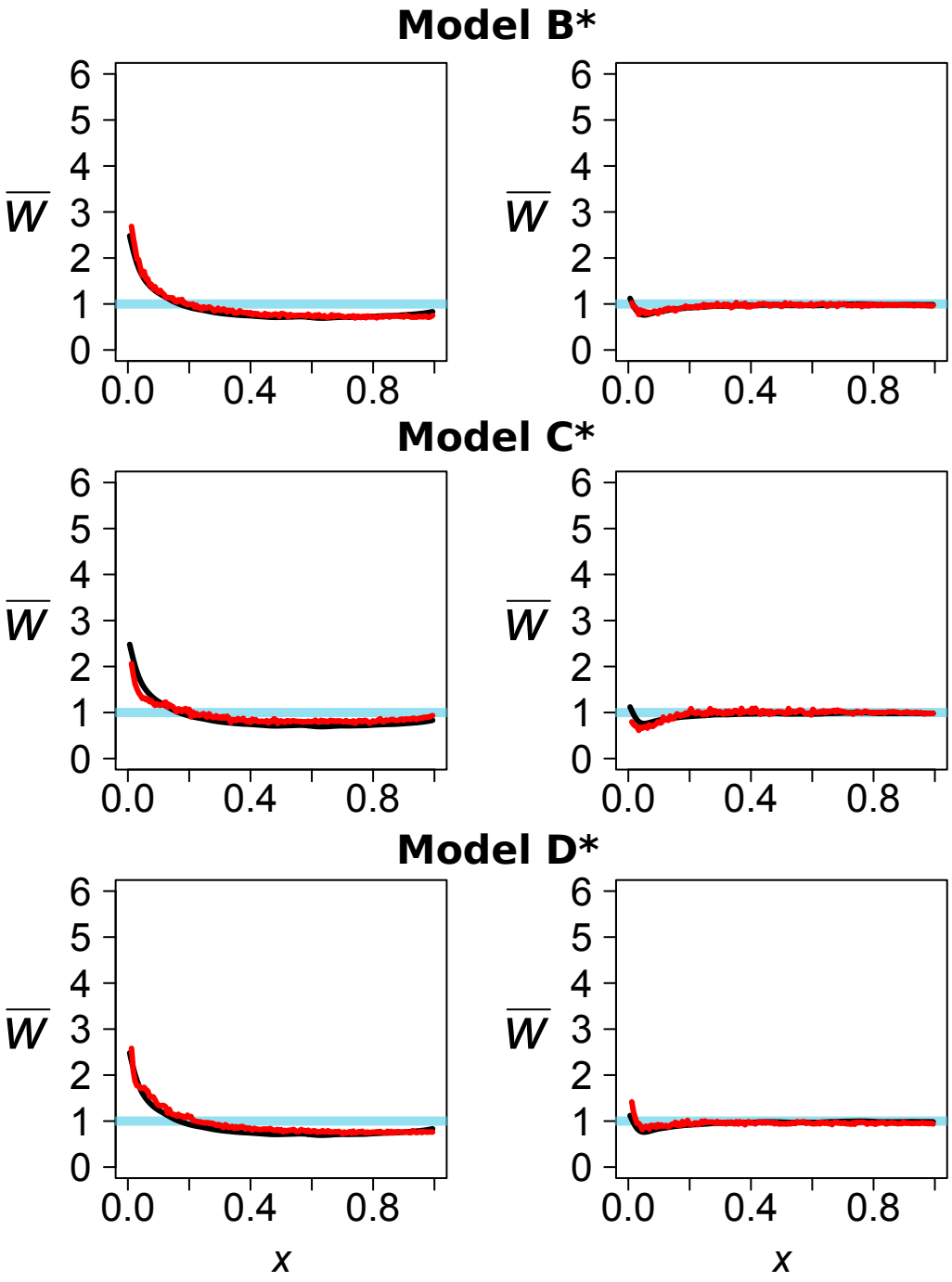


Figure 14: Projections for previous studies (Models B-D) where the parameters for migration or admixture between Europeans and Yorubans have been added or modified for a better fit. The left column is the Yoruba genome projected onto CEU and the right column is the French genome projected onto YRI. LSS scores are in Table 5.

Table 1: LSS comparing the observed projections for reference CEU 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	**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 2: LSS comparing the observed projections for reference CHB 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	**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 3: LSS comparing the observed projections for reference YRI to each other. The diagonals compare that test genome to the $\bar{w}(x) = 1$ line.

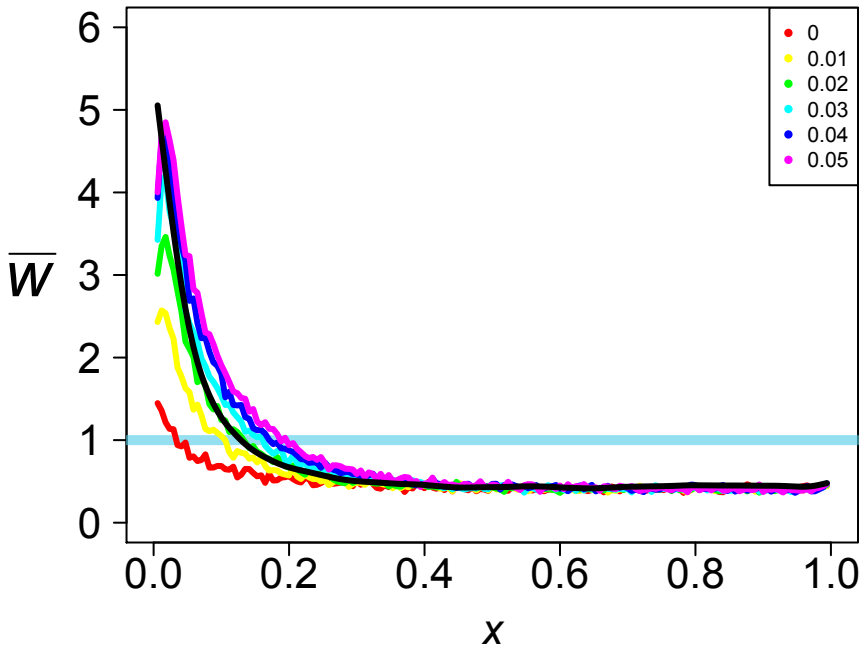
	Mandenka	Papuan	Han	Yoruba	Mbuti	Dinka	Deni	San	Altai	French
Mandenka	**0.12**	1.02	0.92	0.12	3.64	0.23	28.44	4.72	27.68	0.90
Papuan		**1.46**	0.10	1.33	2.95	0.49	22.50	3.21	21.66	0.12
Han			**1.30**	1.22	2.98	0.41	23.50	3.37	22.58	0.08
Yoruba				**0.10**	4.21	0.36	29.90	5.41	29.22	1.20
Mbuti					**3.96**	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 4: LSS comparing the simulated projection from our model (Fig. 10) to the observed projections.

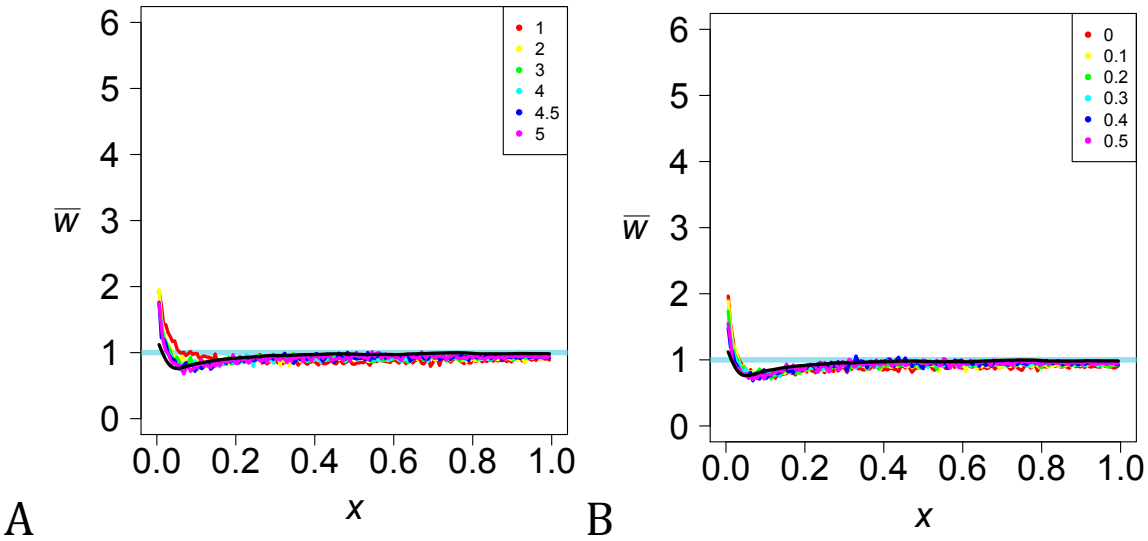
	Reference		
Test	**CEU**	**CHB**	**YRI**
French	*	2.16	1.15
Han	0.73	*	0.94
Papuan	1.28	2.44	1.10
Dinka	2.01	5.83	1.65
Yoruba	1.82	5.96	*
Mandenka	1.60	5.64	0.49
Mbuti	1.14	4.31	0.95
San	1.01	3.69	0.50
Denisovan	2.02	1.20	1.17
Neanderthal	3.80	2.84	1.46

Table 5: LSS comparing the simulated projections for the best estimates from four previous studies (Models A-D) and three previous studies where admixture was added or modified (Models B*-D*) to the observed projections.

	Test/REF	
Model	**Yoruba/CEU**	**French/YRI**
A	0.64	0.29
B	5.55	5.88
C	15.45	0.74
D	13.91	7.32
B*	0.93	0.14
C*	2.24	0.68
D*	3.17	1.20



Supplementary Figure 1: The simulated projections for reference CEU and test Neanderthal for our model when altering the amount of admixture. The black line is the observed projection.



Supplementary Figure 2: 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 and (B) the amount of admixture from Europeans to Yorubans. The black line is the observed projection.

Supplementary Table 1: LSS comparing our model to the observed projections, altering the amount of admixture from Neanderthals to non-Africans.

Test Neanderthal			
Value	**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**

Supplementary Table 2: LSS comparing our model to the observed projections, altering the population increase in the Yoruba population backwards in time and the amount of admixture from Europeans to Yorubans.

Test French			
Value	**YRI**	**Value**	**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