UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title
Using Aspects of the Site Frequency Spectrum to Determine Demographic History in Ancient and Modern Populations

Permalink
https://escholarship.org/uc/item/5qm5v4fw

Author
Yang, Melinda Anna

Publication Date
2015

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Using Aspects of the Site Frequency Spectrum to Determine Demographic History in Ancient and Modern Populations

by

Melinda Anna Yang

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Integrative Biology

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Montgomery Slatkin, Chair
Professor Rasmus Nielsen
Professor Junko Habu

Fall 2015

Using Aspects of the Site Frequency Spectrum to Determine Demographic History in Ancient and Modern Populations

Copyright 2015
by
Melinda Anna Yang

Abstract

Using Aspects of the Site Frequency Spectrum to Determine Demographic History in Ancient and Modern Populations

by

Melinda Anna Yang

Doctor of Philosophy in Integrative Biology
and the Designated Emphasis in
Computational and Genomic Biology

University of California, Berkeley

Professor Montgomery Slatkin, Chair

The site frequency spectrum characterizes the variation found within different studied populations, providing insight on past demographic history. Here, I describe two methods of studying demographic history that condition or transform the site frequency spectrum to provide greater understanding of demographic history. The doubly conditioned site frequency spectrum helps to highlight differences between admixture and ancient structure models, helping to confirm that Neanderthals did admix with the ancestors of non-African humans. Projection analysis compares a single test genome to a reference population and provides insight on the demographic relationship between the reference and test populations analyzed. This method has been applied in humans to better articulate human demographic history and demonstrate the utility of the method. It has also been applied to genomic data from ancient and modern horses, highlighting admixture between the Przewalski's horse and domestic horses, as well as the relationship of several ancient horse genomes to present day horse populations.

To my father, mother and brother

For all your support and love.

Contents

Contents **ii**

List of Figures **iv**

List of Tables **x**

1 Introduction **1**

2 The doubly conditioned frequency spectrum and applications **6**
2.1 Introduction 6
2.2 Materials and Methods 8
2.3 Results 14
2.4 Discussion and Conclusion 19

3 Projection of a Test Genome onto a Reference Population **23**
3.1 Introduction 23
3.2 Analytic Theory 24
3.3 Discussion and Conclusions 34

4 Applications of Projection Analysis in Ancient and Modern Human Populations **36**
4.1 Introduction 36
4.2 Application to Humans, the Altai Neanderthal and the Denisovan 36
4.3 Additional Application to Neanderthals and Ancient Humans 53
4.4 Discussion and Conclusions 57

5 Applications of Projection Analysis in Ancient and Modern Horse Populations **60**
5.1 Introduction 60
5.2 Demographic history of the Przewalski's horse 60
5.3 Demographic history of the Yakut horse 71
5.4 Conclusion 77

A Deriving the conditional derived *sfs* using coalescent theory **78**

B *ms* simulations used in Chapter 2 **80**

Bibliography **82**

List of Figures

1.1 A simple example of the site frequency spectrum (*sfs*). In this figure, (A) gives five genomes and four polymorphic sites and (B) gives the *sfs* associated with (A). 2

2.1 Demographic models relating Neanderthals and modern humans. 9
2.2 The *dcfs* for the CEU population and simulations of recent admixture (a, c, e) and ancient structure (b, d, f). 13
2.3 The *dcfs* for the CHB population and simulations of recent admixture. 15
2.4 The *dcfs* for the JPT population and simulations of recent admixture. 16
2.5 The *dcfs* from Europeans from Complete Genomics (black solid line) and simulations of a demographic history including ancient structure and no bottleneck (A) and a bottleneck at $t_b = 0.1$ (B). 17
2.6 The *dcfs* for the CEU population and simulations of recent admixture and ancient structure with a bottleneck of 0.001. 18
2.7 The *dcfs* for the CEU population and simulations of recent admixture and ancient structure with a bottleneck of 0.01. 19
2.8 The *dcfs* for simulations of recent admixture and ancient structure for no bottleneck or present day gene flow, including misassignment of the ancestral allele from no misassignments to a 25% error rate. 20
2.9 The *dcfs* for simulations of recent admixture of 2% (red), 3% (blue), 5% (black), and 10% (green), for no bottleneck (star), a bottleneck at 0.1 (square), and a bottleneck at 0.03 (circle). 21

3.1 The effect of unidirectional geneflow 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. 25

3.2 The effect of population size changes in a model with two populations that diverged 60 kya.In the figure, (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. 26
3.3 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). 27
3.4 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. 28
3.5 The effect of ghost admixture into the test (A-C) and the reference (D-F). A and D follow the topology in Figure 3.4A; B and E follow the topology in Figure 3.4B; and C and F follow the topology in Figure 3.4C. $t_2 = 400$ kya, $f = 0.02$, and $t_{GF} = 50$ kya. t_3 is varied from 50 to 400. Population sizes remain constant at 10,000. 29
3.6 The effects of ghost admixture into the reference with a bottleneck in the reference occurring 70100 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 3.5. A follows the topology in Figure 3.4A, B follows the topology in Figure 3.4B, and C follows the topology in Figure 3.4C. 30
3.7 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. 31

3.8 Simulated demographic models used to illustrate the effect of ancient samples in a one-population and two-population model.The * represents where the present-day population was sampled and the gray dashed line indicates when the ancient genomes were sampled (0 - 4k gen). Any divergence occurs 2k gen ago. For both OPM and TPM, A has an N_e of 5k, with no population size changes or admixture. B adds a pulse of admixture from the second diverging population. C has no admixture but allows a population size expansion from 500 to 5k in the reference population 750 gen ago. D allows the reverse, a population size decline from 5k to 500 in the reference population 750 gen ago. E has a bottleneck from 5k to 500, 500-1000 gen ago. Any diverging population has the same N_e as the ancestral population. 32
3.9 One Population Model simulated projections for the demographic models tested in Figure 3.8. The key indicates the time the ancient genomes were sampled, in generations. The top row gives the results for when the reference panel is sampled from the present and the test genome is sampled from the past. The bottom row gives the reverse. $\bar{w}$ is the value of the projection and x is the derived allele frequency in the reference population. 33
3.10 Two Population Model simulated projections for the demographic models tested in Figure 3.8. The key indicates the time the ancient genomes were sampled, in generations. The top row gives the results for when the reference panel is sampled from the present and the test genome is sampled from the past. The bottom row gives the reverse. $\bar{w}$ is the value of the projection and x is the derived allele frequency in the reference population. 34

4.1 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 4.1-Table 4.3. 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. . 40
4.2 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 4.5, Table 4.6, and Table 4.7. 41
4.3 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 4.5, and the LSS scores comparing the observed and simulated projections can be found in Table 4.4. . . 42

4.4 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 4.6, and the LSS scores comparing the observed and simulated projections can be found in Table 4.3. . . 43
4.5 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 4.7, and the LSS scores comparing the observed and simulated projections can be found in Table 4.3. . . 44
4.6 The simulated projections for reference CEU and test Neanderthal for our model when altering the amount of admixture ($f_{NEA-ANC1}$). The black line is the observed projection. 46
4.7 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 ($\frac{N_{ANC4}}{N_{YOR}}$) and (B) the amount of admixture from Europeans to Yorubans ($f_{ANC1ANC4}$). The black line is the observed projection. 48
4.8 The demographic models from each of the four previous studies: Gravel et al. [19] (model A), Harris and Nielsen [27] (model B), Excoffier et al. [13] (model C), and Schiffels and Durbin [58] (model D). Shading and symbols have the same meaning as in Figure 4.1, and the triangle indicates growth at the given percentage. 51
4.9 The observed projections (black line) and simulated projections from demographic models inferred from other studies (red line). For each model A-D in Figure 4.6, 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 4.8. 52
4.10 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 4.8. 53
4.11 Projection of ancient hominin genomes onto the European reference panel (refCEU, top row), the Han Chinese reference panel (refCHB, middle row) and the Yoruba reference panel (refYRI, bottom row), with the observed projection in black and the simulated projection in red. The sum of least squares (LSS) score gives the fit between the observed and simulated projections. The mean projection value (MPV) is the mean for $x > 0.5$. 54

4.12 The placement of the five ancient genomes (black circles) in the demographic model described in Figure 4.1shaded gray, Table 4.1 to 4.3 contains parameter values). The time of sampling for these five genomes are included in Table 4.11. Bolded parameters are those that were modified to improve the fit onto the projections (values in Table 4.13). 56

5.1 Projection of each Przewalski's horse genome onto the Przewalski's horse refPRZ reference panel. Projections are separated into genomes with high signals of admixture with domesticated horses (A), moderate signals of admixture (B), or little to no signal of admixture (C) as determined on the basis of ABBA-BABA admixture tests. 62
5.2 Projections of all genomes onto the Przewalski's horse refPRZ reference panel. A. Przewalski's horse test genomes. B. Domesticated horse test genomes. C. Franches-Montagnes test genomes. 63
5.3 Projections of all genomes onto the domesticated horse refDOM reference panel. A. Przewalski's horse test genomes. B. Domesticated horse test genomes. C. Franches-Montagnes test genomes. 64
5.4 Projections of all genomes onto the Franches-Montagnes refFM reference panel. A. Przewalski's horse test genomes. B. Domesticated horse test genomes. C. Franches-Montagnes test genomes. 65
5.5 Observed and simulated projections for Franches-Montagnes and Przewalski's horses as either test or reference genomes. Simulated projections (colored according to key) were obtained for Models A-E as described in Figure 5.1. Observed projections are drawn in black. A. Projection of Przewalski's horse test genome onto Franches-Montagnes reference genome panel (refFM). B. Projection of Franches-Montagnes horse test genome onto Przewalski's horse reference genome panel (refPRZ). 66
5.6 Demographic models tested in projection analyses (adapted from dadi model C). FM: Franches-Montagnes; PRZ: Przewalski's horses; NANC, effective population size at end of population size decay at 10,000 yBP; NPRZ: effective population size of Przewalski's horses; NFM: effective population size of Franches-Montagnes horses; MPF1: migration rate of Przewalski's horses into Franches-Montagnes horses (before the change of migration rate for model C3); MPF2: migration rate of Przewalski's horses into FranchesMontagnes horses after the change of migration rate for model C3; MFP1: migration rate of Franches-Montagnes horses into Przewalski's horses (before the change of migration rate for model C3); MFP2: migration rate of Franches-Montagnes horses into Przewalski's horses after the change of migration rate for model C3; TMIG, the time of migration rate change between MPF1/MFP1 and MPF2/MFP2; FFP: recent admixture fraction. Migration rates are reported as the fraction of the sink population made up by new migrants. 68

5.7 Observed and simulated projections for Franches-Montagnes and Przewalski's horses as either test or reference genomes. Simulated projections were obtained for the modified model C, C1-3. Observed projections are drawn in black. A. Projection of Przewalski's horse test genome onto Franches-Montagnes horse reference genome panel. B. Projection of Franches-Montagnes horse test genome onto Przewalski's horse reference genome panel. 69
5.8 LSS scores for varying recent times of migration and amounts of admixture in Model C4, relative to observed projections on refPRZ. Different colors represent different amounts of admixture as indicated by the key. The x-axis indicates the time of recent admixture from Franches-Montagnes to Przewalski's horse and the y-axis indicates the LSS score for each pairing, where a lower LSS score indicates an improved fit of the simulated projection onto the observed projection. 70
5.9 Estimation of the time of migration rate change, using projections for Franches-Montagnes and Przewalski's horses as either test or reference genomes. Observed (black) and simulated projections (color) allowing for different times of migration rates changing for Franches-Montagnes projected onto Przewalski's horse reference genome panel (A) and Przewalski's horse projected onto Franches-Montagnes horse reference genome panel (B). C. LSS score variation with times of migration rate for both projections. Times shown are in yBP and were taken from time periods described in the section Joint demographic inference using dadi as potentially affecting interbreeding opportunities. All other parameters were the ones best fitting Model C3. 71
5.10 Projections of modern Yakutian horses onto the refYAK panel. The x-axis represents the categories of derived allele frequencies, while $\bar{w}$ represents the corresponding projection ($\bar{w}(x) > 1$ indicates that the test genome has more alleles at that frequency than the reference panel, and vice versa). 73
5.11 Projections of Przewalski's horses onto the refFM panel. See Figure 5.10 for captions. 74
5.12 Projections of modern Yakutian horses onto the refFM panel. See Figure 5.10 for captions. 75
5.13 Projections of non-Yakutian and non-Franches-Montagnes modern horses onto the refYAK panel. See Figure S6.2 for captions. 76
5.14 Projections of ancient horse genomes onto the refYAK panel. See Figure 5.10 for captions. 77

List of Tables

2.1 Parameters Used in *ms* [30] for the Recent Admixture and Ancient Structure Models. 12
2.2 The Average D Statistic for Each Simulated Demographic History in Figure 2.2. 12

4.1 Description of parameters used in the simulation of the 10-population tree in Figure 4.1. 37
4.2 Description of parameters used in the simulation of the 10-population tree in Figure 4.1 (continued from Figure 4.1). 38
4.3 Description of parameters used in the simulation of the 10-population tree in Figure 4.1 (continued from Figure 4.1 and Figure 4.2). 39
4.4 LSS comparing the simulated projection from our model (Figure 4.1) to the observed projections (Figure 4.3, Figure 4.4 and Figure 4.5). 41
4.5 LSS comparing the observed projections for reference CEU to each other. The diagonals compare that test genome to the $\bar{w}(x) = 1$ line. 45
4.6 LSS comparing the observed projections for reference CHB to each other. The diagonals compare that test genome to the $\bar{w}(x) = 1$ line. 45
4.7 LSS comparing the observed projections for reference YRI to each other. The diagonals compare that test genome to the $\bar{w}(x) = 1$ line. 45
4.8 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 47
4.9 LSS (Least Sum of Squares) comparing our model to the observed projections, altering the amount of admixture from Neanderthals to non-Africans ($f_{NEA-ANC1}$). 47
4.10 LSS comparing our model to the observed projections, altering the population increase in the Yoruba population backwards in time ($\frac{N_{ANC4}}{N_{YOR}}$) and the amount of admixture ($f_{ANC1-ANC4}$) from Europeans to Yorubans. 49
4.11 Data used for each ancient genome 52
4.12 Sum of least square scores when comparing projections of test genomes within a reference panel 55
4.13 Parameter values used in simulated demographic model that differs from Table 4.1 through 4.3. 57

5.1 Przewalski's and domesticated horse genomes used as test genomes or in reference panels. 61
5.2 Sum of least squares score (LSS) for the demographic models tested for Przewalski's and Franches-Montagnes horse genomes used either as test or reference. . . 66
5.3 Demographic parameters included in tested models. 67
5.4 Horse genomes used for projection analyses. 72
5.5 Minimum projection values. 75

Acknowledgments

First and foremost, I would like to thank my PhD advisors Montgomery Slatkin and Rasmus Nielsen, for their support as I kept chipping away at the mysteries of human evolutionary genetics. Without their advice, positive feedback, and constant encouragement, this thesis would not be here today. I would also like to thank my third committee member, Junko Habu, for helping me delve into the world of East Asian archaeology. Her support and discussions have helped me retain an anthropological perspective for future research and reminds me of the questions that motivate my research in the first place. Also, I would like to thank my fourth qualifying exam committee member, Tim White—thanks for passing me and teaching me so much about human paleontology.

Amongst so many others, I would like to thank my labmates—you all have made sitting in VLSB 4151 both a huge delight and a terrible distraction...at least I managed to get this thesis done. I would also like to thank the IB community, particularly the IB grad cohort of 2010—my sanity has more or less stayed intact over the years because of so many of the friends I made in this department. I look forward to using all the support, teachings and advice I have received in the last 5.5 years for many years to come.

Chapter 1

Introduction

In population genetics, the demographic history of a population refers to the set of population level events that occur in the past that leads to a population's current biological makeup. Some common events are population size changes, such as bottlenecks, decreases and expansions; and migration, which results in mixing with other closely related populations. Genetic data are ideal for studying demographic history because these demographic events often leave a noticeable trace in present day genomes. Using the pattern of variation left by mutations throughout the genome, we can infer which demographic model best fits the observed patterns.

One of the most widely used methods that has been developed to characterize patterns of mutation is the site frequency spectrum (*sfs*). The *sfs* is a set of summary statistics characterizing the relative abundances of mutations across a set of samples. It requires multiple samples to accurately represent a population, but it is a fast, flexible technique of summarizing variation across the genome.

There are two types: the unfolded and folded *sfs*. The unfolded *sfs* assumes knowledge of the ancestral or derived nature of the mutations analyzed and gives the *sfs* using the frequencies of the derived alleles. The folded *sfs* assumes no such knowledge and gives the *sfs* using the frequencies of the minor alleles. Thus, the folded *sfs* provides less information and folds the unfolded *sfs* in two. From here on out, I will focus on the unfolded *sfs* and use the term *sfs* to refer to the unfolded *sfs*.

A simple example of the *sfs* is given below. Consider a sample of five genomes with four polymorphic sites, L1 through L4 (Figure 1.1). If three individuals have a derived mutation at L1, a single genome has a derived mutation at L2, four individuals have a derived mutation at L3, and a single genome has a derived mutation at L4, then the *sfs* is (2,0,1,1), where each element is the number of sites where i genomes have the derived allele, where i ranges from one genome (frequency of 0.2) to four genomes (frequency of 0.8). When assuming a population of constant size and an infinite sites model, the equilibrium *sfs* is $\frac{\theta}{x}$, where x is the frequency of the derived allele in the sample, $\theta = 4N\mu$, μ is the mutation rate, and N is the population size.

The *sfs* is a powerful set of statistics for understanding genetic variation within a pop-

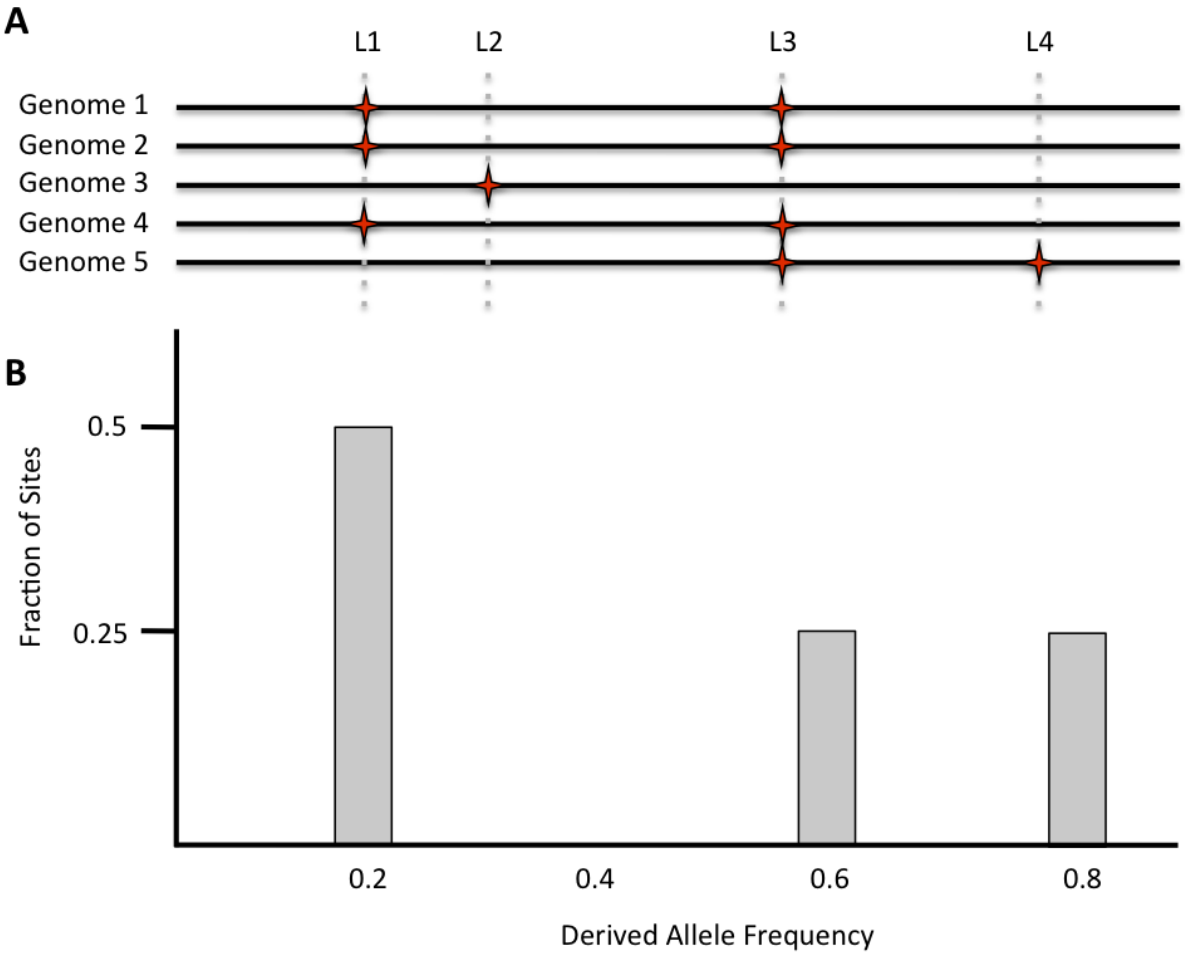


Figure 1.1: A simple example of the site frequency spectrum (*sfs*). In this figure, (A) gives five genomes and four polymorphic sites and (B) gives the *sfs* associated with (A).

ulation. Many analytical approaches have been used to characterize the *sfs* for populations that have undergone different types of events, and two general frameworks of particular use are classical diffusion theory [57] and coalescent theory [68]. These studies [57, 68, 23, 71, 12] have developed a set of analytic theory that can determine the expected *sfs* given different demographic scenarios. With these analytical results, the best demographic model to fit the observed data can be determined using a maximum likelihood framework. Inferences have been expanded to include multiple populations, adding greater dimensionality to the *sfs* and forming the joint frequency spectrum (*jfs*), which gives the relative abundances of mutations within two or more populations.

The *sfs* is informative for both selection and demographic history, and several methods have been developed to use the *sfs* to clarify selection intensities [5, 14, 35] or episodes

of population size change [45, 64, 23] in sampled populations. These two forces often have similar effects, confounding understanding of the true events causing the distortion in the *sfs*. For instance, both negative selection against slightly deleterious mutations and and recent population growth can lead to greater numbers of rare mutations than would be expected [71]. Thus, when testing for selection, one must consider past demographic events, and when one is studying past demographic history, it is important to consider the relative neutrality of the analyzed sites.

While studying the *sfs* directly provides much information on evolutionary history, efforts to selectively use sites with certain criteria or transformations of the *sfs* can provide greater information not observable with only the regular *sfs*. For instance, choosing to group sites by function and determining the *sfs* for each of these groups can help to distinguish signatures in the *sfs* from selection rather than demographic events. Williamson et al. [71] developed a method that could estimate both demographic and selection parameters by deliberately dividing sites into two or more categories. Knowing that some sites, such as non-synonymous sites, have functional consequences potentially causing selection, while other sites are effectively neutral, they divided sites into at least two categories. Once they designated one category as neutral, this set was used to determine the best demographic model to fit the data. Then, they analyzed the other sets, comparing their *sfs* to the expected *sfs* of the best fitting demographic model found for the neutral set. Thus, by categorizing sites based on prior information about associated selective pressures and conditioning the *sfs*, both the demographic and selective forces affecting the data could be determined.

Another category of sites to condition on when creating the *sfs* are those dependant on the allele found in a closely related species. In Chen et al. [8], the close relationship between Neanderthals and humans was used to condition the *sfs*. In 2006, a portion of a Neanderthal specimen's nuclear DNA was sequenced [22], so human genetic variation could be directly compared to the allele present in Neanderthals. Two sets of sites were developed, the first where the Neanderthal site contained a derived allele, and the second where the Neanderthal site contained an ancestral allele. For each of these sets, the *sfs* was determined [8]. Theory was developed to determine what new information could be gleaned from each of these two *sfs*. They found that by separating the data into these two sets, they had a better understanding of the allele age and when changes in allele frequency occurred within the human lineage [8].

In Chapter 2, we take the work of Chen et al. [8] one step further. In the publication of the draft Neanderthal genome in 2010 [21], a test statistic known as the D-statistic was developed, which showed low levels of admixture into non-African humans from Neanderthals. However, an alternative model with no admixture was proposed that also gave the same D-statistic as for admixture [11]. We used the conditional *sfs* for sites where the Neanderthal allele is derived and added a condition that the site was ancestral in an African genome to distinguish between models including and not including admixture. We called this *sfs* the doubly conditioned frequency spectrum (*dcfs*) [75]. By focusing only on sites that are potentially informative for admixture, we enriched for sites that can determine whether a model of genetic admixture is reasonable.

Greater amounts of information can be retrieved when two or more populations are studied. For a single population, the *sfs* can be characterized by a one-dimensional array depicting the abundances across a range of allele frequency categories. When more than one population is studied, an extra dimension is added for each new population. Thus, for two populations, a 2x2 matrix of data is generated, depicting the abundance of sites with x_1 derived allele frequency in the first population and x_2 derived allele frequency in the second population. The utility of the *jfs* has been demonstrated in several studies [8, 24].

In 2009, Gutenkunst et al. developed the software *dadi* (Diffusion Approximation for Demographic Inference), which uses diffusion theory and numerical approximations to simulate the *jfs* under different population models and estimate parameters using a maximum likelihood framework [24]. Their software is very useful in estimating parameters of demographic history in one to three populations, such as population divergence times, population size changes, and admixture.

In Chapter 3, we studied a transformation of the *jfs* that would allow comparison of one population to another. That is, given the frequency spectrum of one population, how different would that of a related population look? We asked how different a genome was from the expected frequency spectrum for a reference population. If the genome was a member of the reference population, then there is no deviation from the expected *sfs*. However, if the genome is not a member of the reference population, then the deviation from the expected *sfs* depends on the demographic relationship between the reference population and the population from which the genome was sampled. We call this method projection analysis [73, 74] and simulated multiple different demographic scenarios to explore how the projection could vary.

In Chapters 4 and 5, we applied the projection method to two different groups: humans and horses. In Chapter 4, we applied the method to modern human populations and several ancient genomes to develop a demographic model that fits well to the observed projections [73, 74]. In Chapter 5, we compared domesticated horses to the Przewalski's horse, studying population size changes and admixture between the two horse populations [56]. We also compared four ancient horse genomes from the Yakut region to present day horses [41], including modern Yakut horse samples.

Both methods used here are enhanced with the addition of ancient genomic data. Since 2010, when a draft nuclear genome of both a Neanderthal and ancient human was published [21], there have been great advances in the field of ancient genomics. Several ancient hominin genomes from Europe [33, 61, 60, 51, 39, 26], central Asia [55, 25, 54, 43, 15, 16, 49], the Americas [53, 50], Australia [52] and Africa [18] have been sequenced and compared to present day populations.

Both methods [75, 73] can be applied to solely modern genomic data, but are much more useful with the addition of ancient data. The *dcfs* for humans requires understanding of the allele in the source population of admixture, which in this case was Neanderthals. The projection analysis compares a single genome to a reference population, which is ideal when studying ancient populations, as most of the data for ancient populations is represented by a single ancient genome.

Conditioning or transforming the *sfs* can provide novel ways of studying a data set, bringing to light new information or lending support to particular hypotheses. Here, we illustrate two methods of manipulating the *sfs* that have led to greater understanding of demographic history, in humans and horses.

Chapter 2

The doubly conditioned frequency spectrum and applications

2.1 Introduction

The interaction between Neanderthals and early modern humans has been a long-standing question in human evolutionary studies. Specimens with morphological traits typical of Neanderthals have been found across Eurasia, from southwest Spain in Europe to southern Siberia in Asia. The first appearance of these traits is as early as 400 kya, and they persist until about 30 kya [36]. Some paleoanthropologists have argued for interbreeding between Neanderthals and early modern humans, using finds such as the child found in Lagar Velho, Portugal that shows a mixture of Neanderthal and early modern human skeletal characteristics [10]. Critics, however, have been skeptical that these finds really suggest potential admixture [65]. Genetic evidence using mitochondrial Neanderthal DNA has consistently shown Neanderthals falling outside the range of modern human variation [37, 59, 20]. This result was taken by many to mean no interbreeding occurred between Neanderthals and early modern humans, although Nordborg [46] showed that low levels of admixture could not be excluded by the mtDNA data.

Green et al. [21] sequenced the first draft of the Neanderthal genome and presented evidence from genomic data that present day non-African human populations share more genetic variants with Neanderthals than did modern African human populations represented by Yorubans. Part of their evidence was based on a 4-taxon statistic, called the D-statistic [21, 55, 11]. The D-statistic quantifies the excess sharing of derived sites between the Neanderthal and any two modern human populations. A nonzero value of D indicates that one of the modern human populations is more similar to the Neanderthal than is the other. Green et al. [21] found that D-statistics indicated greater similarity between Neanderthals and non-African populations than between Neanderthals and African populations. Green et al. [21] proposed a model in which 1-4% of non-African genomes result from admixture from Neanderthals into the ancestors of non-African populations after the separation of Africans

from non-Africans. These results imply that Neanderthals and early modern humans did interbreed. However, recent admixture is not the only hypothesis consistent with the observations. Substructure in early hominin populations in Africa could produce the same patterns [63, 11].

The ancient substructure in Africa model posits that there were two or more subpopulations of hominins in Africa with limited gene flow. Then, ancestors of Neanderthals emigrated from the same subpopulation from which the ancestors of present day non-Africans later emigrated. As a consequence, non-Africans would be slightly more genetically similar to Neanderthals than would Africans because of their more recent common ancestry. After the ancestors of Neanderthals emigrated, the gene flow between the ancestors of present-day Africans and non-Africans would be sufficiently high until the out-of-Africa event, thus making the Africans and non-Africans more genetically similar to one another than either is to Neanderthals. In this model, no later interbreeding between Neanderthals and early modern humans occurred. Durand et al. [11] showed that both models could account for the greater similarity of non-Africans than Africans to Neanderthals.

The model of ancient substructure is consistent with other recent studies. Plagnol and Wall [48], Barreiro et al. [3] and Hayakawa et al. [28] have all suggested the possibility of ancient structure in Africa. Plagnol and Wall showed that non-Africans may have arisen from a Western African subpopulation [48], while Barreiro et al. [3] and Hayakawa et al. [28] observed deep lineages in some genes that seem best explained by ancestral structure in Africa. It is important to be able to distinguish whether the genetic similarity observed by Green et al. [21] is due to recent admixture or ancient structure in Africa.

To distinguish between these two models, we develop here a new approach that relies on the site-frequency spectrum (*sfs*). Durand et al. [11] suggested that the ancient structure model results in more variation in gene tree depth than the recent admixture model. Greater variance in tree depth would alter the frequency spectrum but not the D-statistic. Here, we show that the *sfs* appropriately conditioned can distinguish between recent admixture and ancient structure because it is particularly sensitive to episodes of recent admixture. We construct the *sfs* for non-Africans, conditioning on sites that have the derived allele in the Neanderthal draft genome and the ancestral allele in one randomly sampled African chromosome. This doubly conditioned frequency spectrum (*dcfs*) is enriched for sites in non-African sequences that are Neanderthal- specific. Similarly to the D-statistic, the sites explored are shared derived sites between Neanderthals and non-African humans and are likely to be informative about a recent admixture event.

We derive the analytical expression of the *dcfs* in non-Africans for a null model with no gene flow and compare a series of simulated *dcfs* for demographic models of both recent admixture and ancient structure. In the simulations we allow for a variety of demographic histories, including bottlenecks in population size, ongoing gene flow between present-day human populations, population growth in early humans, varying admixture rates, and different rates of ancient gene flow. The shape of the *dcfs* are observed for each parameter set and compared with the observed non-African *dcfs*.

The observed *dcfs* are computed using four modern human populations from the Com-

plete Genomics Diversity Panel (CGDP) [9] and the draft sequence of the Neanderthal genome [21]. Following Green et al. [21], we chose the Yoruba population (YRI), the Utah residents with European ancestry (CEU), the Japanese (JPT), and the Han Chinese (CHB) from the CGDP to represent the African, the European, and the Asian populations. The Yoruba population is not representative of all the African populations, as current African populations are very diverse [7, 67]. However, the CGDP has several individuals of Yoruba ancestry and like most Africans, the Yoruba population probably had no interactions with Neanderthals. Using these populations, we assess whether the *dcfs* better supports a demographic history of recent admixture or ancient structure.

2.2 Materials and Methods

Complete Genomics Diversity Panel and data processing

The CGDP data we used consists of 46 individuals from nine populations whose genomes were sequenced to an average 45-fold coverage [9]. We used five CEU, four CHB, four JPT, and seven YRI individuals downloaded from the Complete Genomics, Inc. ftp site (ftp://ftp2.completegenomics.com/) in May 2011. Each individual genome was aligned with the reference human genome, hg18. We used the release of the dataset version 1.2 (Software version 1.10, File Format version 1.5).

The Neanderthal sequence was obtained by pooling reads from the three Vindija bones (SLVi33.16, SLVi33.25 and SLVi33.26) that were aligned to hg18 [21]. The Neanderthal data were downloaded from the UCSC genome browser (http://genome.ucsc.edu/Neandertal/) in May 2011. Following Green et al. [21], we used only sites with a mapping quality score of at least 90, a sequence quality higher than 40, and a coverage of at most two, since the average coverage of the draft genome was 1.3.

We assumed that the ancestral state at each of the sites is the reconstructed common ancestor as given in the 1000 Genomes project. This reconstruction is based on an alignment of four species: *Homo sapiens* (human), *Pan troglodytes* (chimp), *Pongo pygmaeus* (orangutan), and *Macaca mulatta* (rhesus macaque). The reconstructed common ancestor was downloaded from the 1000 Genomes ftp site (ftp://ftp-trace.ncbi.nih.gov/1000genomes) in May 2011. Only sites where the human-chimp ancestral sequence, the chimp sequence, and the human-chimp- orangutan ancestral sequence agree were kept. From the merged set of data, we filtered out transition substitutions to reduce the number of sites that are the result of ancient DNA damage [6] and removed all sites that had three or more alleles.

Recent Admixture model

In this model, we assume that there was a single episode of admixture at time t_{GF} in the past ($t = 0$ being the present) from Neanderthals to non-Africans after the migration of humans out of Africa (Figure 2.1A). With probability f, a non-African lineage was derived from a

Neanderthal lineage. The parameter f represents the fraction of the non-African genomes of Neanderthal origin. We define the divergence time of non-African and African populations as $t_H > t_{GF}$. We denote by $t_N > t_H$ the divergence time of Neanderthals and the population ancestral to modern humans. All the populations are assumed to be panmictic. Green et al. [21] proposed this model as the most parsimonious explanation for why Neanderthals share more genetic variants with non-Africans than with Africans.

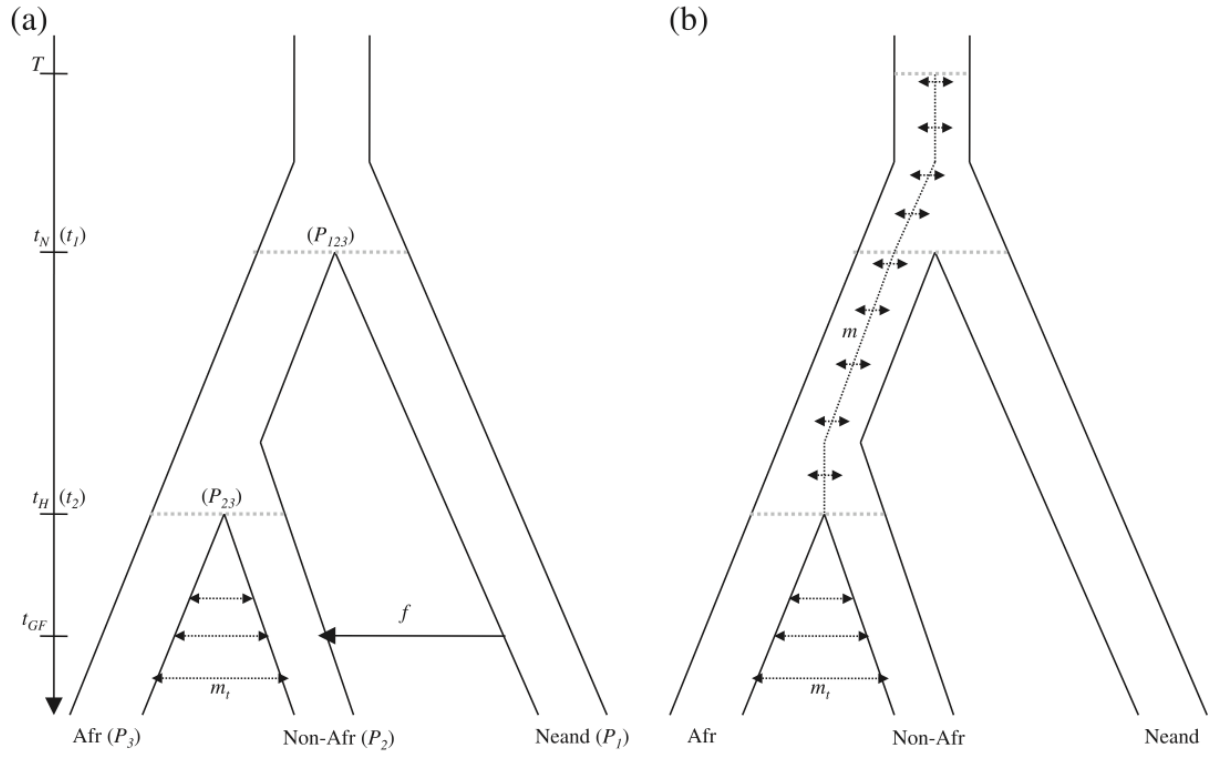


Figure 2.1: Demographic models relating Neanderthals and modern humans.

At time t_H in the past ($t = 0$ being the present), non-African humans (Non-Afr) split from African humans (Afr). After they split, Non-Afr and Afr may exchange migrants at rate m_t. At time t_N, Neanderthals (Neand) split from the ancestors of modern humans. (a) Recent admixture model. At time $t_{GF}|t_H$, Neand admixed with non-Afr at rate f. The variables used in the derivation of the dcfs are shown in parantheses. (b) Ancient structure model. The ancestral population of modern humans is structured in two subpopulations exchanging migrants at rate m. The substructure continues in the ancestral population of Neanderthals and modern humans.

Ancient Structure model

We assume in this model that the population ancestral to modern humans and Neanderthals was divided into two randomly mating subpopulations (Figure 2.1B). We assume that sub-

populations exchanged migrants symmetrically at rate m per generation. At time T in the past, subpopulations merged into one panmictic population. A similar model was proposed in Slatkin and Pollack [63]. Green et al. [21] noted that this model could explain the extra similarity of Neanderthals with non-Africans, and Durand et al. [11] showed that D statistics could not distinguish between ancient structure and recent admixture for plausible demographic parameter ranges.

Analytical Expression for the *dcfs*

We detail the coalescent derivation of the dcfs under a three-population tree model with no gene flow. We denote P_1, P_2, and P_3, three closely related populations, which correspond to Neanderthals, non-Africans, and Africans, respectively (Figure 2.1A). Going backward in time, populations P_2 and P_3 split at time t_2 from ancestral population P_{23}. Population P_1 splits from population P_{23} at time t_1. We denote P_{123} the population ancestral to P_1 and P_{23}. Using an infinite sites model, we assume we have sampled one chromosome from P_1 and P_3 and n chromosomes from P_2. Using diffusion theory, Chen et al. [8] showed that the frequency spectrum in P_2 conditioned on a chromosome from P_1 carrying the derived allele was uniform. For completeness, we rederive this result using coalescent theory (Appendix A).

We generalize the method described in Appendix A to derive the frequency spectrum in P_2 given that a chromosome in P_1 carries the derived allele and a chromosome in P_3 carries the ancestral allele. This is the doubly conditioned frequency spectrum (*dcfs*).

Assume we have sampled n chromosomes in P_2, one in P_1, and one in P_3, P_2 and P_3 split at time t_2 and P_1 splits from P_{23} at time t_1. At time t_2, there are k lineages from P_2 and one from P_3. As shown in Appendix A, the probability that m of these $k+1$ lineages and the lineage in P_1 carry the derived allele is uniform on m:

$$\text{Prob}(m, P_1 = \text{der}|k) = \theta' \tag{2.1}$$

where θ' is a constant that depends on t_1, t_2, n, and the effective ancestral population sizes. The probability that $j = m$ derived lineages are among the k lineages from P_2 and that the lineage from P_3 has the ancestral state is

$$\text{Prob}(j, P_1 - \text{der}, P_3 - \text{anc}|k) = \theta'\left(1 - \frac{m}{k+1}\right) \tag{2.2}$$

Then, the probability that i chromosomes in P_2 and the chromosome in P_1 carry the derived allele and the chromosome in P_3 carries the ancestral allele is

$$\text{Prob}(i, P_1 - \text{der}, P_3 - \text{anc}|j, k, n) = \theta'\left(1 - \frac{j}{k+1}\right)\frac{k-1}{n-1}H(n-2, i-1, k-2)(j-1) \tag{2.3}$$

where $H(n, i, k)$ is the hypergeometric distribution with parameters (n, i, k). Summing over j, we obtain

$$\text{Prob}(i, P_1 - \text{der}, P_3 - \text{anc}|k, n) = \theta' \frac{k-1}{n-1}\left(1 - \frac{n-k}{(k+1)(n-2)}\right) - \theta' \frac{(k-1)(k-2)}{(k+1)(n-2)} i \quad (2.4)$$

The exact *dcfs* is obtained by averaging over k. It can be written as

$$dcfs(i) = \text{Prob}(i, P_1 - \text{der}, P_3 - \text{anc}|n) = \alpha' - \beta' i \quad (2.5)$$

where α' and β' are positive constants that depend on t_1, t_2, n, and effective population sizes in P_{23} and P_{123}. Another *dcfs* of interest is the one obtained when conditioning on a chromosome from P_1 carrying the ancestral allele and a chromosome from P_3 carrying the derived allele, therefore changing the order of conditioning in the previous derivation. At time t_2, the probability that m out of k lineages carry the derived allele and that one chromosome from P_1 carries the ancestral allele is

$$\text{Prob}(m, P_1 - \text{anc}) = \frac{\theta}{m} - \theta' \quad (2.6)$$

Thus, the probability that $j = m - 1$ lineages from P_2 and the lineage from P_3 carry the derived allele and the lineage from P_1 carries the ancestral allele is

$$\text{Prob}(j, P_1 - \text{anc}, P_3 - \text{der}|k) = \left(\frac{\theta}{j+1} - \theta'\right)\frac{j+1}{k+1} \quad (2.7)$$

Using the same derivation as before, we find again that

$$dcfs(i) = \text{Prob}(i, P_1 - \text{anc}, P_3 - \text{der}|n) = \alpha" - \beta" i \quad (2.8)$$

where α" and β" are positive constants that depend on t_1, t_2, n, and effective population sizes in P_{23} and P_{123}. It is important to note that equations 2.5 and 2.8 are valid regardless of the population size histories (bottlenecks, growth, etc.) of P_1, P_2, P_3, P_{23}, and P_{123}.

We assume we have a sample of n chromosomes from a non-African population and one chromosome each from a Neanderthal and an African population. Assuming the chimp has the ancestral allele, we are concerned with the frequency of the derived allele in non-Africans. In particular, we denote $dcfs(x)$ the expected frequency the derived allele appears x number of times in the non-African population, conditioned on the Neanderthal chromosome carrying the derived allele and the African chromosome carrying the ancestral allele. We showed in equation 2.5 that under a model with no recent admixture or ancient structure, the $dcfs(x)$ is given by

$$dcfs(x) = \alpha - \beta x \quad (2.9)$$

where α and β are positive constants that depend on the effective population sizes and population divergence times. This result is also valid for the ancient structure model with no ancestral gene flow ($m = 0$). Such a model is equivalent to a model with no gene flow

but in which the African population is the outgroup to Neanderthals and non-Africans. Equation 2.9 is also valid for the ancient structure model with large m because such a model is equivalent to the null model of no gene flow between Neanderthals and modern humans.

Parameter	Recent Admixture	Ancient Structure
θ ($4N\mu$)	20	20
Admixture rate (f)	{0.02, 0.03, 0.05, 0.1}	NA
Admixture time (t_{GF})	0.05	NA
African/Non-African coalescence time (t_H)	0.1125	0.1125
Neanderthal/human coalescence time (t_N)	0.3	0.3
Ancient migration time (T)	NA	{0.3, 0.35, . . . , 0.8}
Ancient migration rate ($4Nm$)	NA	{0,1, . . . ,10}
Bottleneck time (t_b)	{0.03, 0.1}	{0.03, 0.1}
Bottleneck effect (b)	0.01	0.01
Ongoing gene flow ($4Nm_t$)	{0, 1, 5}	{0, 1, 5}
Time of population growth (t_g)	0.1150	0.1150
Amount of population growth (g)	100	100

NOTE.—"μ" refers to the neutral mutation rate for the entire locus.

Table 2.1: Parameters Used in *ms* [30] for the Recent Admixture and Ancient Structure Models.

Calculating the *dcfs* from the CGDP Data

In order to build the *dcfs*, at each site, we sampled one read at random for the Neanderthal and called it ancestral if it matched the reconstructed ancestor or derived if it did not. Similarly for the Yoruba, at each position, we sampled one chromosome at random and compared it to the reconstructed ancestor. We then counted the number of derived alleles in the ten CEU chromosomes, the eight CHB chromosomes, and the eight JPT chromosomes for sites at which the Neanderthal draft genome carried the derived allele and the Yoruba chromosome carried the ancestral allele.

		D Statistic	
Changing Demographic Parameters		Recent Admixture, $f = 0.05$	Ancient Structure, $4Nm = 4$, $T = 0.6$
No ongoing gene flow	No bottleneck	0.0531	0.0489
	Bottleneck later than admixture, $t_{GF} = 0.03$	0.0533	0.0496
	Bottleneck earlier than admixture, $t_{GF} = 0.1$	0.0521	0.0500
Ongoing gene flow, $4Nm = 1$	No bottleneck	0.0528	0.0375
	Bottleneck later than admixture, $t_{GF} = 0.03$	0.0463	0.0401
	Bottleneck earlier than admixture, $t_{GF} = 0.1$	0.0487	0.0413
Ongoing gene flow, $4Nm = 5$	No bottleneck	0.0318	0.0185
	Bottleneck later than admixture, $t_{GF} = 0.03$	0.0365	0.0172
	Bottleneck earlier than admixture, $t_{GF} = 0.1$	0.0342	0.0224

Table 2.2: The Average D Statistic for Each Simulated Demographic History in Figure 2.2.

Simulations

To simulate models of recent admixture and ancient structure (Figure 2.1), samples were generated in the coalescent simulator *ms* [30]. Unless otherwise specified, we assumed an effective population size of $N = 10,000$ for all populations and a generation time of 25 years per generation. In each replicate of both models, the simulated sample consisted of one Neanderthal chromosome, one YRI chromosome, and either eight (CHB or JPT), or ten (CEU) chromosomes of non-African origin.

In the recent admixture model, t_H was set to 4,500 generations ago (112.5 kya, [40]), and t_N was set to 12,000 generations ago (300 kya, [21]). The t_{GF} parameter was set to 2,000 generations (50 kya), and f was chosen to be 0.05.

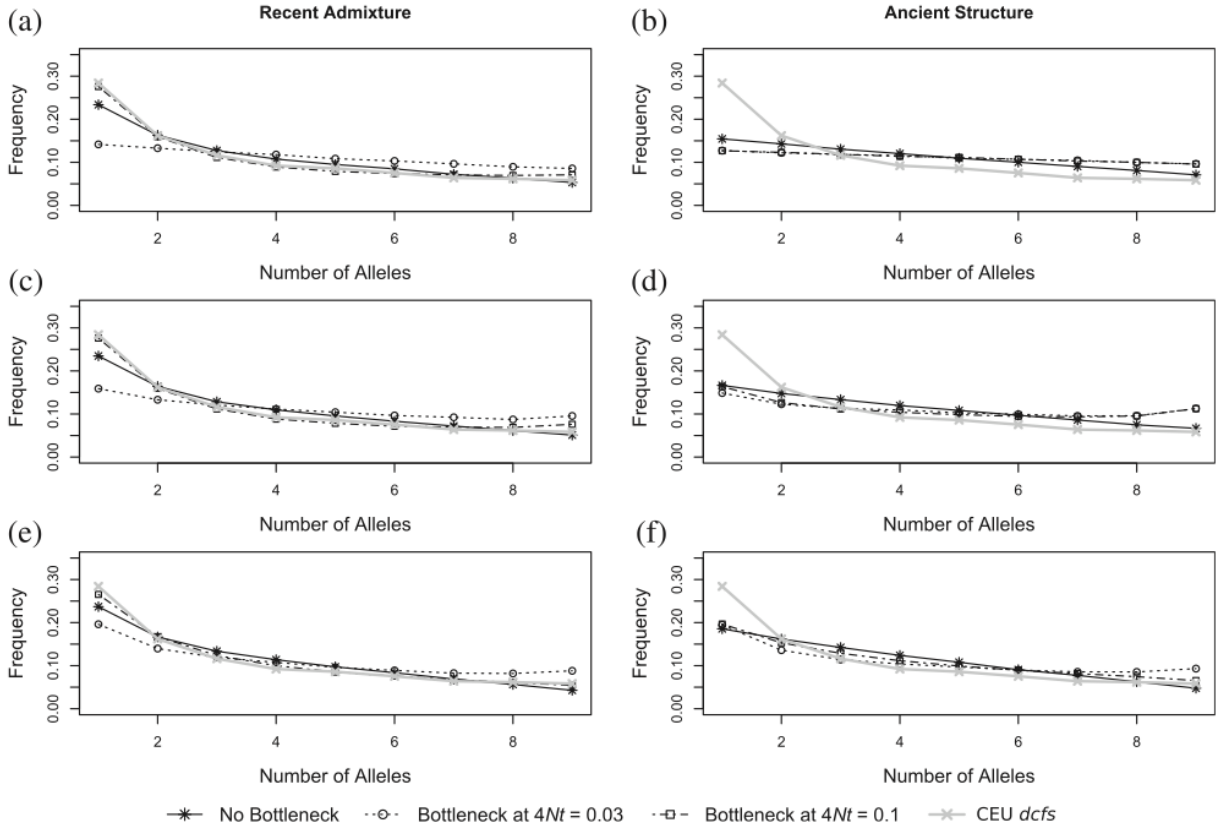


Figure 2.2: The *dcfs* for the CEU population and simulations of recent admixture (a, c, e) and ancient structure (b, d, f).

Parameters used are shown in Table 2.1, including no ongoing gene flow (a, b), ongoing gene flow of $4Nmt = 1$ (c, d), and ongoing gene flow of $4Nmt = 5$ (e, f). The time of admixture in the recent admixture model was $4Nt_{GF} = 0.05$. The admixture rate simulated was 0.05.

In the ancient structure model, T was varied between 12,000 and 32,000 generations ago, in steps of 2,000 generations. The intensity of ancient migration m was set to $4Nm =$

$0, 1, ..., 10$. The non-African and Neanderthal populations split 12,000 generations ago (t_N), and t_H, the population split time between YRI and the non-African populations, was 4,500 generations ago. The parameter values for the ancient structure model were chosen so that the D statistic did not differ significantly from the observed D statistics (Durand et al. 2011).

For each model, a bottleneck reducing the effective size of the non-African populations by a factor of 100 (b) for 100 generations was set to (t_b) 1,200 generations (30 kya) and 4,000 generations ago (100 kya). The two times of the bottleneck allowed it to be either before or after the time of admixture. We also considered ongoing symmetric gene flow between YRI and the non-African populations with rates $4Nmt = 1$ and $4Nmt = 5$. We simulated population growth in humans 100 generations before the YRI and non-African split, increasing the population size 100-fold in humans. We simulated the *dcfs* for other admixture rates of 2%, 3%, and 10% admixture with no bottleneck and a bottleneck younger or older than the time of admixture. The parameter values for *ms* are shown in Table 2.1. The *ms* commands used to generate the simulated sequences can be found in Appendix B.

For each set of parameters, 1 million replicates were run and the non-African *dcfs* were estimated averaging over all the replicates. To allow comparison between different sets of parameters, the *dcfs* were normalized by the number of segregating sites in each simulation. Finally, the shapes of the *dcfs* were characterized for each model for each set of parameters.

To test the effect of misassignment of the ancestral allele on the *dcfs*, we simulated a model with no bottleneck, population growth, or present-day gene flow and incorrectly assigned the ancestral allele for 0.5%, 1%, 5%, and 25% of the analyzed sites.

2.3 Results

Distinguishing Recent Admixture and Ancient Structure

To be consistent with observed D statistics, we chose our parameters so that our simulations yield D statistics between 1% and 10% (Table 2.2). Three main shapes characterized the *dcfs* for the simulated data: linear (Equation 2.8), L- shaped (an excess of singletons compared with the null hypothesis of no gene flow), or U shaped (excess of both rare and common alleles). The exact forms depended on the parameter values.

In the simplest case of no bottleneck or ongoing gene flow between modern humans, the *dcfs* had an L-shape for the recent admixture model simulations with $f = 0.05$ (Figure 2.2A, 2.3A and 2.4A). The *dcfs* for the ancient structure model decreased linearly, regardless of the values of m and T (Figure 2.2B and 2.5B). Adding a bottleneck in the ancient structure model reduced the slope of the line, but the *dcfs* was still linear (Figure 2.2B and 2.5B). Note that this result is the theoretical result for the ancient structure model with very low or very high values of m, and this results holds independently of population size fluctuations (Equation 2.8). For the recent admixture model, the placement of the bottleneck before or

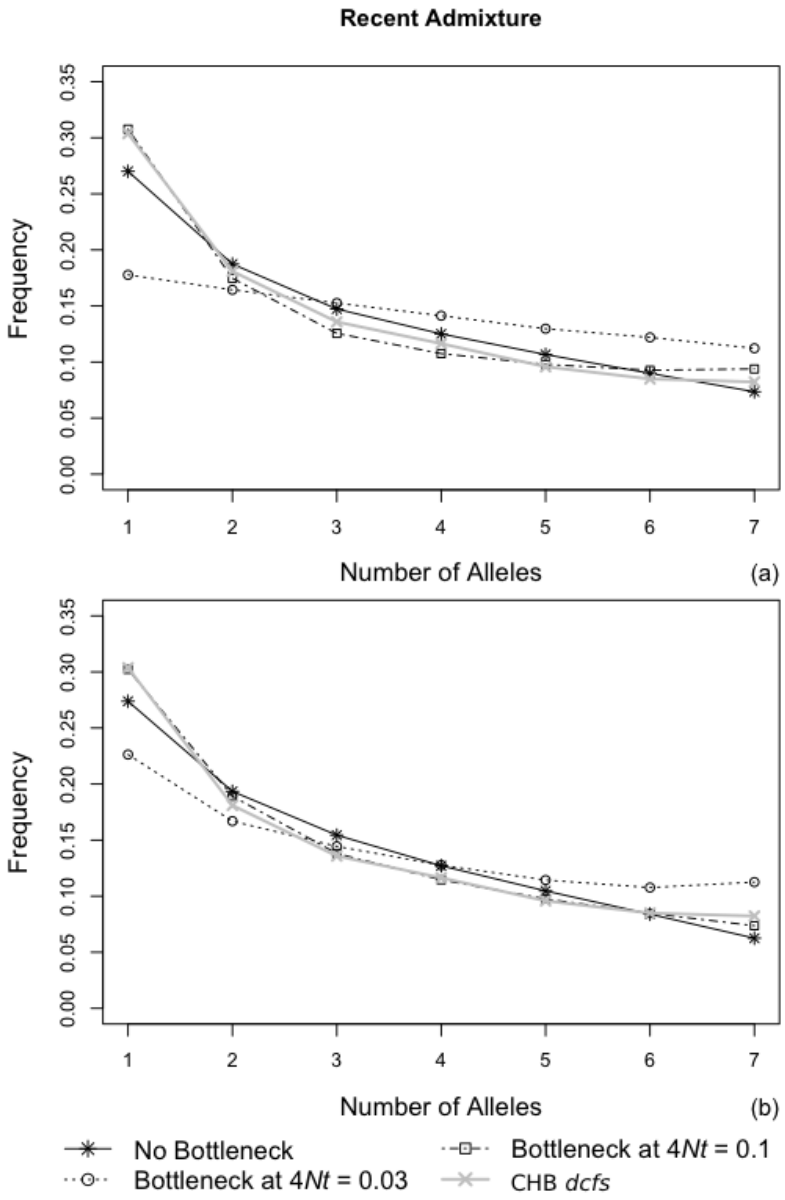


Figure 2.3: The *dcfs* for the CHB population and simulations of recent admixture.

Parameters used are shown in Table 2.1, where the parameters includes no ongoing gene flow (a) and ongoing gene flow of $4Nmt = 5$ (b). The time of admixture is $4Nt_{GF} = 0.05$, the admixture rate simulated is 0.05, and no population growth occurred.

after the admixture event had an effect. When $t_b > t_{GF}$, the *dcfs* was still L-shaped, whereas when $t_b ¡ t_{GF}$, the shape was linear for the *dcfs* (Figure 2.2A).

Including ongoing gene flow between the non-African populations and the YRI population without a bottleneck resulted in a steeper linear *dcfs* for the model of ancient structure and

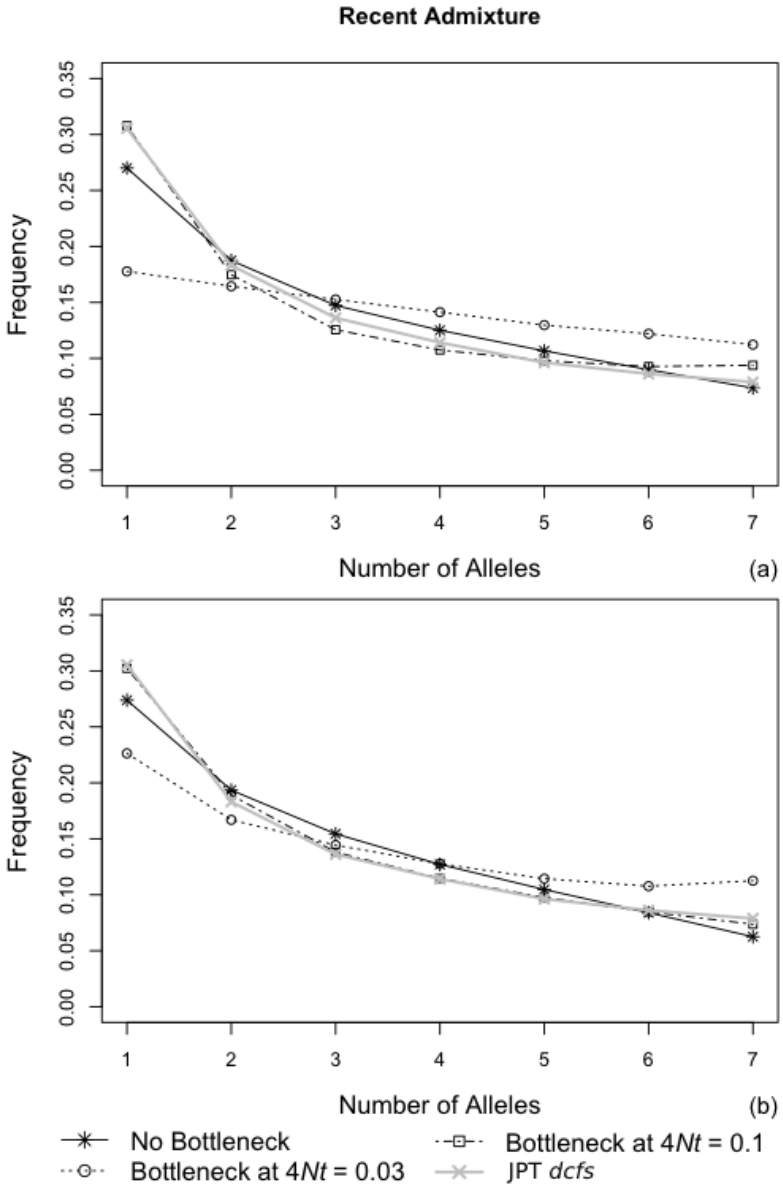


Figure 2.4: The *dcfs* for the JPT population and simulations of recent admixture.

Parameters used are shown in Table 2.1, where the parameters includes no ongoing gene flow (a) and ongoing gene flow of $4Nmt = 5$ (b). The time of admixture is $4Nt_{GF} = 0.05$, the admixture rate simulated is 0.05, and no population growth occurred.

a slightly more pronounced L-shape for a model of recent admixture, with little difference from a model containing no ongoing gene flow (Figure 2.2C-F, 2.3B and 2.4B). Including a bottleneck before or after admixture and ongoing gene flow of $4Nmt = 1$ gave a U-shaped *dcfs* in both models, though when $t_b > t_{GF}$, there was a slightly greater excess of rare alleles

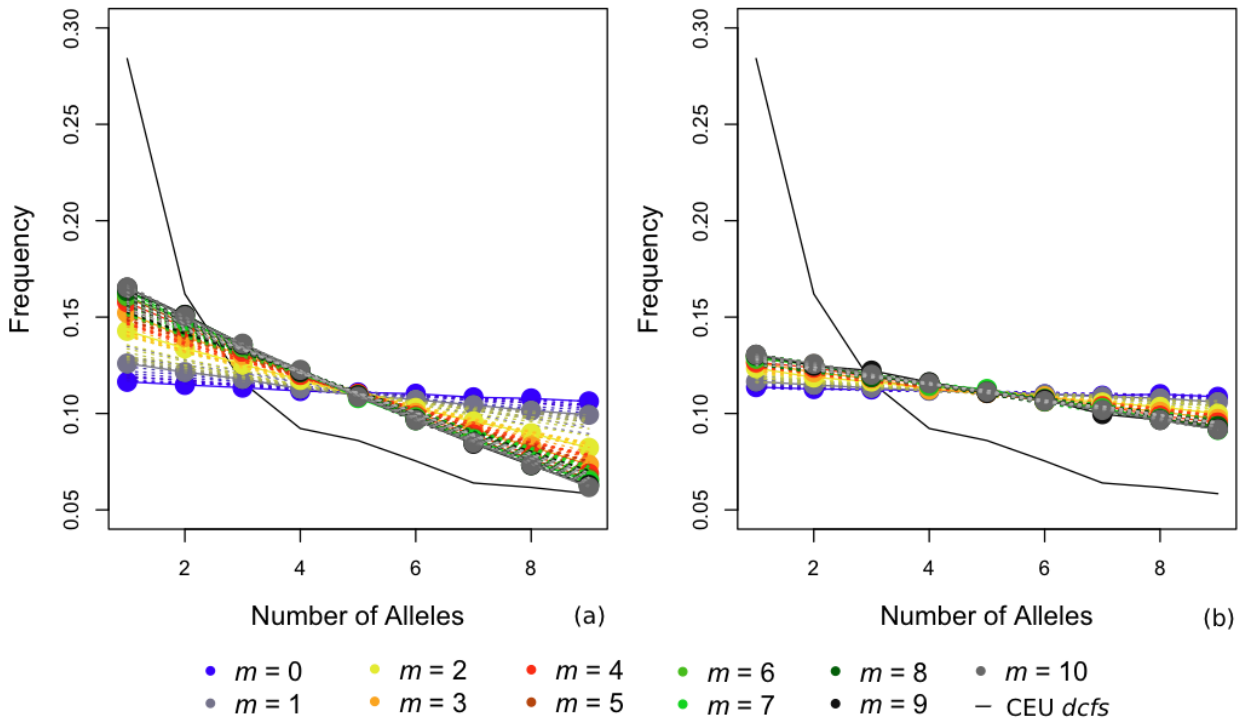


Figure 2.5: The *dcfs* from Europeans from Complete Genomics (black solid line) and simulations of a demographic history including ancient structure and no bottleneck (A) and a bottleneck at $t_b = 0.1$ (B).

121 demographic histories were modeled. The lines with the shaded circles represent each of the migration rates (0 to 10, step size of 1) with the oldest time of population structure at 0.8. The dotted lines represent the different times of population structure (0.3 to 0.75, step size of 0.05), with the slope becoming steeper as T approaches 0.3.

(Figure 2.2C and 2.2D). Increasing the level of ongoing gene flow to $4Nmt = 5$ makes each observed shape steeper, but in both models, when $4Nt_b = 0.1$, the U-shape was lost (Figure 2.2E, 2.2F, 2.3B and 2.4B). The inclusion of population growth in humans still retained the L- shape in the recent admixture model and the linear shape in the ancient structure model. Population growth made little difference in the shape of the curve for the different bottleneck times without increasing the severity of the bottleneck (Figure 2.6 and 2.7).

Misassignment of the ancestral allele did not change the shape of the dcfs for the ancient structure model. As the error rate increased from 0.5% to 25%, the slope was steeper but the ancient structure dcfs remained linear (Figure 2.8).

The observed *dcfs* for the 10 CEU haplotypes from the CGDP had an L-shape (Figure 2.2, gray curve), showing an excess of rare alleles but not higher frequency of sites carrying nearly fixed derived alleles. The shape did not fit any of the simulations for the ancient structure model (Figure 2.2B, 2.2D and 2.2F) for all m and T (Figure 2.5). The *dcfs* for

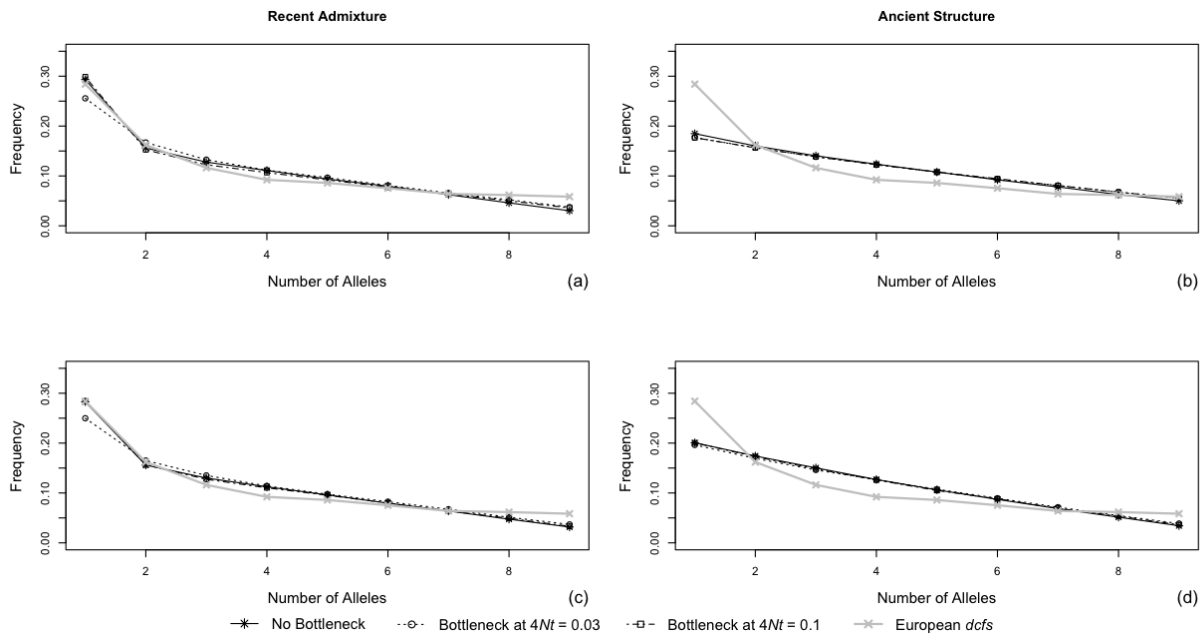


Figure 2.6: The *dcfs* for the CEU population and simulations of recent admixture and ancient structure with a bottleneck of 0.001.

Parameters used are shown in Table 2.1, including no ongoing gene flow (A, B) and ongoing gene flow of $4Nmt = 5$ (C, D). For the recent admixture model (A, C), the time of admixture is $4Nt_{GF} = 0.05$ and the admixture rate simulated is 0.05. For the ancient structure model (B, D), the ancient migration rate is $4Nm = 4$ and the end of ancient migration is $T = 0.6$. Population growth occurred 100 generations prior to the split of Africans and non-Africans, and a bottleneck reduced the population size of non-Africans to 0.001 its previous size.

the eight Han Chinese and eight Japanese chromosomes from the CGDP were also L-shaped and thus were also a better fit for the recent admixture model than the ancient structure model (Figure 2.3 and 2.4).

Using an admixture rate of about 5% provided a much better fit to the observed data when combined with an older bottleneck (Figure 2.9) than any model containing ancient structure only. Because many parameters can affect the fit of the simulated results to the observed data, this does not imply that the true admixture rate is 5%. Rather, the results suggest that scenarios with recent admixture are plausible. Various versions of the recent admixture model gave very good fits to the observed data, which was not the case for the ancient structure model for any set of parameter values we used.

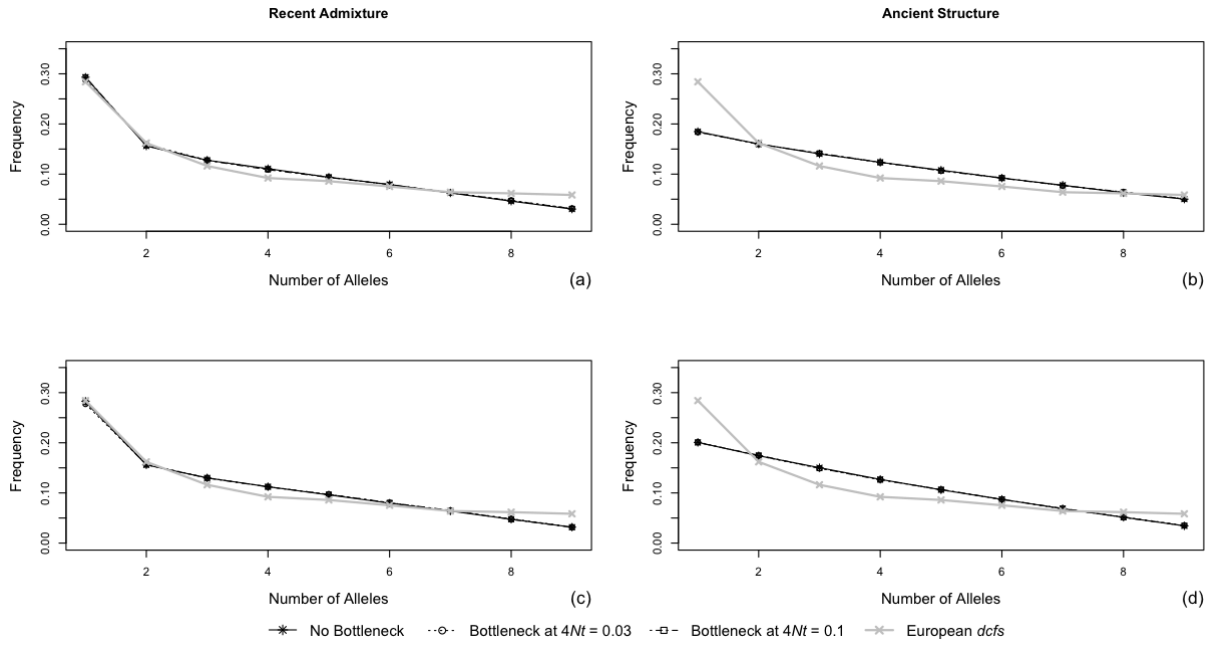


Figure 2.7: The *dcfs* for the CEU population and simulations of recent admixture and ancient structure with a bottleneck of 0.01.

Parameters used are shown in Table 2.1, including no ongoing gene flow (A, B) and ongoing gene flow of $4Nmt = 5$ (C, D). For the recent admixture model (A, C), the time of admixture is $4Nt_{GF} = 0.05$ and the admixture rate simulated is 0.05. For the ancient structure model (B, D), the ancient migration rate is $4Nm = 4$ and the end of ancient migration is $T = 0.6$. Population growth occurred 100 generations prior to the split of Africans and non-Africans, and a bottleneck reduced the population size of non-Africans to 0.01 its previous size.

2.4 Discussion and Conclusion

Support for the Recent Admixture Hypothesis

The *sfs* for the derived European (CEU), Chinese (CHB), and Japanese (JPT) alleles, conditioned on the ancestral YRI allele and the derived Neanderthal allele (*dcfs*), did not have the shape predicted by the ancient structure model for any set of parameters explored. The actual dcfs showed a noticeable excess of rare alleles that was not observed in any of the simulations for ancient structure but was observed in most of the simulations for recent admixture. An excess of rare alleles has been a characteristic feature that many past studies have described. The excess of rare alleles in the dcfs is unlikely to be due to errors in sequencing [31]. Misassignment of the ancestral allele under a model of ancient structure was simulated and does not alter our conclusions. The high coverage of the genomic data from the CGDP makes it highly unlikely that sequencing error in those genomes is large enough to affect our conclusions.

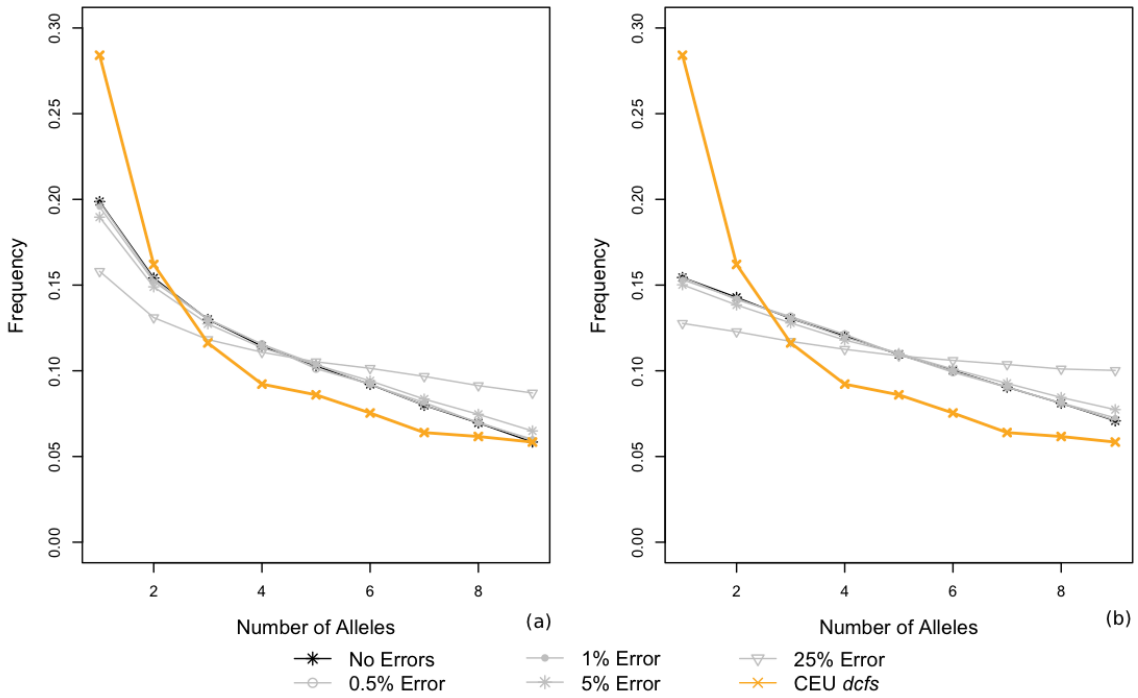


Figure 2.8: The *dcfs* for simulations of recent admixture and ancient structure for no bottleneck or present day gene flow, including misassignment of the ancestral allele from no misassignments to a 25% error rate.

For the recent admixture model (a), the time of admixture is $4Nt_{GF} = 0.05$ and the admixture rate simulated is 0.05. For the ancient structure model (b), the ancient migration rate is $4Nm = 4$ and the end of ancient migration is $T = 0.6$.

In the unconditional *sfs*, an excess of rare alleles can also be due to recurrent selective sweeps [34] and logistic population growth [17]. However, these explanations for an excess of rare alleles is observed when no conditioning is used and do not account for an excess of rare alleles specifically at sites derived in the Neanderthal draft genome and ancestral in the YRI chromosomes. We have shown that population size fluctuations do not affect linearity in the *dcfs* in the absence of gene flow. Thus, the excess of rare alleles in the actual *dcfs* is due to recent admixture and not ancient structure, recent population growth, or recurrent sweeps.

To further understand the recent admixture model, we compared the effect of a bottleneck in non-Africans before or after the admixture occurred. We found that a bottleneck before admixture provides a better fit than a bottleneck after admixture. Although we did not attempt to estimate parameters, we found that in all simulations, the shape of *dcfs* when the bottleneck was before admixture was more similar to the *dcfs* calculated from the data than the shape of the *dcfs* for a younger bottleneck. For the parameters we kept constant,

an admixture rate of about 5% provided a good fit, but this is unlikely to be true for other plausible demographic parameters. For instance, population growth before the divergence between the Yoruba and the non-African populations may make it difficult to distinguish between younger and older bottlenecks. However, we can be confident that the shape of the dcfs is unlikely to change dramatically if parameters remain within a plausible range. The lack of an excess of common alleles in the actual dcfs seem to suggest little, if any, ongoing gene flow between the Yoruba and non-African populations, but this effect may be more difficult to observe, as the excess seen in the simulation is small and is lost with higher rates of ongoing gene flow.

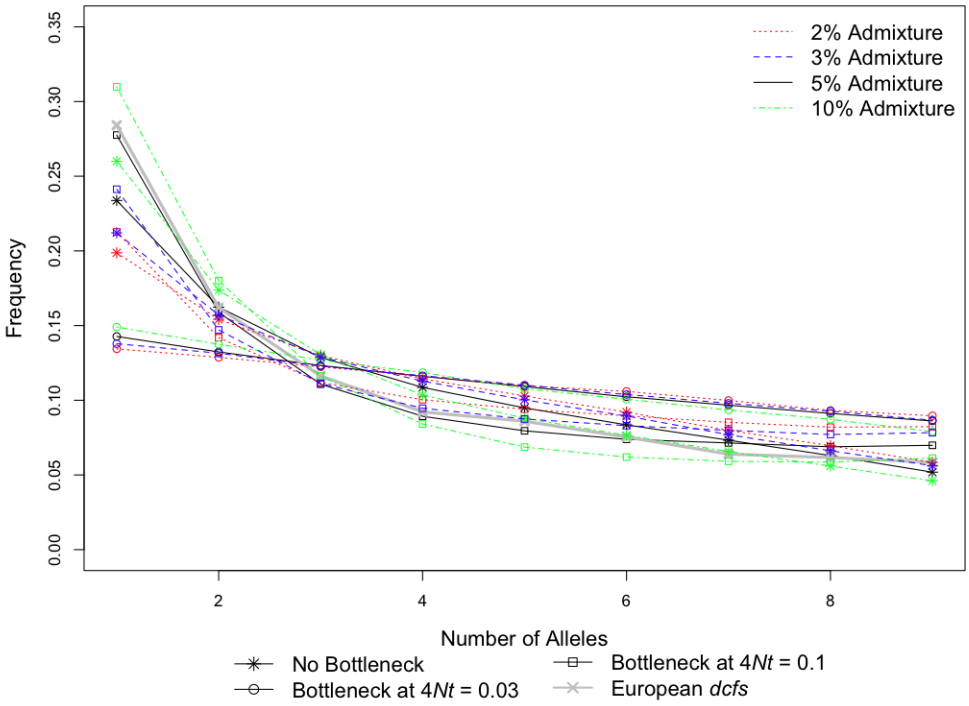


Figure 2.9: The *dcfs* for simulations of recent admixture of 2% (red), 3% (blue), 5% (black), and 10% (green), for no bottleneck (star), a bottleneck at 0.1 (square), and a bottleneck at 0.03 (circle).

The time of admixture is $4Nt_{GF} = 0.05$. The European *dcfs* is in gray, and it is best matched by a 5% admixture rate and a bottleneck at 0.1. With a younger bottleneck at 0.03, there is much less of an increase in rare alleles for higher admixture rates compared to when there is no bottleneck or an older bottleneck.

The simulations showed clear trends in the *dcfs* indicating that although the D statistic cannot distinguish between different demographic models, the *dcfs* can. The differences in the *dcfs* show that the greater variation in gene tree depth in the ancient structure model compared with the recent admixture model can be used to distinguish between these

two models. The data from the Neanderthal genome and the CGDP showed a markedly different *dcfs* compared with any of the *dcfs* simulated for the ancient structure model, suggesting that the model of ancient structure does not accurately approximate the history of humans and Neanderthals. The results, however, do support the hypothesis of recent admixture. Interbreeding between Neanderthals and early modern humans is currently the most parsimonious and plausible explanation for the observed excess of genetic similarity between Neanderthals and non-Africans.

Chapter 3

Projection of a Test Genome onto a Reference Population

3.1 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. [24] 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 [27] 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. [13] 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 [58] recently introduced the multiple sequential Markovian coalescent (MSMC) model, which is a generalization of the pairwise sequential Markovian coalescent model [40]. 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 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 explore the effects of using ancient samples, as both the reference panel and the test genome.

3.2 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 $w = \frac{1}{2x}$; and if it is homozygous derived, then $w = \frac{1}{x}$. The projection $\bar{w}(x)$ 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, $\bar{w}(x) = 1$ independently of x if the test genome is randomly sampled from the reference population. Therefore, deviation of $\bar{w}(x)$ 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 t in the past, and that there has been no admixture between them since that time. The results of Chen et al. [8] show that in this model $\bar{w}(x) = e^{-\frac{\tau}{2N}}$ 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* [24]. Models with more than three populations were simulated using *fastsimcoal2* [13].

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.

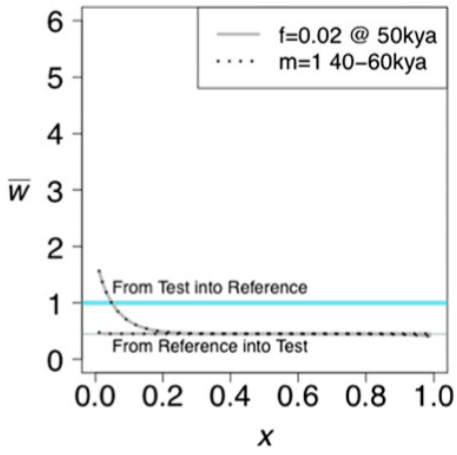


Figure 3.1: The effect of unidirectional geneflow on the projection of a test genome onto a reference population. Two kinds of gene flow were assumed: either a single pulse of admixture of strength f or a period of immigration at a rate m per generation. Both populations are of constant size $n = 10,000$. The divergence time is 400 kya.

Two populations

We first consider two populations of constant size that separated t 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 3.1 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: $\bar{w}(x)$ 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 3.2A, black line) or ancestral population (Figure 3.2A, blue line), but not when there is a bottleneck in the test population (Figure 3.2A, red line). The reason for the humped

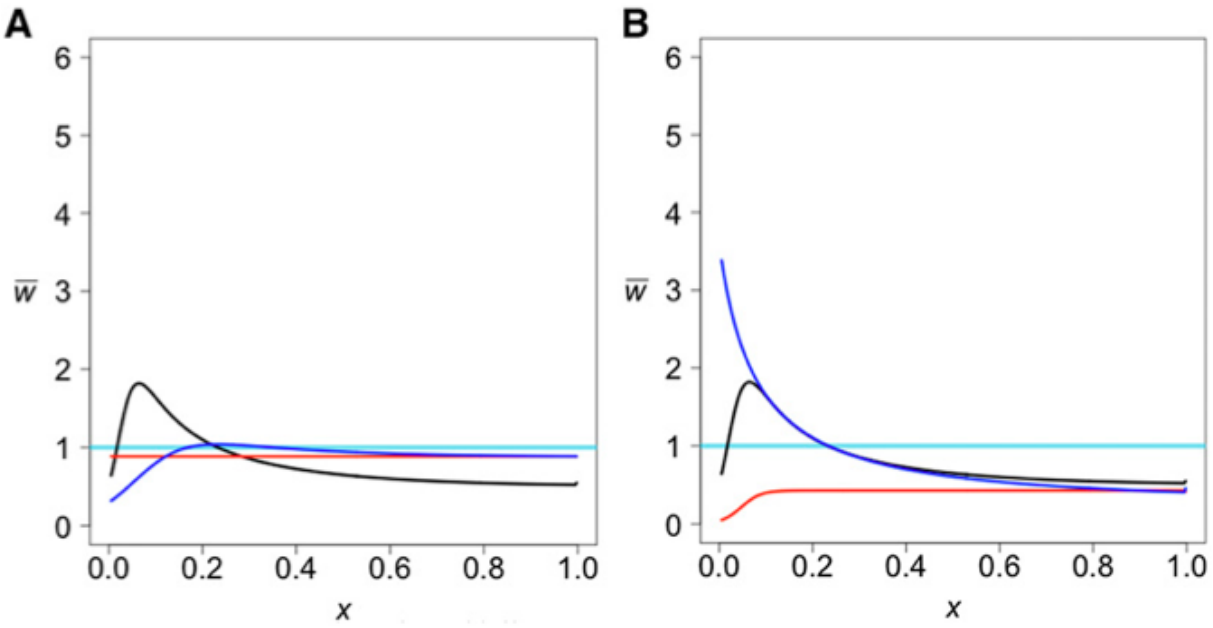


Figure 3.2: The effect of population size changes in a model with two populations that diverged 60 kya.In the figure, (A) A bottleneck occurs in the reference population (black), the test population (red), or the ancestral population (blue). During the bottleneck, the population size is reduced from 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.

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. 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 3.2B, blue line). When the reference population expands, a slight decrease in rare alleles is observed (Figure 3.2B, red line).

A bottleneck followed by admixture amplifies the effect of admixture (Figure 3.3A, black line) while admixture that occurs before or during the bottleneck does not change the shape of the projection as much (Figure 3.3A, 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 3.3B).

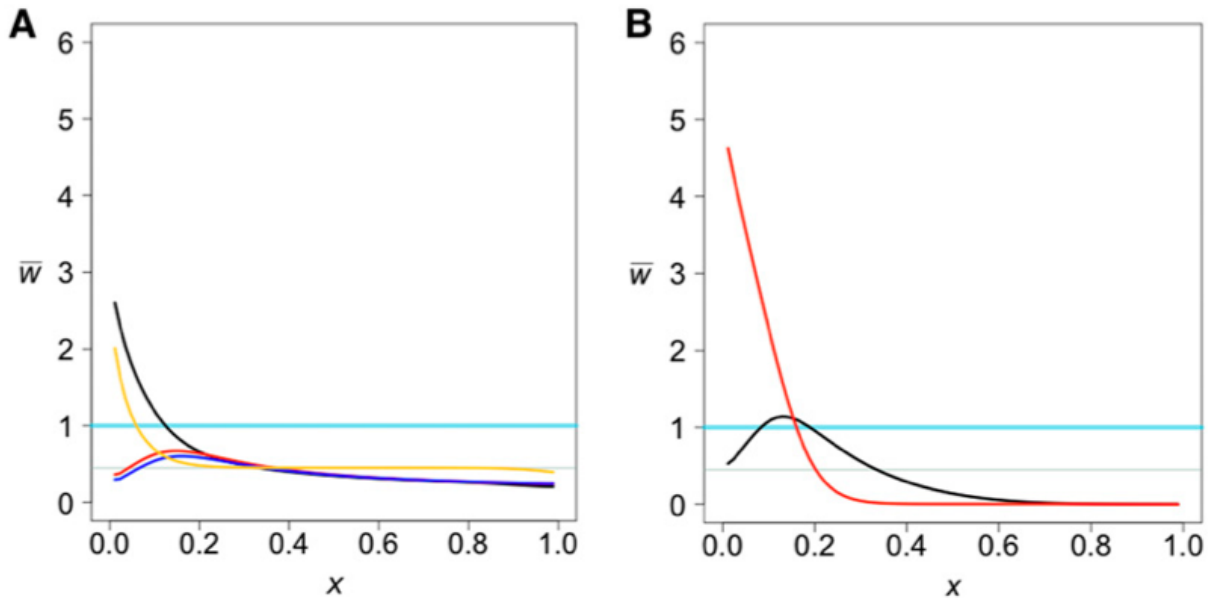


Figure 3.3: The combined effect of a bottleneck and admixture. The divergence time for both models is 100 kya. (A) The yellow projection represents no bottleneck but admixture of $f = 0.02$ at 40 kya. The other projections include admixture at 40 kya (black), 80 kya (red), and 120 kya (blue) of 0.02 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).

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 [4] 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 3.4). 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 3.5A-Figure 3.5C), 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 3.4A and Figure 3.5D), the effect is similar to that of gene flow directly from the test into the reference population (Figure 3.1). The increase of $\bar{w}(x)$ 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 3.5D).

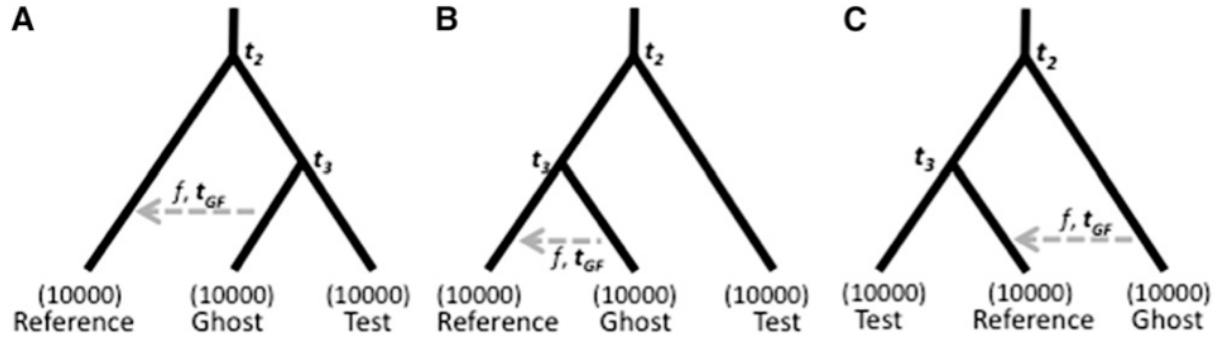


Figure 3.4: Illustration of three possible population relationships in which there is a pulse of admixture of intensity f at time t_{GF} in the past from the ghost population into the reference population. t_2 and t_3 are the times of population separation. In each topology, either the test and ghost (A), reference and ghost (B), or the test and reference (C) are more closely related to each other than the third population.

In the second topology (Figure 3.4B), the reference and ghost populations are sister groups. Here, gene flow from the ghost population also increases $\bar{w}(x)$ 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 $\bar{w}(x)$ 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 3.5E). When the reference and test populations are sister groups (Figure 3.4C), and the ghost population is an outgroup, a dip is observed for low frequencies (Figure 3.5F).

If there is a bottleneck in the reference population after admixture, the effect (Figure 3.6A) is similar to that seen in the two-population case (Figure 3.3). The signal of admixture is amplified. In the case where the reference and ghost populations are sister groups (Figure 3.6B), 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

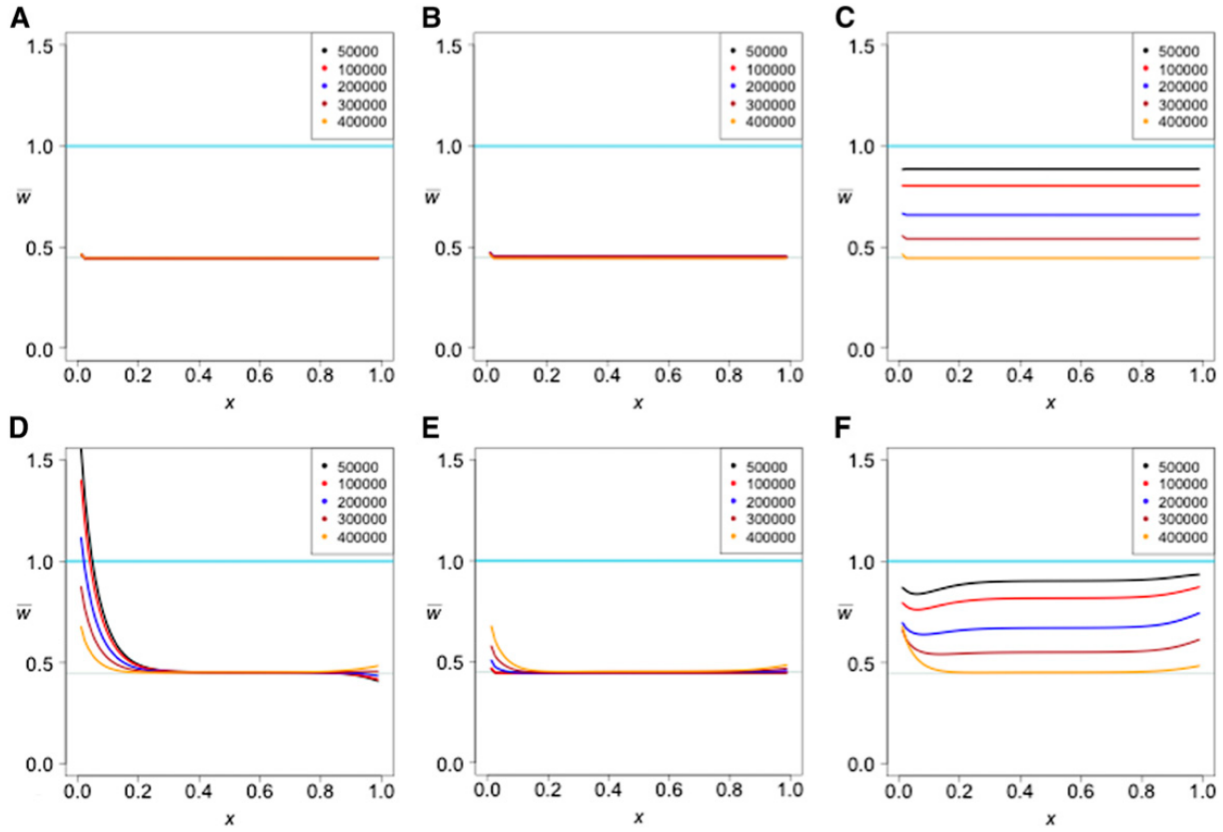


Figure 3.5: The effect of ghost admixture into the test (A-C) and the reference (D-F). A and D follow the topology in Figure 3.4A; B and E follow the topology in Figure 3.4B; and C and F follow the topology in Figure 3.4C. $t_2 = 400$ kya, $f = 0.02$, and $t_{GF} = 50$ kya. t_3 is varied from 50 to 400. Population sizes remain constant at 10,000.

and test populations are sister groups, the humped shape remains, but the effect is reduced as the time of divergence increases (Figure 3.6C), 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. [29] show that ancestral misidentification occurs at levels of approximately 1-5% in human genome data sets. We use *ms* [30] to simulate two simple demographic 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 3.1. We allowed for 0, 0.1, 1, or 10% of the sites to

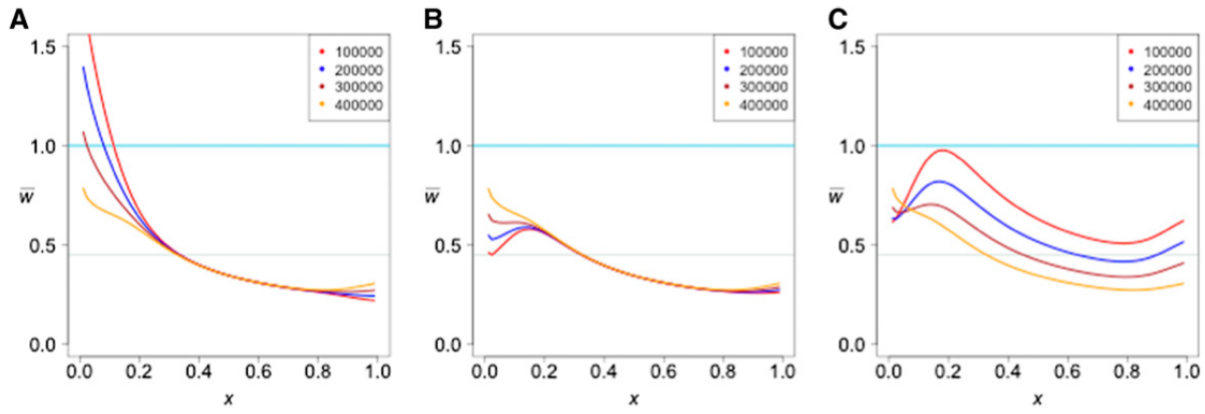


Figure 3.6: The effects of ghost admixture into the reference with a bottleneck in the reference occurring 70100 kya changing the reference population size from 10,000 to 1000 and back to 10,000. t_3 is varied from 100 to 400 kya. All other parameters are the same as in Figure 3.5. A follows the topology in Figure 3.4A, B follows the topology in Figure 3.4B, and C follows the topology in Figure 3.4C.

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 [29], the projection shows a similar result (Figure 3.7).

Simulations of Ancient Samples

To simulate demographic scenarios including ancient samples, we used *fastsimcoal2* (version 2.1, [13]) to model several demographic histories, from which samples were taken to form a reference panel of $n = 200$ and a test genome to project onto the reference panel. For each simulation, we projected an ancient sample onto a modern population or a modern sample onto an ancient population. The ancient samples were taken at 500, 1000, 2000, 3000 and 4000 generations ago (ga). Unless otherwise indicated, the effective population size was 5000. We considered two demographic models: a one-population model (OPM, Figure 3.8, OPM A-E) where the ancient sample was directly ancestral to the present-day population, and a two-population model (TPM, Figure 3.8, TPM A-E) where the ancient sample belongs to a sister population that diverged from the present-day population.

In OPM A, no population size change or migration was applied to the population. In OPM B, we applied a pulse of admixture of 0.05 at 750 ga from an unsampled population into the present-day population. We then allowed a population size expansion from 500 to 5,000 at 750 ga (OPM C), a population size decline from 5,000 to 500 at 750 ga (OPM D), and a bottleneck 500 to 1,000 ga, where the population reduces from 5,000 to 500, before recovering to 5,000 (Figure 3.8, OPM C-E). In the two-population model, the same five scenarios were simulated. Again, we considered no population size changes or migration (TPM A), before

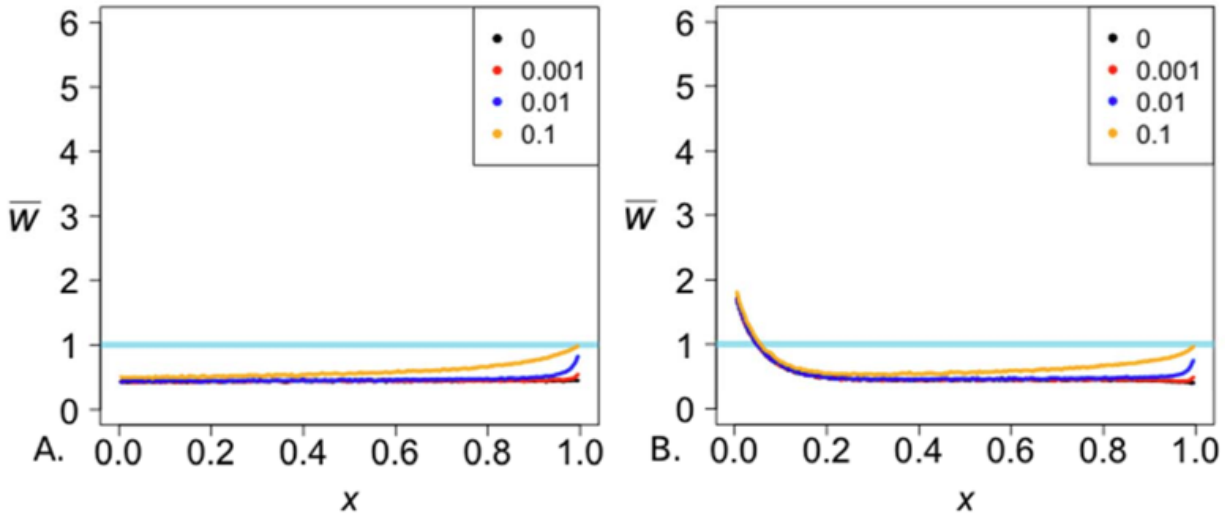


Figure 3.7: Simulated projections with ancestral misidentification. The proportion of sites misidentified is given in the legend. For both models, the population is a constant size of 10,000 and the time of divergence is 400 kya. In part B, there is an additional admixture event of 0.02 from the test to the reference population at 50 kya.

adding migration from the sister population into the present-day population (TPM B). The three population size changes occur only in the present-day population (Figure 3.8, TPM C-E).

For the one-population model, when the reference panel is from the present and the test genome is ancient, the projection's shape does depend on the sampling time (Figure 3.9, top row). In Figure 3.9, the projection of an ancient sample onto a reference panel comprised of members of the descendant population decreases with the age of the sample. When there are no population size changes or migration (Figure 3.9, OPM A), the projection follows the $\bar{w}(x) = e^{-\frac{t}{2N}}$ line, where t is the age of the ancient sample, not the time of population divergence. Small amounts of admixture from an unsampled population have no effect on the projection (Figure 3.9, OPM B). Population size changes show different levels of effect for different sampling times. When there is a population expansion, the projection decreases for small x (Figure 3.9, OPM C), while when there is a population decline, the projection increases for x (Figure 3.9, OPM D). A bottleneck results in a humped shape similar to that observed when the test genome is sampled from a related population that diverged prior to the bottleneck (Figure 3.9, OPM E). Changes in the sampling time results in slight changes in the shape of the projection, but the projection retains the characteristic shape for that type of population size change.

The mirror scenario, where the reference panel consists of ancient samples and the test genome is sampled from the present, looks markedly different (Figure 3.9, bottom row).

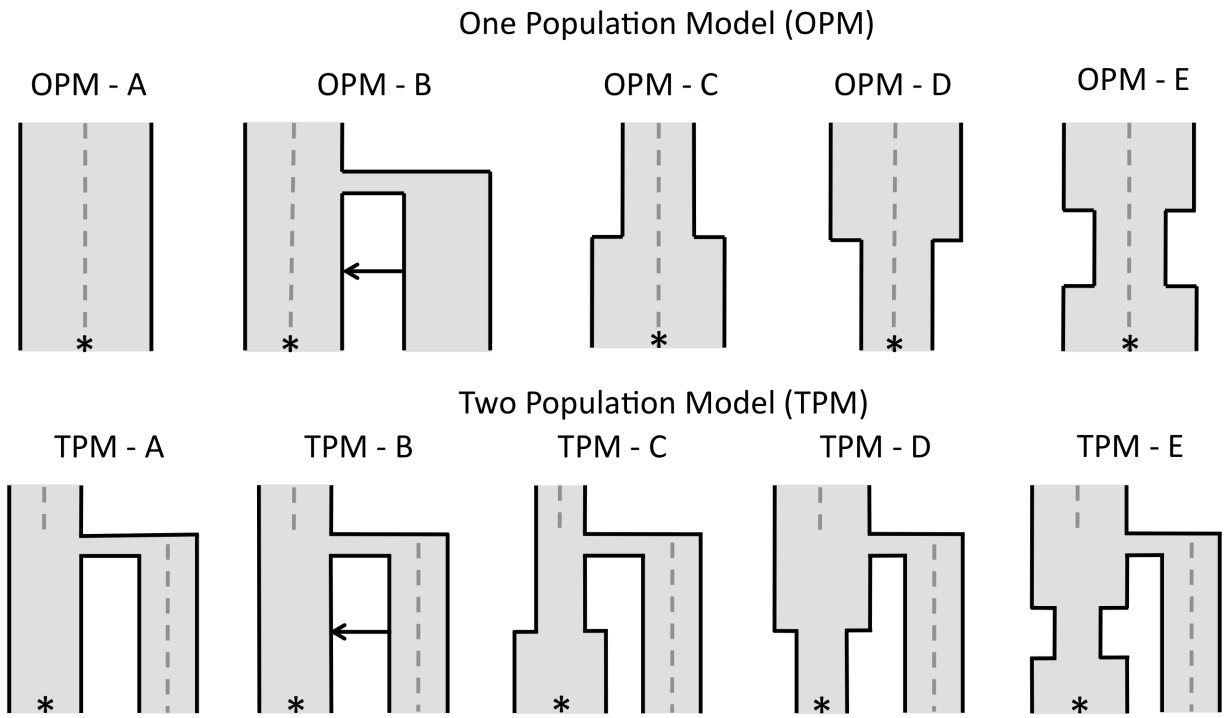


Figure 3.8: Simulated demographic models used to illustrate the effect of ancient samples in a one-population and two-population model.The * represents where the present-day population was sampled and the gray dashed line indicates when the ancient genomes were sampled (0 - 4k gen). Any divergence occurs 2k gen ago. For both OPM and TPM, A has an N_e of 5k, with no population size changes or admixture. B adds a pulse of admixture from the second diverging population. C has no admixture but allows a population size expansion from 500 to 5k in the reference population 750 gen ago. D allows the reverse, a population size decline from 5k to 500 in the reference population 750 gen ago. E has a bottleneck from 5k to 500, 500-1000 gen ago. Any diverging population has the same N_e as the ancestral population.

Here, the present-day test genome looks no different from the ancient population upon which it is projected. This is reasonable because the main contribution to deviations in the projection from is from new mutations in the reference population that are not found in the test population. When the reference panel is made up of ancient samples, there are no new mutations in the reference population that are not also in the present-day population from which the test genome is sampled. Thus, using an ancient reference panel and a test genome from the descendant population will not give insight into the demographic changes that the population has undergone between the time of sampling and the present-day.

In the two-population model, the results for the projection are very different than that found for the one-population model. The simplest scenario (Figure 3.10, TPM A) highlights a difference in the projection relative to OPM A (Figure 3.9). In TPM A, the projection is

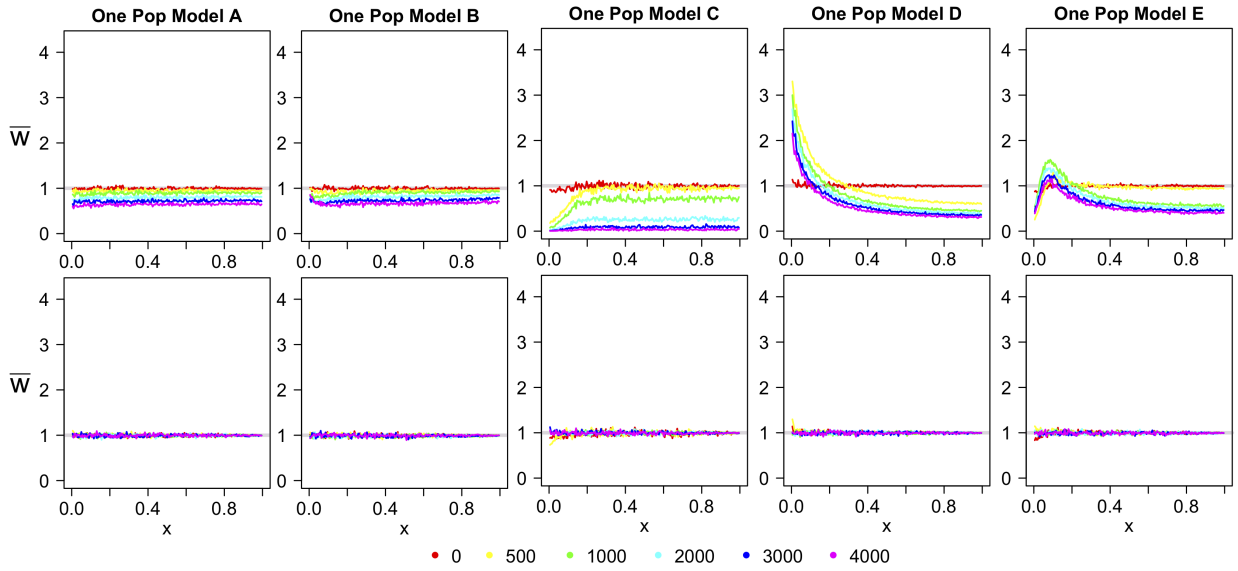


Figure 3.9: One Population Model simulated projections for the demographic models tested in Figure 3.8. The key indicates the time the ancient genomes were sampled, in generations. The top row gives the results for when the reference panel is sampled from the present and the test genome is sampled from the past. The bottom row gives the reverse. $\bar{w}$ is the value of the projection and x is the derived allele frequency in the reference population.

lower for ancient samples, until the time of sampling is younger than the time of divergence. When the time of sampling is younger than the time of divergence, the projection no longer changes as the sampling time changes—it looks the same as if the test genome was sampled from the present-day. Thus, if the time of sampling is known, the projection can determine whether an ancient sample is directly ancestral to a present-day population or a member of a related population that diverged before the time of sampling.

A pulse of admixture from the test population into the reference population shows an increase in rare alleles, but only if the test genome was sampled after the time of divergence (Figure 3.10, TPM B). Population size changes show the characteristic effects (decline in rare alleles for population expansion; increase in rare alleles for population decline; humped effect for population bottleneck; Figure 3.10, TPM C-E). Similar to the TPM A case, the projections for test genomes sampled more recently than the time of divergence look the same as for when the test genome was sampled in the present.

In the two-population model, when the reference panel consists of ancient samples and the test genome is sampled from the present-day, the projection is again different than the reverse (Figure 3.10, bottom row). As the reference panel is sampled closer to the time of divergence, the projection moves closer to the $\bar{w}(x) = 1$ line and away from the expected $\bar{w}(x) = e^{-\frac{t}{2N}}$ if the reference panel was sampled from the present. Once the reference panel is sampled from a time at least as old as the time of divergence, the projection acts similarly

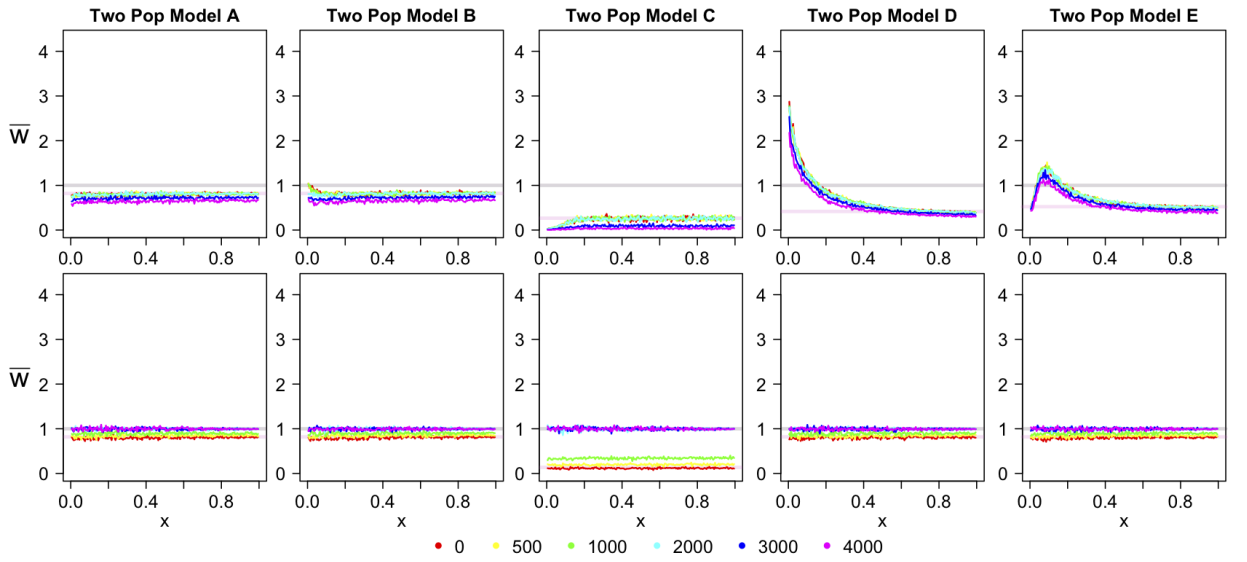


Figure 3.10: Two Population Model simulated projections for the demographic models tested in Figure 3.8. The key indicates the time the ancient genomes were sampled, in generations. The top row gives the results for when the reference panel is sampled from the present and the test genome is sampled from the past. The bottom row gives the reverse. $\bar{w}$ is the value of the projection and x is the derived allele frequency in the reference population.

as in OPM A; the test genome looks as if it was sampled from the reference population—that is, $\bar{w}(x) = 1$ for all x.

To conclude, the shape of the projection can be affected by the time of sampling. Particularly, the dynamics are notably different when the ancient samples are directly ancestral to the present-day samples and when they belong to a sister population that diverged from the present-day population. In the following chapter, we highlight when this distinction can be made with ancient hominin data.

3.3 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.

For scenarios involving ancient samples, projection analysis also provides information. Where the ancient population is directly ancestral to the modern population, if the test

genome is ancient and the reference panel is modern, the projection reflects the changes in the reference panel since the sampling time. However, when the test genome is modern and the reference panel is ancient, the projection of the test genome is on the line $\bar{w}(x) = 1$, despite the time that has passed since the ancient genomes were present.

In the alternate scenario where the ancient population is a member of a sister population, if the test genome is ancient and the reference panel is modern, the projection looks the same as when the test genome is sampled from the present. In the reverse situation when the test genome is modern and the reference panel is ancient, the projection of the test genome moves closer to the $\bar{w}(x) = 1$ as the reference panel is nearer to the time of divergence.

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. [24], Harris and Nielsen [27], Excoffier et al. [13], and Schiffels and Durbin [58]. 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.

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 $\bar{w}(x)$ 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 $\bar{w}(x)$ on small x. Relatively large samples from the reference population (approximately 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.

Projections provide a visually appealing method of comparing a single genome against a set of genomes belonging to a well-studied reference population. When genomes sampled are ancient, the projection can distinguish between several different demographic scenarios, providing further insight into potential demographic models to test using a more statistically rigorous analyses. Future studies of ancient genomes may find projections useful as a test for the ancestral relationship between the ancient sample and present-day populations. While not a method of demographic inference, the projections shape provides clues as to the direction of further model testing using formal demographic inference tools, such as *dadi* [24] or *fastsimcoal2* [13].

Chapter 4

Applications of Projection Analysis in Ancient and Modern Human Populations

4.1 Introduction

In the previous chapter, we illustrated using analytic theory and simulations that using projection analysis can inform on the demographic relationship between two populations. Here we apply this method in human populations, considering both present day and ancient genomic data.

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. We use projection analysis to test demographic inferences for European and Yoruba populations obtained from the four previous studies described above.

4.2 Application to Humans, the Altai Neanderthal and the Denisovan

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) [1]. For test genomes, we used the high-coverage Denisovan genome [43], the high-coverage Neanderthal genome [49], and some of the high-coverage present-day human genomes sequenced by Meyer et al. [43]. 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. [43]. These labels are provided in a note in Table 4.1. 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.

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_{ANC1}/N_{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_{ANC2}/N_{HAN}	0.1	0.01–1	Prüfer *et al.* (2014)	East Asian population expansion
	N_{ANC3}/N_{PAP}	0.1	0.01–1	Prüfer *et al.* (2014)	Papuan population expansion
	N_{ANC4}/N_{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_{ANC5}/N_{ANC1}	4	1.0–10	Gravel *et al.* (2011); Harris and Nielsen (2013); Prüfer *et al.* (2014)	Non-African population decline
	N_{ANC6}/N_{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 4.1: Description of parameters used in the simulation of the 10-population tree in Figure 4.1.

To show that projections give insight into human demographic history, we developed a

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 4.2: Description of parameters used in the simulation of the 10-population tree in Figure 4.1 (continued from Figure 4.1).

10-population demographic history with realistic parameters taken from the literature and adjusted using different curve-fitting techniques (Table 4.1 and Figure 4.1). The initial parameter ranges that we chose were informed by a variety of previous studies, as noted

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	$f_{MAN–YOR}$	0.1	0–0.15	Prüfer *et al.* (2014)	With the close relationship between these two populations, admixture was allowed.
	$f_{YOR–MAN}$	0.1	0–0.15		
	$f_{FRE–HAN}$	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.
	$f_{HAN–FRE}$	0.01	0–0.15		
	$f_{FRE–YOR}$	0.001	0-0.15		
	$f_{YOR–FRE}$	0.005	0–0.15		
	$f_{YOR–HAN}$	0.003	0–0.15		
	$f_{SAN–YOR}$	0.05	0–0.15		
	$f_{MBU–YOR}$	0.05	0–0.15		
	$f_{DIN–YOR}$	0.01	0–0.15		
	$f_{ANC1–YOR}$	0.01	0–0.15		
	$f_{ANC1–ANC4}$	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.
	$f_{ANC4–ANC1}$	0.2	0–0.5		
	$f_{NEA–DEN}$	0.01	0–0.05	Prüfer *et al.* (2014)	Low amounts of admixture from archaic hominins were added.
	$f_{DEN–PAP}$	0.03	0–0.05	Prüfer *et al.* (2014)	
	$f_{NEA–ANC1}$	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.

Table 4.3: Description of parameters used in the simulation of the 10-population tree in Figure 4.1 (continued from Figure 4.1 and Figure 4.2).

in Table 4.1. 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* [24] and the Broyden-Fletcher-Goldfarb-Shanno algorithm [44], we estimated several demographic parameters simultaneously that gave the best-fitting projection for the two populations. For more than two populations, we used *fastsimcoal2* [13] 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 4.1 is not meant to be optimal. Instead,

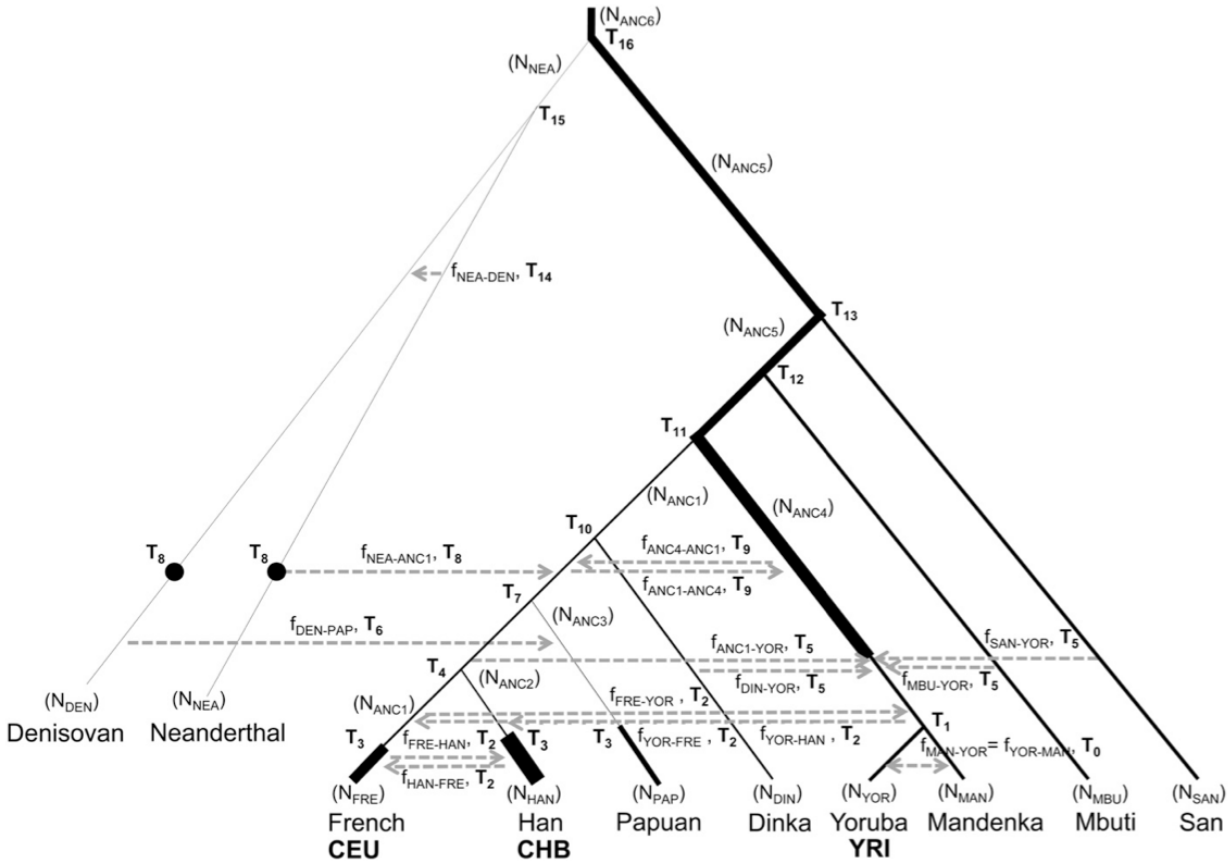


Figure 4.1: A model of human demographic history for 10 populations that, when simulated, gave projections similar to the observed projections. The populations in boldface type are the reference populations, and the row above them indicates the population origin of each test genome. The values used are found in Table 4.1-Table 4.3. 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.

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 4.2, 4.3, 4.4 and 4.5 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

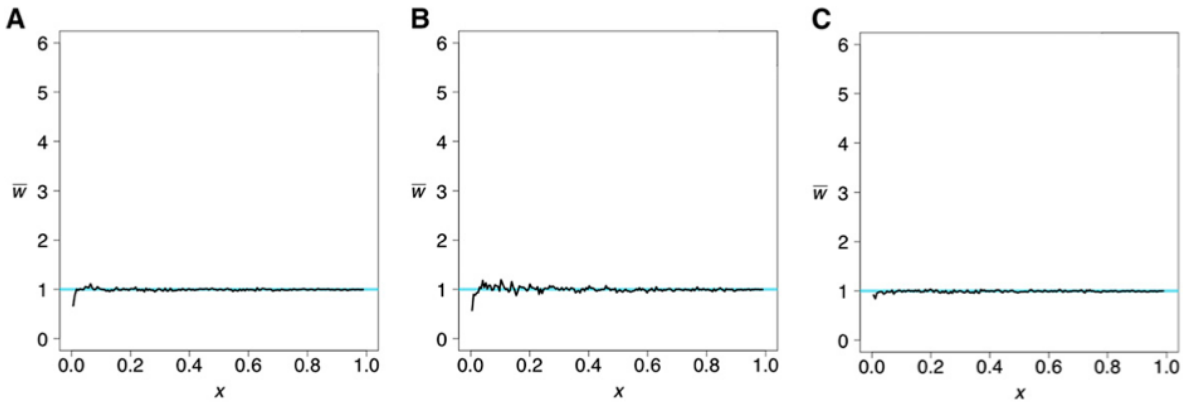


Figure 4.2: The projections of French onto CEU (A), Han onto CHB (B), and Yoruba onto YRI (C). The sum of LSS scores comparing the observed projection to the line are found in Table 4.5, Table 4.6, and Table 4.7.

	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

Table 4.4: LSS comparing the simulated projection from our model (Figure 4.1) to the observed projections (Figure 4.3, Figure 4.4 and Figure 4.5).

the 10-population demographic scenario described in Figure 4.1, which are represented by the red curves in Figure 4.3, Figure 4.4, and Figure 4.5. Table 4.5, Table 4.6 and Table 4.7 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.

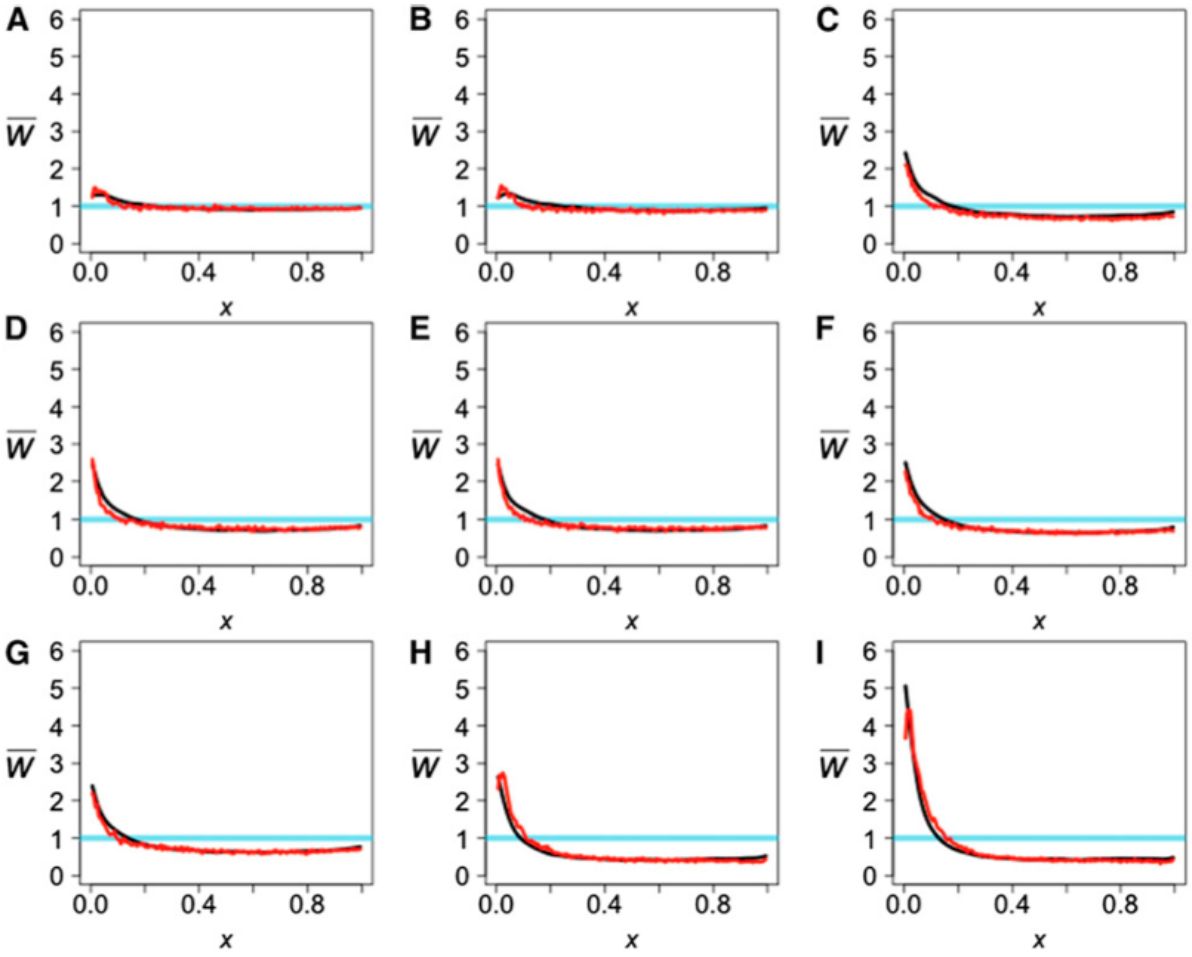


Figure 4.3: The observed projection (black line) and simulated projection from our model (red line) for the CEU reference population. The test genomes are Han (A), Papuan (B), Dinka (C), Yoruba (D), Mandenka (E), Mbuti (F), San (G), Denisovan (H), and Neanderthal (I). The LSS scores comparing the observed projections to each other and the expectation can be found in Table 4.5, and the LSS scores comparing the observed and simulated projections can be found in Table 4.4.

Comparison of a test genome with the same population

In Figure 4.2A, the projection of the French genome onto CEU fits the expectation except for small x. Similar deviations are seen in Figure 4.2B in the projection of the Han genome onto CHB and, to a lesser extent, in Figure 4.2C 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 in Yang et al. [73] for details.

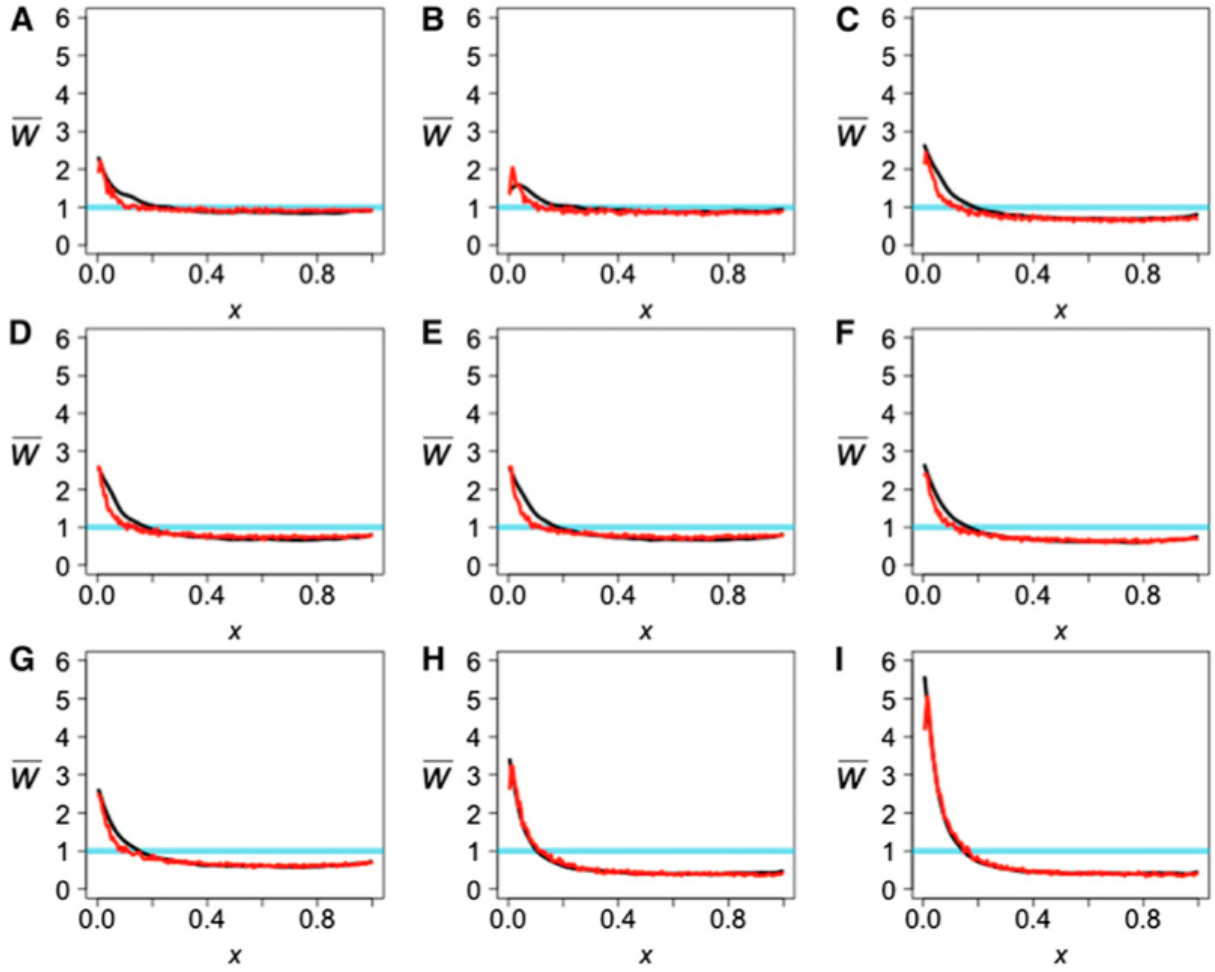


Figure 4.4: The observed projection (black line) and simulated projection from our model (red line) for the CHB reference population. The test genomes are French (A), Papuan (B), Dinka (C), Yoruba (D), Mandenka (E), Mbuti (F), San (G), Denisovan (H), and Neanderthal (I). The LSS scores comparing the observed projections to each other and the expectation can be found in Table 4.6, and the LSS scores comparing the observed and simulated projections can be found in Table 4.3.

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 3.3A, black curve). The projections of the Altai Neanderthal onto CEU and CHB show a large excess of rare alleles (Figure 4.3I and Figure 4.4I), which requires the combination of a bottleneck in the ancestors of non-Africans and admixture from Neanderthals into non-Africans after that bottleneck.

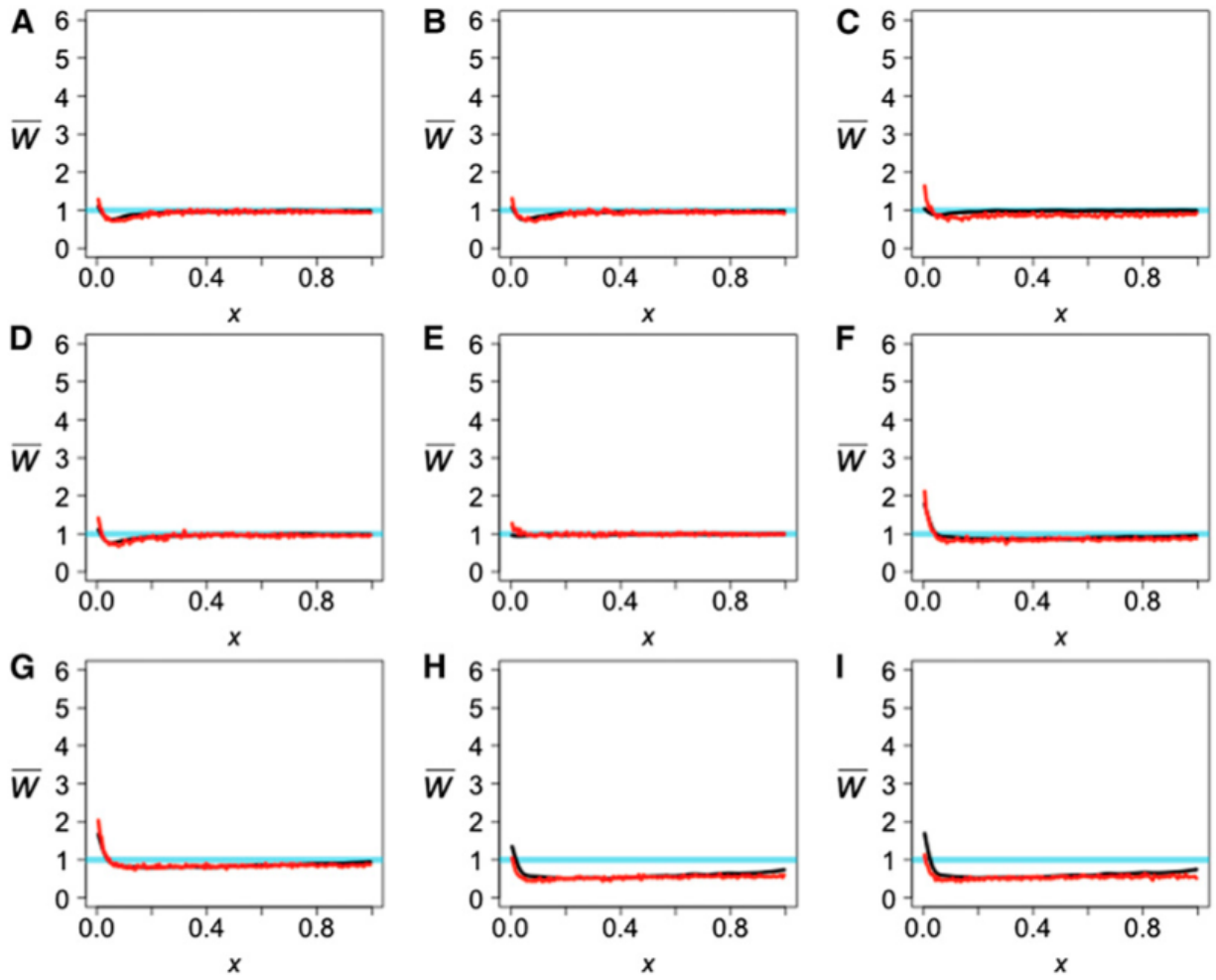


Figure 4.5: The observed projection (black line) and simulated projection from our model (red line) for the YRI reference population. The test genomes are Han (A), Papuan (B), Dinka (C), French (D), Mandenka (E), Mbuti (F), San (G), Denisovan (H), and Neanderthal (I). The LSS scores comparing the observed projections to each other and the expectation can be found in Table 4.7, and the LSS scores comparing the observed and simulated projections can be found in Table 4.3.

Including both processes in our model, we obtain good fits to the observed projections (Table 4.4, Figure 4.3I, and Figure 4.4I). When admixture is omitted, the result is a decrease in the excess of rare alleles and a worse fit (Table 4.9 and Figure 4.6).

Similarly, the projections of the Denisovan genome onto CEU and CHB (Figure 4.3H and Figure 4.4H) are consistent with the three-population analysis shown in Figure 3.4A and Figure 3.5D. 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

	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 4.5: 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	**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 4.6: 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	**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 4.7: LSS comparing the observed projections for reference YRI to each other. The diagonals compare that test genome to the $\bar{w}(x) = 1$ line.

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 4.4,

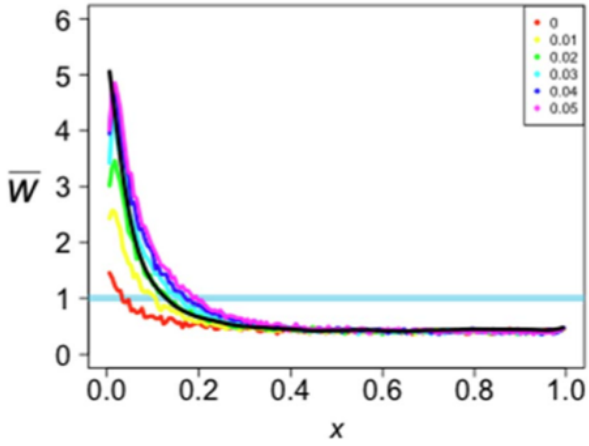


Figure 4.6: The simulated projections for reference CEU and test Neanderthal for our model when altering the amount of admixture ($f_{NEA-ANC1}$). The black line is the observed projection.

Figure 4.3H, and Figure 4.4H). 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 4.5I) and Denisovan (Figure 4.5H) 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 [49]. These projections are consistent with the signal of Neanderthal introgression being carried by recent admixture from the ancestors of Europeans and East Asians into the ancestors of the Yoruba population. In our model (Figure 4.1), 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 4.5, 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. [49] show that the inbreeding coefficient was $\frac{1}{8}$. 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 4.4A) differs from the projection of the Han genome onto CEU (Figure 4.3A). 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

	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

Table 4.8: 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 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 4.9: LSS (Least Sum of Squares) comparing our model to the observed projections, altering the amount of admixture from Neanderthals to non-Africans ($f_{NEA-ANC1}$).

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 4.4A). When these events are included in our model, the resulting projections are relatively close to the observed projections (Table 4.4).

The Papuan demographic history modeled here includes divergence from the ancestors of Europeans and East Asians and a bottleneck and population expansion (Figure 4.1). 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. [72], but not by others [43, 49]. We made this assumption because we followed Gravel et al. [19] in assuming that Europeans and East Asians diverged relatively recently. With admixture from Denisovans 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 4.4, Figure 4.3B, and Figure 4.4B).

Relationship between non-Africans and YRI

The projections of the Papuan, French, and Han genomes onto YRI (Figure 4.5A, 4.5B, and 4.5D) are similar despite the difference between the Han and Papuan projections onto CEU (Figure 4.3A and 4.3B). 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 $\bar{w}(x) = 1$ for larger x, and they lead to a good fit to the observed projections (Table 4.4 and Figure 4.5A, 4.5B, and 4.5D). 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 4.10 and Figure 4.7).

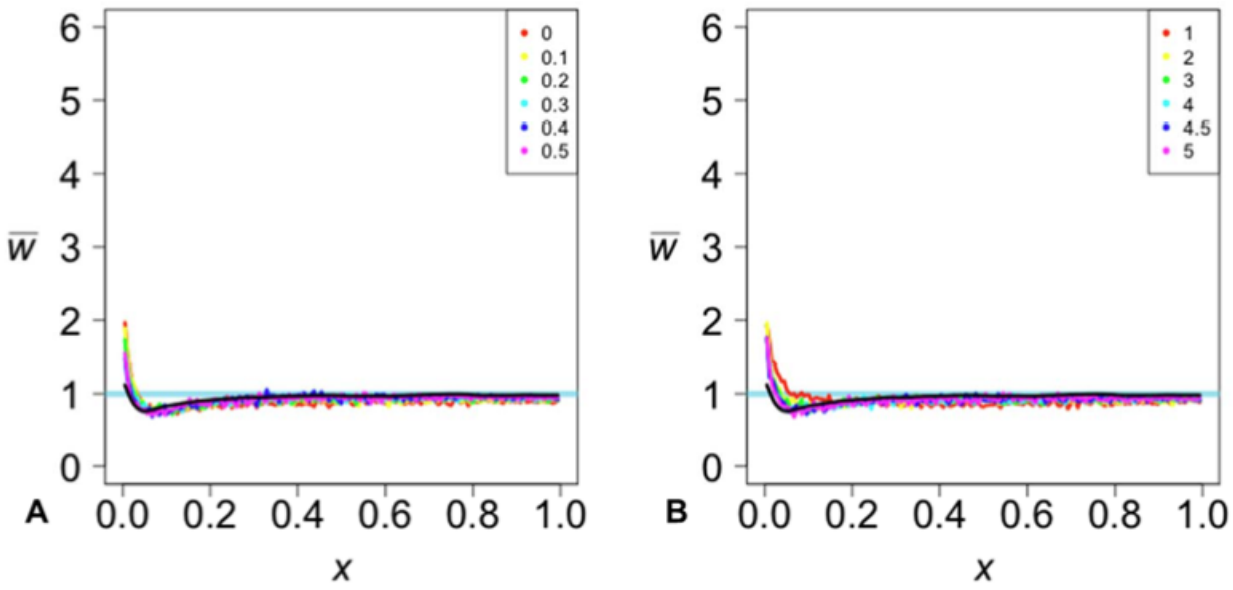


Figure 4.7: The simulated projections for reference YRI and test French for our model when altering (A) the population increase in the Yoruba population backwards in time ($\frac{N_{ANC4}}{N_{YOR}}$) and (B) the amount of admixture from Europeans to Yorubans ($f_{ANC1ANC4}$). The black line is the observed projection.

African projections onto CEU and YRI

The projections of all five African genomes—San, Yoruba, Mandenka, Dinka, and Mbuti—onto CEU (Figure 4.3C-4.3G) are similar to one another and similar to their projections onto CHB (Figure 4.4C-4.4G). All these projections are consistent with low levels of admixture from the African populations into the ancestors of Europeans and East Asians. Previous

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

Table 4.10: LSS comparing our model to the observed projections, altering the population increase in the Yoruba population backwards in time ($\frac{N_{ANC4}}{N_{YOR}}$) and the amount of admixture ($f_{ANC1-ANC4}$) from Europeans to Yorubans.

analyses [38, 43, 47, 49] 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 results similar to the observed projections (Table 4.4 and Figure 4.3C to Figure 4.3G). 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 4.3D and 4.3G), 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 4.5C, Figure 4.5E to Figure 4.5G) illuminate the relationship between these four African populations and the Yoruba. Other studies [67, 43] 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 that it is less diverged from YRI than is the San, agreeing with the model proposed in other studies [67, 43, 49]. The Mandenka projection falls nearly on the $\bar{w}(x) = 1$ line, 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 4.1) gives a close fit to the observed projections (Table 4.4 and Figure 4.5C, Figure 4.5E to Figure 4.5G).

Test of Published Models

We used observed projections to test for consistency with inferred demographic parameters from four studies [19, 13, 27, 58] for European and Yoruba populations. All four studies applied their methods to these two populations.

We obtained projections by using *fastsimcoal2* [13] to simulate 1 million SNPs with the estimated demographic parameters from each of these four models. The demographic parameters used are shown in Figure 4.8. 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 AC, the time of divergence of Europeans and Yoruba is approximately 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 [58] say that such migration probably occurred.

The simulated projections show that model A gives the best fit to the observed projections (Table 4.8 and Figure 4.9). For model A, increasing the rate of recent migration from Yoruba to Europeans from 0.000025 migrants/generation to 0.00005 migrants/generation led to a slightly better fit (Table 4.8 and Figure 4.10). 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

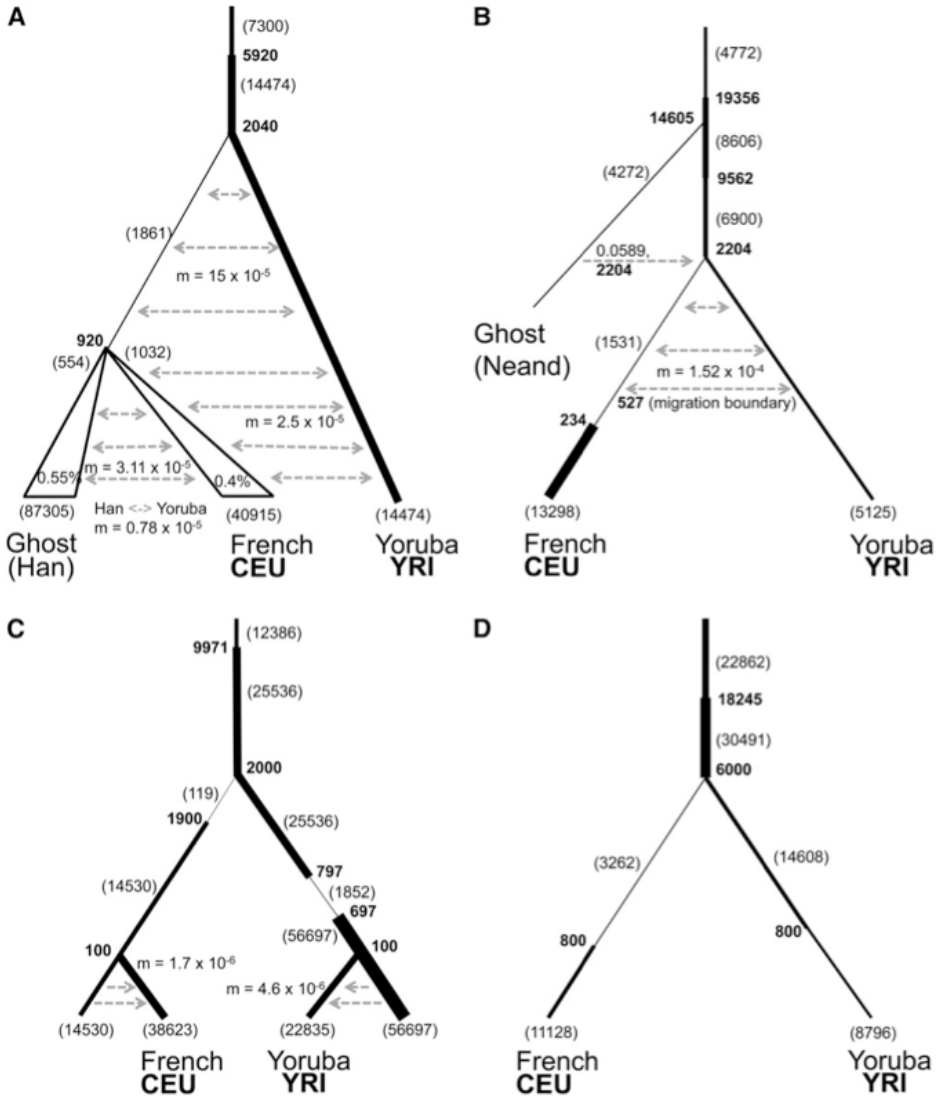


Figure 4.8: The demographic models from each of the four previous studies: Gravel et al. [19] (model A), Harris and Nielsen [27] (model B), Excoffier et al. [13] (model C), and Schiffels and Durbin [58] (model D). Shading and symbols have the same meaning as in Figure 4.1, and the triangle indicates growth at the given percentage.

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

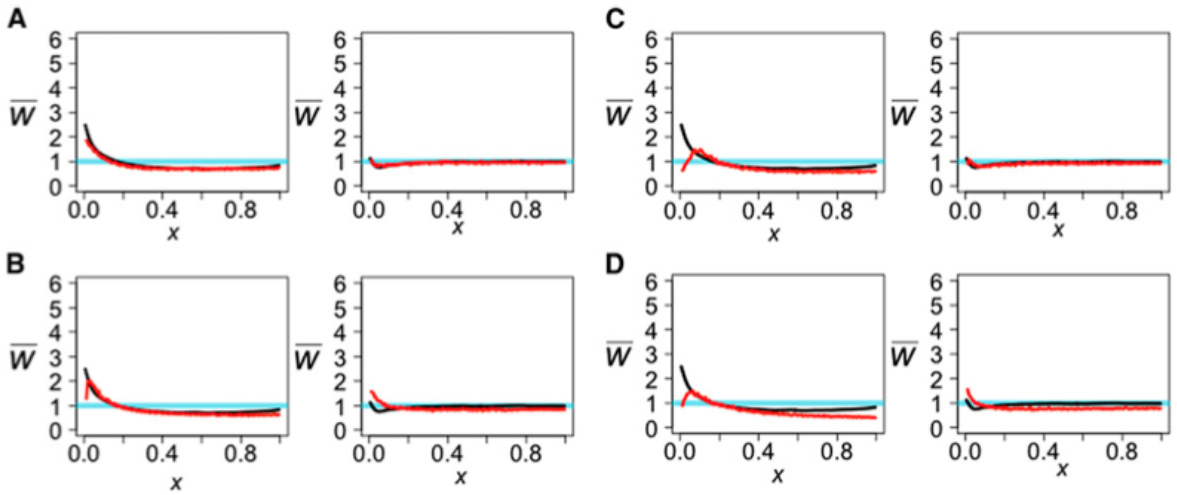


Figure 4.9: The observed projections (black line) and simulated projections from demographic models inferred from other studies (red line). For each model A-D in Figure 4.6, the left projection is the Yoruba genome projected onto CEU and the right projection is the French genome projected onto YRI. LSS scores are in Table 4.8.

to Yoruba and 0.0003 migrants/generation from Yoruba to Europeans results in a better fit (models A*-D*, Table 4.8; Figure 4.10).

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.

Sample	Date Used[a]	Covg[b]	Reference	MinCov[c]	MaxCov[c]
Vindija	40000	1.3	Green et al. 2010	1	4
Mezmaiskaya	65000	0.5	Prüfer et al. 2014	1	3
Ust-Ishim	45000	42	Fu et al. 2014	21	64
Loschbour	6000	22	Lazaridis et al. 2014	5	25
Stuttgart	5000	19	Lazaridis et al. 2014	7	34

[a] in thousands of years, roughly taken from the date ranges found in the reference.

[b] the average coverage given in the reference.

[c] the 2.5% and 97.5% interval cutoffs for the coverage that were used in the analysis.

Table 4.11: Data used for each ancient genome

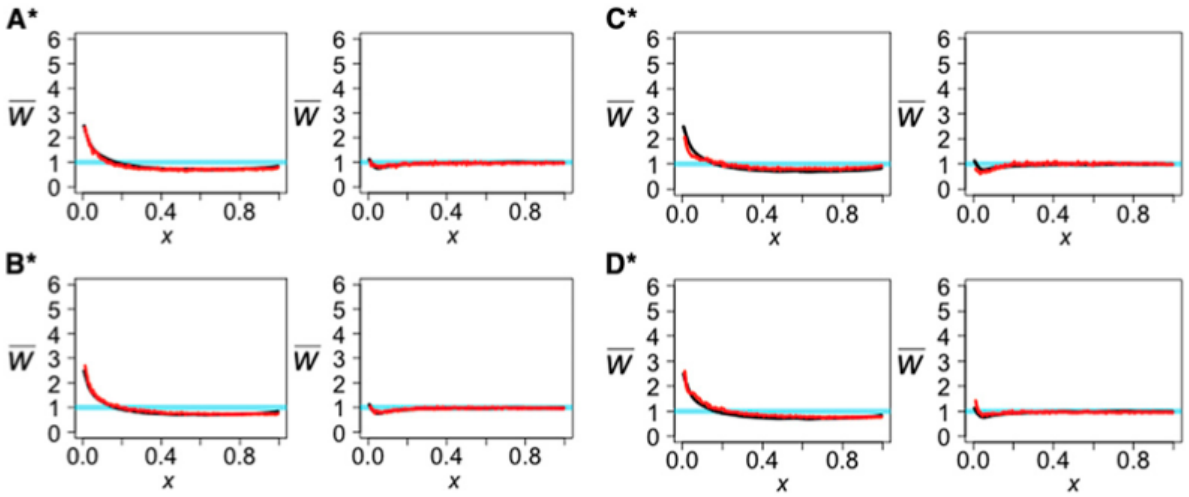


Figure 4.10: Projections for previous studies (models A-D) where the parameters for migration or admixture between Europeans and Yorubans have been added or modified for a better fit. For each model A*-D*, the left projection is the Yoruba genome projected onto CEU and the right projection is the French genome projected onto YRI. LSS scores are in Table 4.8.

4.3 Additional Application to Neanderthals and Ancient Humans

In addition to the Altai Neanderthal and the Denisovan, we further explored what we could learn about demographic relationships when using ancient samples. Five ancient genomes were compared to present-day human populations using projection analysis. Of the five, two are Neanderthal and three are ancient modern humans. Table 4.11 indicates the sampling time, as indicated by the study in which the genome was sequenced. The Vindija Neanderthal was the original Neanderthal genome sequenced [21], and the Mezmaiskaya Neanderthal was sequenced by Prüfer et al. [49]. The three ancient modern humans used in this study are the Ust-Ishim [15], the Loschbour and the Stuttgart genomes [39]. The Ust-Ishim individual died 45 kya, and is equally distant from all present-day non-Africans, with some greater admixture into present-day East Asians [15]. The Loschbour and Stuttgart genomes date to around 7-8 kya, in Central Europe. The Loschbour individual was found in a hunter-gather site, while the Stuttgart individual was associated with the Linearbandkeramik farming culture. Both of these genomes are of West Eurasian ancestry and are members of different populations that contributed to present-day European populations [39].

We project these five genomes onto three reference panels representing Europeans (CEU), Han Chinese (CHB) and the Yoruba (YRI) populations. To calculate the projection, we modified the analysis from that found in Yang et al. [73] to use reads instead of genotypes

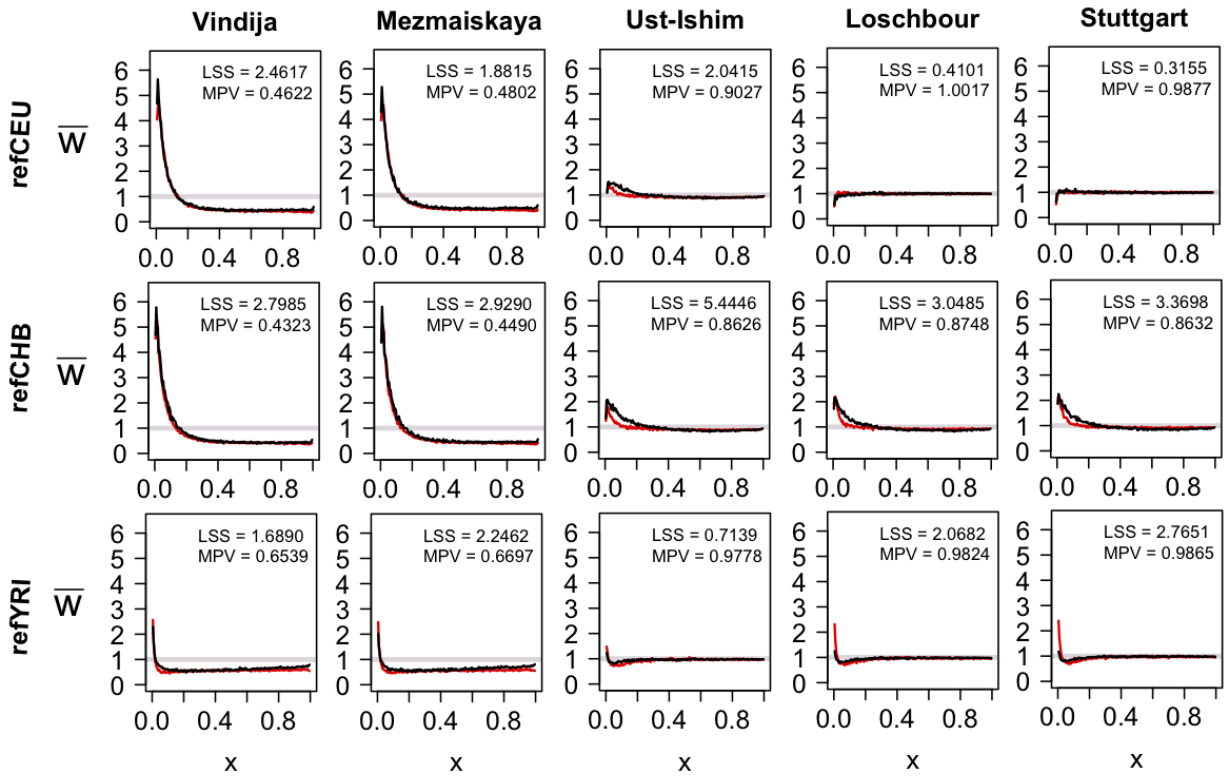


Figure 4.11: Projection of ancient hominin genomes onto the European reference panel (refCEU, top row), the Han Chinese reference panel (refCHB, middle row) and the Yoruba reference panel (refYRI, bottom row), with the observed projection in black and the simulated projection in red. The sum of least squares (LSS) score gives the fit between the observed and simulated projections. The mean projection value (MPV) is the mean for $x > 0.5$.

called from the reads, in order to more accurately assess low coverage samples. We used the CEU, CHB and YRI panels from Phase 3 of the 1000 Genomes Project as the reference panels ([2]). We considered only biallelic sites where the mutation was a transversion. We filtered out any sites where the mapping quality was less than 30, and for each ancient genome we filtered for sites where the coverage was within the 2.5% to 97.5% interval of the coverage distribution unique to each sample (Table 4.11, minCov and maxCov). The derived allele frequency of the reference panel was determined by using the genotypes assessed in the Phase 3 panels and the ancestral allele called in the Phase 3 1000 Genomes data set. For each site, the test genome was called derived or ancestral by choosing randomly from the set of reads for that site. The projection was calculated across all autosomal sites that were not filtered out by the above criteria. A minimum projection value (MPV) was calculated using the average projection for $x > 0.5$. The projections within each panel were compared

to each other and to the line using the sum of least squares (LSS) score (Table 4.12).

refCEU	*Vindija*	*Mezmaiskaya*	*Ust-Ishim*	*Loschbour*	*Stuttgart*
Vindija	**145.14**	0.61	106.95	156.16	146.29
Mezmaiskaya		**130.95**	94.62	141.32	131.80
Ust-Ishim			**4.74**	6.86	4.30
Loschbour				**0.47**	0.55
Stuttgart					**0.30**
refCHB	*Vindija*	*Mezmaiskaya*	*Ust-Ishim*	*Loschbour*	*Stuttgart*
Vindija	**170.61**	0.39	96.64	94.02	92.28
Mezmaiskaya		**160.12**	88.77	86.33	84.65
Ust-Ishim			**14.99**	0.69	0.74
Loschbour				**13.61**	0.35
Stuttgart					**14.52**
refYRI	*Vindija*	*Mezmaiskaya*	*Ust-Ishim*	*Loschbour*	*Stuttgart*
Vindija	**33.09**	0.23	26.88	27.09	27.84
Mezmaiskaya		**30.46**	24.49	24.66	25.34
Ust-Ishim			**0.75**	0.10	0.12
Loschbour				**0.84**	0.09
Stuttgart					**0.84**

Bolded values indicate the score relative to the line $\overline{w(x)} = 1$, while regular values indicate the score relative to each other, within the reference panel.

Table 4.12: Sum of least square scores when comparing projections of test genomes within a reference panel

In the projections, there are several notable characteristics (Table 4.12; Figure 4.11, black curve). First, with respect to the reference panel refCEU, the projections for the ancient samples can be divided into three main groups. The Neanderthals have the lowest projections, with MPV values of 0.4622 and 0.4802 (Figure 4.11, top row). Both Neanderthals show a substantial increase in rare alleles and have very similar projections (LSS=0.61, Table 4.12). The Ust-Ishim shows the next lowest MPV of 0.9027 (Figure 4.11, top row) with minor deviations from a horizontal line likely indicative of population size changes in the refCEU population. The Loschbour and Stuttgart genomes lie almost on the line (Table 4.12; LSS = 0.47 and 0.30, Figure 4.11, top row), with a slight decrease for small x.

For the refCHB reference panel (Table 4.12; Figure 4.11, middle row), the projections for the Neanderthal genomes are nearly identical to that observed for the refCEU panel (MPV values of 0.4323 and 0.4490, Figure 4.11, middle row). The Ust-Ishim, Loschbour and Stuttgart projections all indicate they are not members of the Han Chinese population (MPV = 0.8626, 0.8748 and 0.8632, Figure 4.11 middle row). LSS values for each projection are all very high, ranging from 13.61 to 170.61, further supporting that none of these ancient genomes are directly ancestral to the Han Chinese (Table 4.12). Finally, for the refYRI

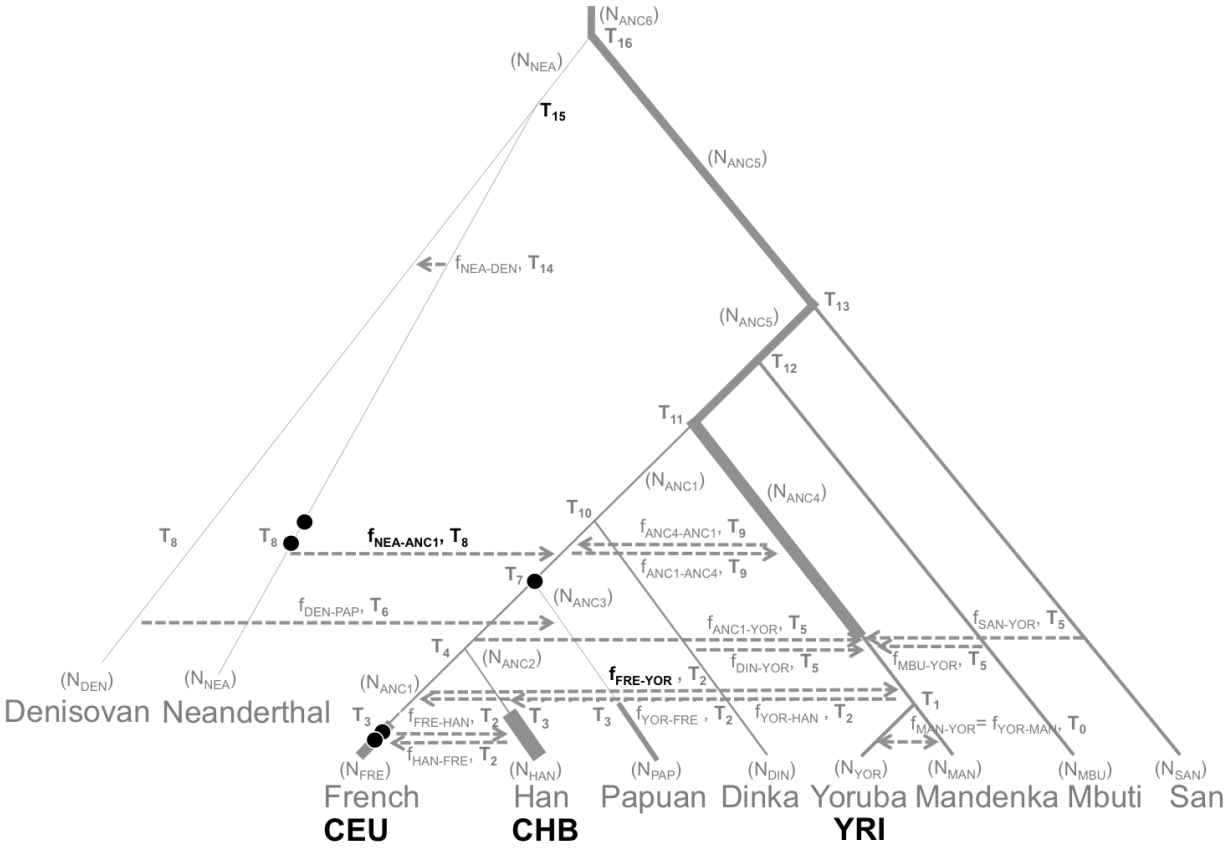


Figure 4.12: The placement of the five ancient genomes (black circles) in the demographic model described in Figure 4.1shaded gray, Table 4.1 to 4.3 contains parameter values). The time of sampling for these five genomes are included in Table 4.11. Bolded parameters are those that were modified to improve the fit onto the projections (values in Table 4.13).

panel, the projections are unusual (Figure 4.11, bottom row), but similar to that observed by Yang et al. [73]. The Neanderthals have a higher projection onto the refYRI panel (MPV values of 0.6539 and 0.6697, Figure 4.11, bottom row) than the non-Africans. The higher MPVs are probably because the Yoruba did not undergo the same bottleneck detected in non-Africans. For non-Africans, the projection increases for large x, which was shown in simulations of Yang et al. [73] to be due to high levels of ancient admixture between the ancestral Yoruba and non-African populations, as well as a population decline in the Yoruba population. This results in a closer fit to the line and lower LSS scores (Table 4.12), despite the fact that these genomes are not ancestral to the present-day Yoruba population. The shape of these projections is very similar to those for present-day non-Africans relative to the refYRI panel [73].

Comparing the projections to a simulated demography

To gain greater perspective on how the projections of these ancient genomes relate to human demographic history, we compared the ancient genomes to simulated projections taken from a demographic model. We used the demographic model that best fit the set of projections for modern humans published in Yang et al. [73], which included eight populations of European, African, East Asian and Papuan origin, and the Altai Neanderthal and Denisovan. For each ancient genome, we simulated the same demographic model, adding a single simulated sample retrieved at the time indicated in Table 4.11, where one generation is assumed to be 25 years. The Neanderthals were placed on the Neanderthal lineage, the Ust-Ishim genome shared a common ancestor with Europeans and East Asians, and the Loschbour and Stuttgart genomes were placed on the European lineage (Figure 4.12), in accordance with the conclusions of their respective studies [21, 49, 16, 39].

Using *fastsimcoal2* (ver 2.1, [13]) and Brents algorithm, the time (T_8) and amount ($f_{NEA-ANC1}$) of Neanderthal admixture, the time of Neanderthal divergence ($T_1$5) and the recent admixture from Europeans to Yoruba ($f_{FRE-YOR}$) were allowed to vary to improve the fit of the projections (Figure 4.12, bolded). The LSS was calculated when each simulated and real projection was compared (Figure 4.11, LSS score in top right corner). Using a time of Neanderthal divergence of 610,175 yrs, with admixture into non-Africans 38,950 yrs ago of 0.018, and recent admixture 7,500 years ago from Europeans to Africans of 0.02 (Table 4.13), the simulated and observed projections exhibited low LSS scores (Figure 4.11).

Parameter	Value Used
T_8	38950
T_{16}	610175
$f_{NEA\text{-}ANC1}$	0.018
$f_{FRE\text{-}YOR}$	0.02

Table 4.13: Parameter values used in simulated demographic model that differs from Table 4.1 through 4.3.

4.4 Discussion and Conclusions

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. [24] and Gravel et al. [19] 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 [27] 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. [13] 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 [58] 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 projections of the Neanderthals all show a very similar projection to each other with respect to each reference panel, despite the differences in sampling time. They also look very similar to the Altai Neanderthal and Denisovan projections analyzed in Yang et al. [73]. Therefore, these genomes belong to a sister group and the reconstructed demographic history that recovers the observed projections also places them all in a sister group. These results concur with the conclusions of previous studies [55, 43, 49]. The increase in rare alleles for their projections onto the refYRI panel was recovered by including some recent admixture from Europeans to the Yoruba population. Another scenario that was not illustrated here is direct admixture from Neanderthals or a sister group to Neanderthals directly into the ancestral Yoruba population. This is unlikely, as recent studies have proposed recent admixture from non-African to African populations [70, 69], and another [18] has shown that there is European gene flow back into many regions of Africa. While we simulated direct admixture from the CEU population to the Yoruba, the admixture may have come from a population distinct from the ones to which the Loschbour and Stuttgart genomes belong. Accounting for this may improve the fit of the Loschbour and Stuttgart projections onto the refYRI panel.

The Ust-Ishim genome is different from both the European and East Asian panels, showing it is likely not a member of either population, but it behaves similarly to other non-Africans with respect to the Yoruba panel. When a simulated ancient sample was placed directly ancestral to Europeans and East Asians 45 kya, the simulated projection was very similar to the observed projection, illustrating that the shape in the projection can largely be

attributed to the population size changes in Europeans and East Asians after the Ust-Ishim was sampled.

The Loschbour and Stuttgart genomes sit on the $\bar{w}(x) = 1$ line when projected onto the refCEU panel, but not when projected onto the refCHB or refYRI panel. The projections show that the Loschbour and Stuttgart could be considered the same population as present-day Europeans. Lazaridis et al. [39] showed that both of these genomes are members of different ancestral source populations for present-day Europeans. Though Europeans are composed of several different source populations, the projections analyzed only shows that these two genomes are ancestral to Europeans, but it does not specify whether there are other ancestral populations also.

We studied the projections of several ancient hominin genomes. Neanderthals were not directly ancestral to modern humans. The Ust-Ishim projection looks ancestral to both Europeans and East Asians, and the Loschbour and Stuttgart projections suggest that they are ancestral to Europeans, but not to East Asians or the Yoruba.

Chapter 5

Applications of Projection Analysis in Ancient and Modern Horse Populations

5.1 Introduction

The projection analysis described in Chapter 3 was also applied to ancient and modern genomic data from the *Equus* genus. Here, we describe the demographic relationship between domesticated horse breeds, the Przewalski's horse, Yakut horses and four ancient horse genomes.

5.2 Demographic history of the Przewalski's horse

We used the genome projection method, where test genomes are projected onto reference populations as described in Chapter 3 to explore models of demographic history for the domesticated and Przewalski's horses.

Test horse genomes and reference panels

We used genomes obtained from 15 present-day Przewalski's horses and 28 domesticated horses representing a range of domesticated horse breeds and types (Table 5.1). Horse genomes were used either as test genomes, or grouped into reference panels: refPRZ for the 14 non-hybrid Przewalski's horse genomes, refFM for the Franches-Montagnes horse genomes, and refDOM for the full range of domesticated horses, including Franches-Montagnes horses (Table 5.1).

Horse ID	Horse type	Reference	Reference panel(s)
SB159	Przewalski	This study	refPRZ
SB274	Przewalski	This study	refPRZ
SB281	Przewalski	This study	refPRZ
SB285	Przewalski	This study	refPRZ
SB293	Przewalski	This study	refPRZ
SB339	Przewalski	This study	refPRZ
SB524	Przewalski	This study	refPRZ
SB528	Przewalski	This study	refPRZ
SB533	Przewalski	This study	refPRZ
SB615	Przewalski	This study	refPRZ
SB966	Przewalski	This study	refPRZ
Prz_D2630	Przewalski	Do et al., 2014	refPRZ
Prz_D2631	Przewalski	Do et al., 2014	refPRZ
Prz_Przewalski	Przewalski	Orlando et al., 2013	refPRZ
Arabian	Arabian	Orlando et al., 2013	refDOM
Fjord	Norwegian Fjord	Orlando et al., 2013	refDOM
Icelandic	Icelandic	Orlando et al., 2013	refDOM
KB7754	Mongolian	This study	refDOM
Mng_D2628	Mongolian	This study	refDOM
Mng_D2929	Mongolian	Do et al., 2014	refDOM
Mor_EMS595	Morgan	This study	refDOM
Qrt_A5964	Quarter	This study	refDOM
Qrt_A5659	Quarter	This study	refDOM
Qrt_A1543	Quarter	This study	refDOM
Qrt_A2085	Quarter	This study	refDOM
Std_M977	Standardbred	This study	refDOM
Std_M5256	Standardbred	This study	refDOM
Std_M1009	Standardbred	This study	refDOM
Std_Standardbred	Standardbred	Orlando et al., 2013	refDOM
Thoroughbred	Thoroughbred	Wade et al., 2009	refDOM
Mon_FM1798	Franches-Montagnes	This study	refDOM, refFM
Mon_FM1932	Franches-Montagnes	This study	refDOM, refFM
Mon_FM1948	Franches-Montagnes	This study	refDOM, refFM
Mon_FM2218	Franches-Montagnes	This study	refDOM, refFM
Mon_FM1190	Franches-Montagnes	This study	refDOM, refFM
Mon_FM1041	Franches-Montagnes	This study	refDOM, refFM
Mon_FM1951	Franches-Montagnes	This study	refDOM, refFM
Mon_FM0467	Franches-Montagnes	This study	refDOM, refFM
Mon_FM_431	Franches-Montagnes	This study	refDOM, refFM
Mon_FM1785	Franches-Montagnes	This study	refDOM, refFM
Mon_FM1030	Franches-Montagnes	This study	refDOM, refFM
Mon_FM0450	Franches-Montagnes	This study	refDOM, refFM

Table 5.1: Przewalski's and domesticated horse genomes used as test genomes or in reference panels.

Projections of domesticated and Przewalski's horse test genomes to the reference panels

To determine the projections for each test genome and reference panel, we used only biallelic sites, with ancestry determined by comparison with an outgroup genome, i.e., that of the *Equus asinus somaliensis* individual Shakira [32]. Using a panel of outgroup species to determine ancestry did not change the projection results (data not shown).

The projection, $\bar{w}(x)$, as described in Chapter 3, explores how different a test genome is from a set of genomes from a reference population (reference panel), where x is the derived allele frequency in the reference panel. For sites belonging to each derived allele frequency category of the reference panel, the expected number of sites with a derived allele in a genome from the reference population can be calculated. The observed number of test genome sites with a derived allele is compared to the expected number to determine the projection. In the projection, a value greater than one means that it is more likely than expected that a derived allele will be observed in the test genome for that frequency category, and a value less than one means it is less likely than expected that a derived allele will be observed in the test genome for that frequency category. A member of the reference population would give a projection of $\bar{w}(x) = 1$ for every frequency category. Depending on the demographic relationship between the test and reference populations, the projection may deviate from $\bar{w}(x) = 1$ in a variety of ways. To calculate the projection, we retained sites that were variable in at least one of the genomes sampled (Table 5.1). The projection was determined for every individual against one of three reference panels, refPRZ, refFM and refDOM, for sites that are polymorphic in the reference panel.

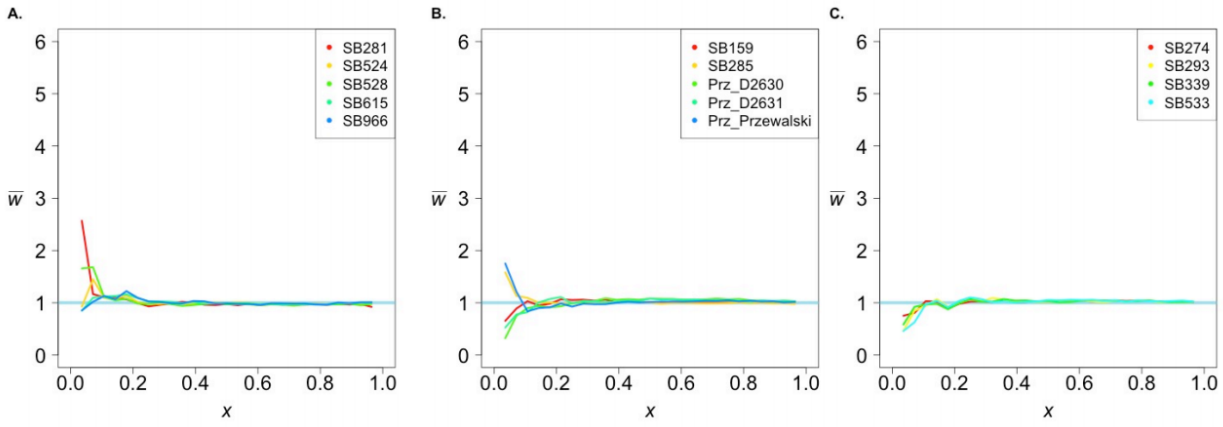


Figure 5.1: Projection of each Przewalski's horse genome onto the Przewalski's horse refPRZ reference panel. Projections are separated into genomes with high signals of admixture with domesticated horses (A), moderate signals of admixture (B), or little to no signal of admixture (C) as determined on the basis of ABBA-BABA admixture tests.

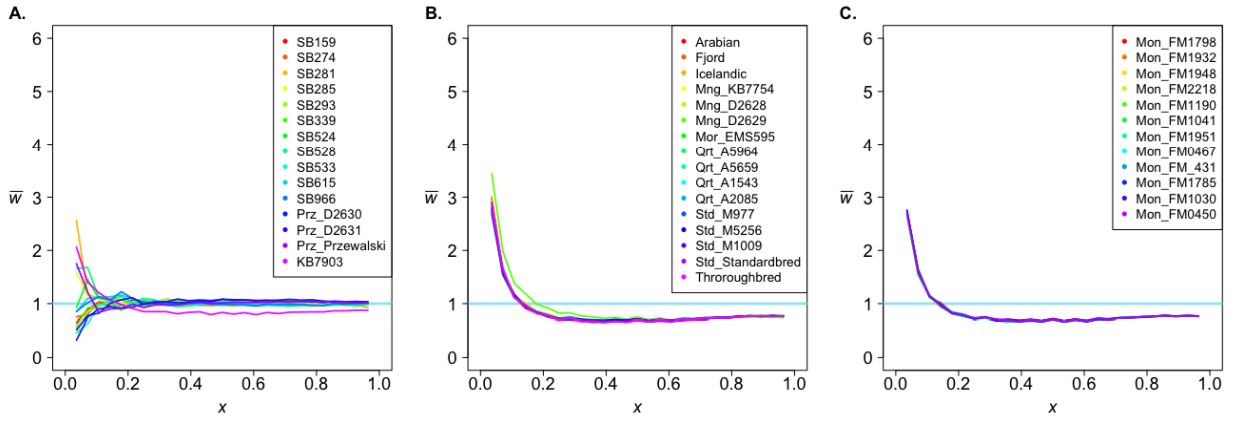


Figure 5.2: Projections of all genomes onto the Przewalski's horse refPRZ reference panel. A. Przewalski's horse test genomes. B. Domesticated horse test genomes. C. Franches-Montagnes test genomes.

We first compared each of the Przewalski's horse genomes to the refPRZ reference panel. The only projection that did not lie on the $\bar{w}(x) = 1$ line was, as expected, for the hybrid KB7903 (Figures 5.1 and 5.2). We found that generally, for intermediate and common alleles, the projection lied on the $\bar{w}(x) = 1$ line, confirming that the test genomes are members of the reference population. However, there was much variation for rare alleles. A dip at low frequencies is unsurprising, as a finite reference population size may lead to inflated counts of low frequency alleles (see the appendix in [73]). An increase at low frequencies was found in seven of the fourteen non-hybrid Przewalski's horse genomes. The excess of rare alleles could be due to recent admixture from domesticated horses into these particular individuals. When these results were compared to the admixture results using the D or the f_3 statistics (S4.3 in [56]), it could be seen that all seven genomes with an increase in the projection for low frequency alleles corresponded to a higher signal of admixture with all or some of the domesticated panel (Figure 5.1A and 5.1B, Figure 5.2A).

Specifically, for five Przewalski's horse genomes that exhibit significant amounts of admixture, projections clearly showed an increase above $\bar{w}(x) = 1$, i.e., SB281, SB524, SB528, SB615 and SB966 (Figure 5.1A). Five other Przewalski's horse genomes showed very slight amounts of admixture and varied in their projection, i.e., SB159, SB285, Prz_D2630, Prz_D2631 and Prz_Przewalski (Figure 5.1B), while the four remaining Przewalski's horses showed non-significant levels of admixture according to the D and f_3 statistics and we observed projections below $\bar{w}(x) = 1$, i.e., SB274, SB293, SB339 and SB533 (Figure 5.1C).

The refDOM panel, which includes more individuals, allows a greater resolution of the projection for less common alleles. However, refDOM is composed of genomes from several different horse breeds/types, which does not satisfy the assumption that the reference population is a randomly mating population. Indeed, the test projections for the non-Franches-Montagnes domesticated horses compared to refDOM showed high variability, and

many projections deviated from the line $\bar{w}(x) = 1$ (Figure 5.3B). The projections of the Franches-Montagnes horse genomes fell on the $\bar{w}(x) = 1$ line, but there was an important decrease in rare allele frequencies (Figure 5.3C), suggesting that the reference panel likely has high population structure. The projections of the Franches-Montagnes genomes onto refFM (Figure 5.4C) fell on the line $\bar{w}(x) = 1$ (Figure 5.4C), with a much smaller decrease for rare alleles, while the same was not observed for other domesticated horses (Figure 5.4B). To avoid the effects of population structure within the reference panel, we focused on the projections for refFM, rather than refDOM, in further analyses.

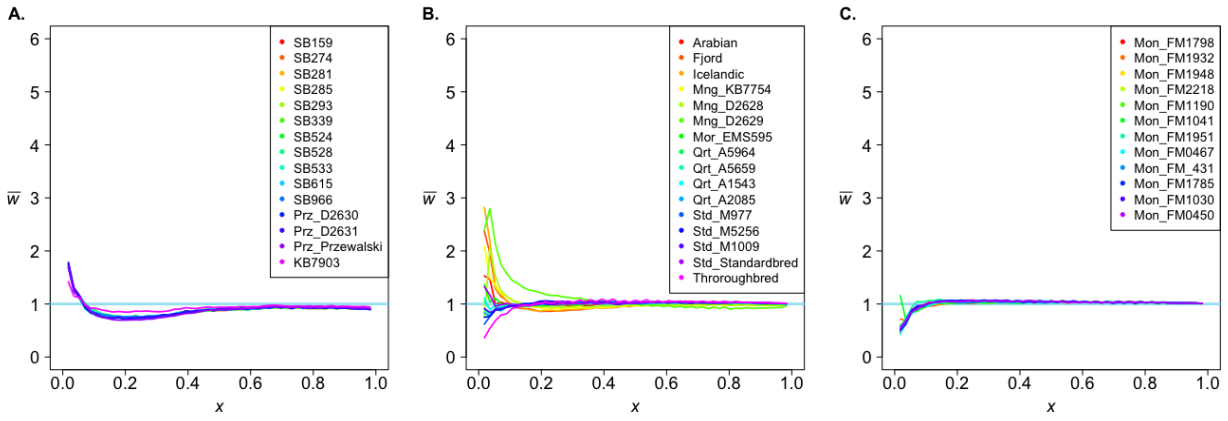


Figure 5.3: Projections of all genomes onto the domesticated horse refDOM reference panel. A. Przewalski's horse test genomes. B. Domesticated horse test genomes. C. Franches-Montagnes test genomes.

When comparing all the non-Franches-Montagnes domesticated horses to refFM, the projection looked very similar, for all horses but one Mongolian horse genome (Mng_D2629, Figure 5.4B). Relative to the other domesticated horse genomes, this genome showed an increase in less common alleles in the projection. The other difference observable was a slightly lower minimum projection for the Norwegian Fjord and Icelandic horse genomes (mean minimum projection value, MMPV= 0.7526), relative to the other non-Franches-Montagnes domesticated horses (MMPV= 0.8145). This is in keeping with the known phylogenetic tree for domesticated horses, where these two breeds are the most diverged from the other horse breeds [42]. Aside from Mng_D2629, the Mongolian horses have the next lowest projection values (MMPV= 0.7875).

When the Przewalski's horse genomes are compared to refFM there is little variation in the projection for different genomes (similarly for refDOM, Figure 5.3A), but for one horse, the hybrid KB7903 (Figure 5.4A). The hybrid has a projection that is closer to the $\bar{w}(x) = 1$ line (minimum projection value= 0.7676), indicating that it is more similar to the domesticated breeds than the other Przewalski's horses. The MMPV for the Przewalski's horse genomes (0.6907, sd= 0.0243) is far lower than that observed for horses in the panel of

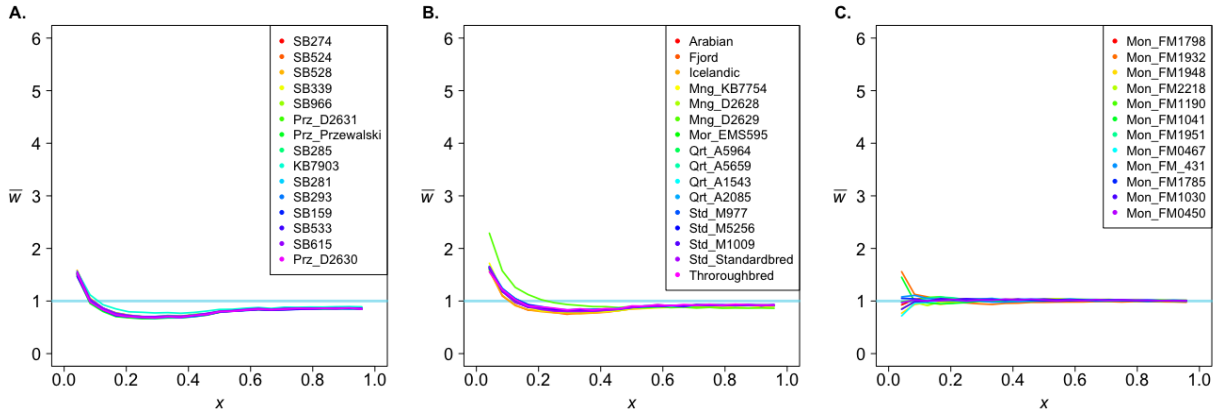


Figure 5.4: Projections of all genomes onto the Franches-Montagnes refFM reference panel. A. Przewalski's horse test genomes. B. Domesticated horse test genomes. C. Franches-Montagnes test genomes.

domesticated horses (MMPV= 0.8068, sd= 0.0278). This shows that the other domesticated horses are more closely related to the Franches-Montagnes breed than the Przewalski's horses.

When the domesticated horse breeds are compared to refPRZ (Figure 5.2B and Figure 5.2C), there is little variation in the projection for all but Mng_D2629. The minimum projection value is quite low (MMPV= 0.6685). This may be lower than the reverse value (for Przewalski's horse genomes relative to refFM) because of a smaller effective population size for Przewalski's horses.

Testing demographic models for domesticated and Przewalski's horses

Though there is no method for parameter estimation developed using projections, comparison of the simulated projection for parameters estimated from other models to the observed projections can give greater insight into demographic history. Here, we focused on the demographic history of two sampled populations of Przewalski's horses and FranchesMontagnes. We used PSMC [40] estimates of past horse population size and the maximum likelihood estimates calculated in dadi from each of five population models described in greater detail in S5.1 in [56]. For each model, we determined the simulated projection using *fastsimcoal2* [13]. We compared these projections to the observed projection using the sum of least squares (LSS) score to assess how well each model recovers the projection. The observed projections were determined using SB159 as the test Przewalski's horse genome, and Mon_FM0450 as the test Franches-Montagnes genome. refPRZ and refFM were used as the reference panels.

For all models, simulated projections were not similar to the observed projections (Figure 5.5), with LSS values all greater than 0.8 (Table 5.2). The best-fit models were those taking into account past gene flow, in particular those considering asymmetric gene flow (Model

Models	refPRZ	refFM
A: No gene flow	3.4988	0.9024
B: Symmetric gene flow	1.9589	2.1137
C: Asymmetric gene flow	0.8347	3.0062
D: Gene flow from PRZ to FM	1.4468	2.5717
E: Gene flow from FM to PRZ	3.4925	0.8734

Table 5.2: Sum of least squares score (LSS) for the demographic models tested for Przewalski's and Franches-Montagnes horse genomes used either as test or reference.

C) in the Przewalski-to-domesticated horse direction (Model D), in accordance with the results of dadi analyses [56]. As in dadi analyses, Model C was the best of the demographic models tested, we next focused on modifying parameters in Model C to improve the fit to the observed projections.

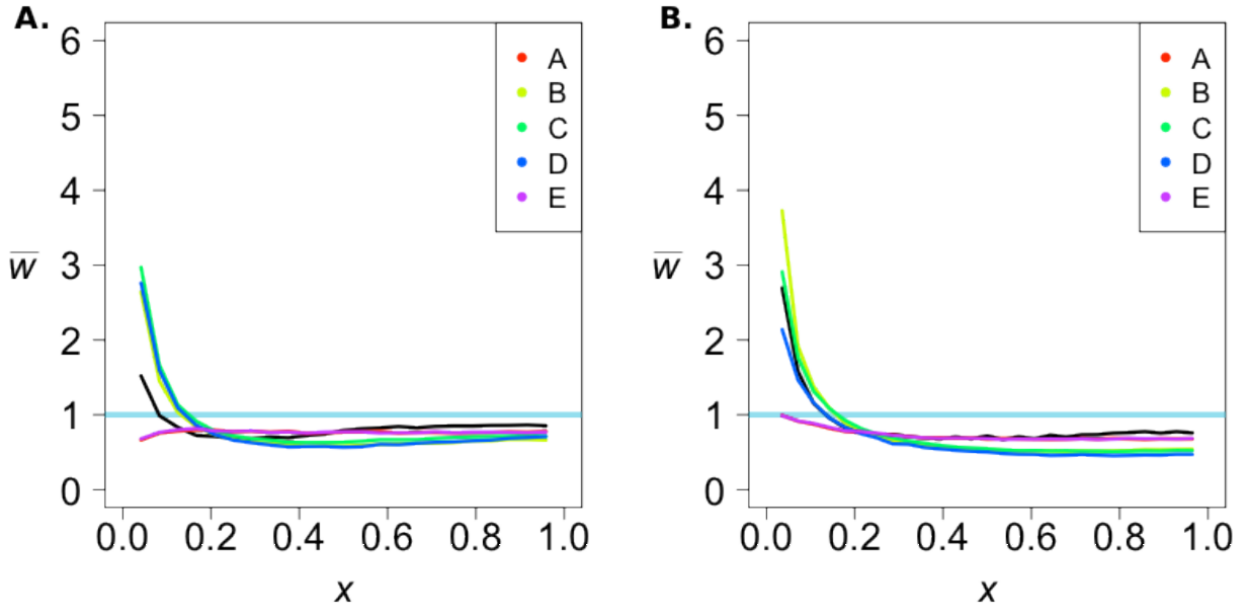


Figure 5.5: Observed and simulated projections for Franches-Montagnes and Przewalski's horses as either test or reference genomes. Simulated projections (colored according to key) were obtained for Models A-E as described in Figure 5.1. Observed projections are drawn in black. A. Projection of Przewalski's horse test genome onto Franches-Montagnes reference genome panel (refFM). B. Projection of Franches-Montagnes horse test genome onto Przewalski's horse reference genome panel (refPRZ).

Estimating the time of change in migration rate between domesticated and Przewalski's horses

Using the model that fits best according to dadi analyses (see Model C in S5.2 of [56]), we considered whether an improved fit of the model could be obtained considering changes in population decay and migration rate (Figure 5.6). We used *fastsimcoal2* [13], Brents algorithm and the LSS score to vary one model parameter at a time and improve the fit. Simulated parameters are reported and the LSS scores for the adjusted models C are reported in Table 5.3, with the projections shown in Figure 5.7. Four models were considered (Figure 5.6). Model C1 only considered a single asymmetric migration, the rate of which was estimated here. Model C2 changed the size of the exponential population decay and the effective population size of Przewalski's horse, allowing a single asymmetric migration rate as in Model C (Model C2_1) and allowing the asymmetric migration rate to vary (Model C2_2). Model C3 considered no population size changes but allowed two different asymmetric migration rates in the past and present, in order to test for potential consequences of past climatic changes. We allowed the migration rate to change once, keeping it constant before and after the time of migration rate change. Finally, Model C4 allowed both population size changes similar to that in Model C2 and two different asymmetric migration rates, plus a pulse of admixture from Franches-Montagnes to Przewalski's horses at some recent time in the past.

	Parameters								LSS	
Model	NPRZ	NANC	TMIG (years)	FFP (%)	MFP1 (x10^{-5})	MPF1 (x10^{-5})	MFP2 (x10^{-5})	MPF2 (x10^{-5})	refPRZ	refFM
C1	3,183	3,000	n/a	0	32.00	24.00	n/a	n/a	0.2887	0.4378
C2_1	4,500	6,000	n/a	0	0.54	4.16	n/a	n/a	0.4545	0.2180
C2_2	4,500	6,000	n/a	0	6.90	1.70	n/a	n/a	0.0586	0.0439
C3	3,183	3,000	23,200	0	45.00	90.00	34.00	0.43	0.3454	0.2487
C4	4,500	6,000	23,200	4	80.00	40.00	0.50	1.30	0.0657	0.0066

"LSS": Sum of least squares score; "NPRZ": effective population size of Przewalski's horse today; "NANC": effective population size of both horse populations 10,000 yBP; "MPF1": migration rate of Przewalski's horses into Franches-Montagnes horses before the change of migration rate; "MFP1": migration rate of Franches-Montagnes horses into Przewalski's horses before the change of migration rate; "MPF2": migration rate of Przewalski's horses into Franches-Montagnes horses after the change of migration rate; "MFP2": migration rate of Franches-Montagnes horses into Przewalski's horses after the change of migration rate; "TMIG": time for the change of migration rate. Migration rates are reported as the fraction of the sink population made up by new migrants. "FFP": recent migration pulse of Franches-Montagnes into Przewalski's horses 1,200 yBP.

Table 5.3: Demographic parameters included in tested models.

In Model C1, the simulated projections did not fit the observed projection (Table 5.3). No changes to the migration rate could recover the observed projections, partly because the simulated projections for the given effective population size and divergence time was closer to $\bar{w}(x) = 1$ than the observed projections (Figure 5.7). In Model C2, the best fit was obtained when the population size exponentially decayed to 6,000 rather than 3,000 at 10,000 yBP in both horse populations, and when the Przewalski's horse instantaneously changed to a present day effective population size of 4,500 rather than 3,183. While fixing the migration rate to that of Model C in Model C2 did not substantially improve the projection, when the migration rate was also allowed to vary, the simulated projection fitted well to the observed

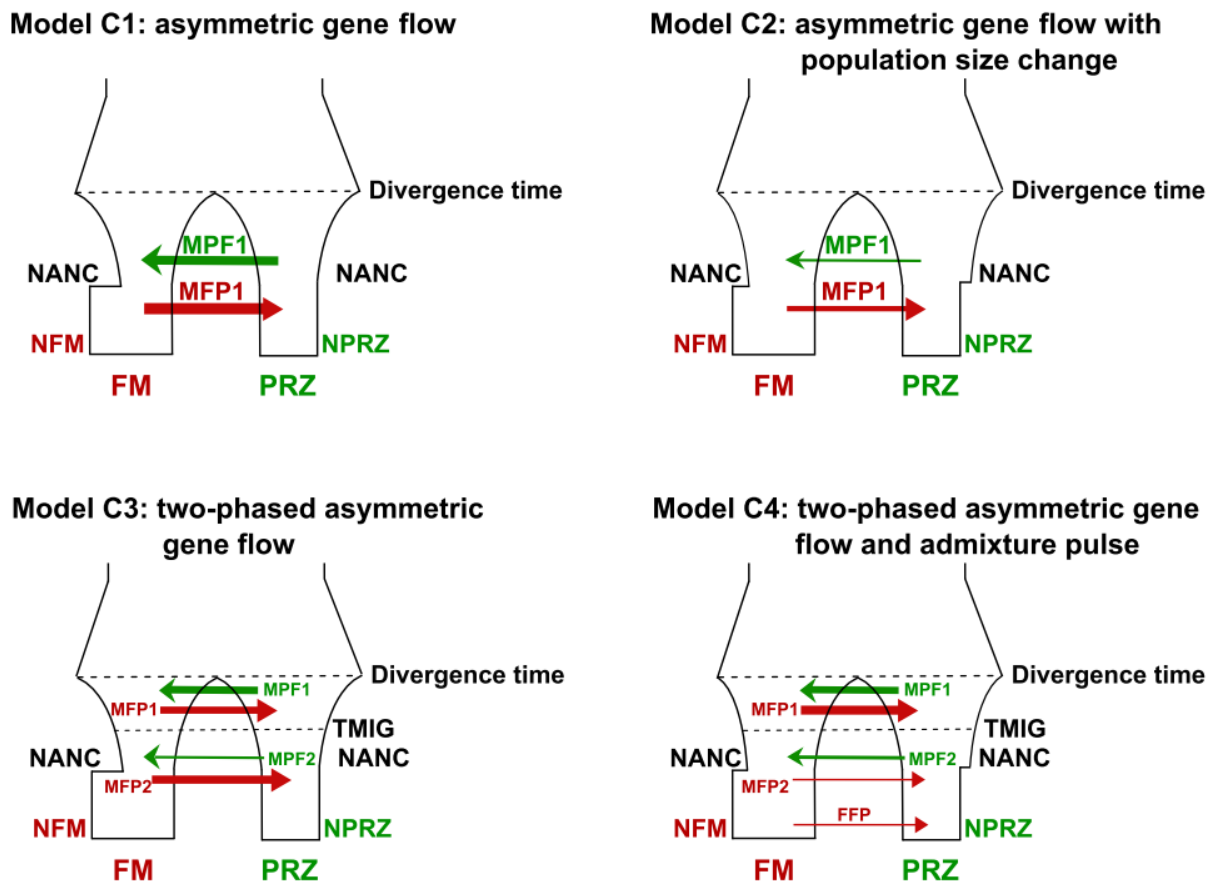


Figure 5.6: Demographic models tested in projection analyses (adapted from dadi model C). FM: Franches-Montagnes; PRZ: Przewalski's horses; NANC, effective population size at end of population size decay at 10,000 yBP; NPRZ: effective population size of Przewalski's horses; NFM: effective population size of Franches-Montagnes horses; MPF1: migration rate of Przewalski's horses into Franches-Montagnes horses (before the change of migration rate for model C3); MPF2: migration rate of Przewalski's horses into FranchesMontagnes horses after the change of migration rate for model C3; MFP1: migration rate of Franches-Montagnes horses into Przewalski's horses (before the change of migration rate for model C3); MFP2: migration rate of Franches-Montagnes horses into Przewalski's horses after the change of migration rate for model C3; TMIG, the time of migration rate change between MPF1/MFP1 and MPF2/MFP2; FFP: recent admixture fraction. Migration rates are reported as the fraction of the sink population made up by new migrants.

projection (Table 5.3). In Model C2_2, we find the migration rate from Franches-Montagnes to Przewalski's horses to be four times higher than the reverse, which contrasts with previous dadi analyses favouring models where the rate of migration is lower from domesticated to Przewalski's horses than the reverse [56]. This is largely due to the increase in rare alleles

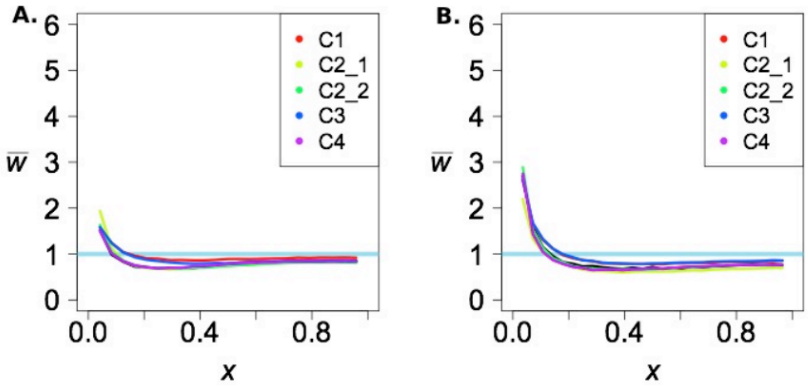


Figure 5.7: Observed and simulated projections for Franches-Montagnes and Przewalski's horses as either test or reference genomes. Simulated projections were obtained for the modified model C, C1-3. Observed projections are drawn in black. A. Projection of Przewalski's horse test genome onto Franches-Montagnes horse reference genome panel. B. Projection of Franches-Montagnes horse test genome onto Przewalski's horse reference genome panel.

observed in the projection onto refPRZ but that is not observed in the projection onto refFM.

In Model C3 also, no simulated projection was found to fit well to the observed projections without adjusting the population size (Table 5.3). However, the observed projections could be recovered when the population size changes used in Model C2 were combined with a change in the migration rate in Model C4 (Table 5.3). In particular, the observed projection onto refFM was very closely fitted by the simulated projection in Model C4 (Table 5.3).

In Model C4, a recent pulse of admixture was also included to improve the fit of the projection onto refPRZ. The high increase in rare alleles observed in this projection may reflect recent gene flow. We tested a range of recent times and amount of admixture, and found that the LSS for Model C4 substantially improved for an admixture rate of about 4% and a time of admixture of about 1,200 yBP (Figure 5.8). The best fitting time for recent admixture is found older than that expected based on admixture tests and on the knowledge of the captive history of the Przewalski's horses in the last 200 yBP. This could be due the assumption of random mating in Przewalski's horses being violated as the refPRZ panel is composed of 14 Przewalski's horse individuals, of which less than half show the signals of admixture.

By allowing the pulse of admixture and the change in migration rate, and by adjusting the past population size decay in both horse populations and the present day effective population size of Przewalski's horse, the observed projections could be recovered (Table 5.3). Furthermore, the more recent migration rate showed a pattern similar to that found in Model C, with the migration rate from Przewalski's horses to Franches-Montagnes 2.6-fold higher than the reverse. The more ancient gene flow from Franches-Montagnes to Przewalski's horses was 2.0 fold higher that of the reverse migration, with the migration rates 31- to 160-fold higher than in recent times (Table 5.3). Ancestral higher amounts of migration

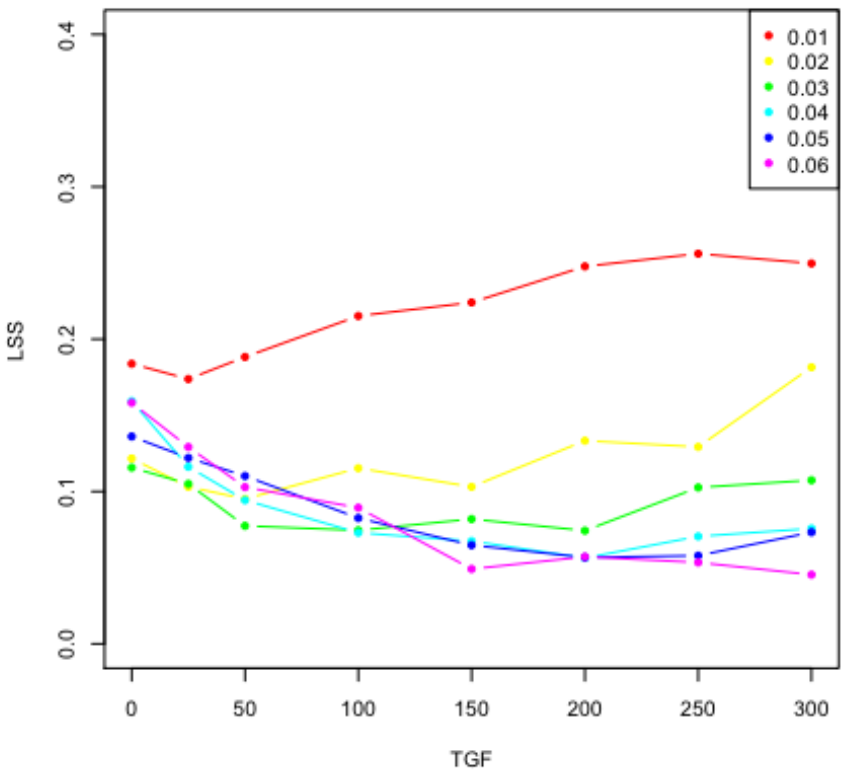


Figure 5.8: LSS scores for varying recent times of migration and amounts of admixture in Model C4, relative to observed projections on refPRZ. Different colors represent different amounts of admixture as indicated by the key. The x-axis indicates the time of recent admixture from Franches-Montagnes to Przewalski's horse and the y-axis indicates the LSS score for each pairing, where a lower LSS score indicates an improved fit of the simulated projection onto the observed projection.

further improved the fit of the projections (Model C2 to Model C4), particularly for the projection onto the Franches-Montagnes, suggesting that the two horse populations had greater contact in the past and grew more reproductively isolated towards the present. The time of migration rate change that fitted best varied for the two projections, but, for both, best fits were found around 19,000 yBP to 23,200 yBP, which coincides with the middle-end of the Last Glacial Maximum (Figure 5.9). These climatic changes likely had an impact on the level of contact between these two horse species.

The results of the projection analyses show some overlap with the results of the dadi demographic inferences [56] suggesting structured high gene flow between the ancestral populations of domesticated and Przewalski's horses followed by a drastic reduction in the migration rate after the LGM.

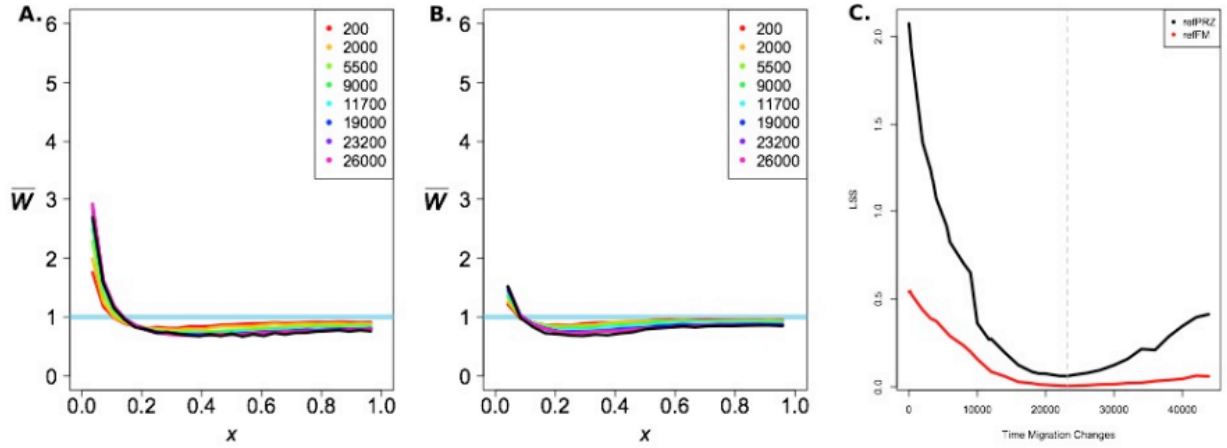


Figure 5.9: Estimation of the time of migration rate change, using projections for Franches-Montagnes and Przewalski's horses as either test or reference genomes. Observed (black) and simulated projections (color) allowing for different times of migration rates changing for Franches-Montagnes projected onto Przewalski's horse reference genome panel (A) and Przewalski's horse projected onto Franches-Montagnes horse reference genome panel (B). C. LSS score variation with times of migration rate for both projections. Times shown are in yBP and were taken from time periods described in the section Joint demographic inference using dadi as potentially affecting interbreeding opportunities. All other parameters were the ones best fitting Model C3.

5.3 Demographic history of the Yakut horse

We also analyzed Yakut horses and four ancient horses sampled from the same region, relative to the Przewalski's horse and present day domestic horse breeds. The Yakut horses are physiologically capable of living through extremely cold temperatures, but is little is known about their demographic history and how quickly they adapted to the cold climate. Here, we examine whether Yakut horses are descended from ancient horse populations sampled in the area, or whether they are more related to domestic horse breeds, using projection analysis as described in Chapter 3. For this analysis, we compared the genomes detailed in Table 5.4, and grouped them according to their corresponding population.

Test horse genomes and reference panels

We first considered the refDOM panel, which includes all of the 27 horse genomes from non-Yakutian domesticated breeds. By including a high number of individuals, this panel may provide greater resolution of the projection for rare alleles. However, it is composed of genomes from several horse breeds, which violates the random mating assumption made for the reference population. Therefore, to avoid the confounding effects originating from

Horse ID	Horse type	Reference	Reference panel(s)
Yak1	Yakutian	this study	refYAK
Yak2	Yakutian	this study	refYAK
Yak3	Yakutian	this study	refYAK
Yak4	Yakutian	this study	refYAK
Yak5	Yakutian	this study	refYAK
Yak6	Yakutian	this study	refYAK
Yak7	Yakutian	this study	refYAK
Yak8	Yakutian	this study	refYAK
Yak9	Yakutian	this study	refYAK
CGG10022	Ancient Taymyr peninsula	Orlando et al. 2013 (13)	n/a
CGG10023	Ancient Taymyr peninsula	Schubert et al. 2014 (16)	n/a
CGG101397	Ancient Yakutia	this study	n/a
Batagai	Ancient Yakutia	this study	n/a
Prz_D2630	Przewalski	Do et al. 2014 (19)	n/a
Prz_D2631	Przewalski	Do et al. 2014 (19)	n/a
Prz_Przewalski	Przewalski	Orlando et al. 2013 (13)	n/a
Arabian	Arabian	Orlando et al. 2013 (13)	refDOM
Fjord	Norwegian Fjord	Orlando et al. 2013 (13)	refDOM
Icelandic	Icelandic	Orlando et al. 2013 (13)	refDOM
Mng_D2628	Mongolian	Do et al. 2014 (19)	refDOM
Mng_D2629	Mongolian	Do et al. 2014 (19)	refDOM
Mor_EMS595	Morgan	this study	refDOM
Qrt_A5964	Quarter	this study	refDOM
Qrt_A5659	Quarter	this study	refDOM
Qrt_A1543	Quarter	this study	refDOM
Qrt_A2085	Quarter	this study	refDOM
Std_M977	Standardbred	this study	refDOM
Std_M5256	Standardbred	this study	refDOM
Std_M1009	Standardbred	this study	refDOM
Std_Standardbred	Standardbred	Orlando et al. 2013 (13)	refDOM
Thoroughbred	Thoroughbred	Wade et al. 2009 (29)	refDOM
Mon_FM1798	Franches-Montagnes	this study	refDOM, refFM
Mon_FM1932	Franches-Montagnes	this study	refDOM, refFM
Mon_FM1948	Franches-Montagnes	this study	refDOM, refFM
Mon_FM2218	Franches-Montagnes	this study	refDOM, refFM
Mon_FM1190	Franches-Montagnes	this study	refDOM, refFM
Mon_FM1041	Franches-Montagnes	this study	refDOM, refFM
Mon_FM1951	Franches-Montagnes	this study	refDOM, refFM
Mon_FM0467	Franches-Montagnes	this study	refDOM, refFM
Mon_FM_431	Franches-Montagnes	this study	refDOM, refFM
Mon_FM1785	Franches-Montagnes	this study	refDOM, refFM
Mon_FM1030	Franches-Montagnes	this study	refDOM, refFM
Mon_FM0450	Franches-Montagnes	this study	refDOM, refFM

Table 5.4: Horse genomes used for projection analyses.

the refDOM structure, we focused our subsequent projections onto refFM and refYAK populations. In panel refFM, we only considered a subset of refDOM, exclusively comprised of 12 Franches-Montagnes horses. Whereas panel refYAK grouped all of the nine modern Yakutian horse genomes characterized in this study.

Projection results

Projections of modern Yakutian horse genomes onto the reference panels, refYAK and refFM

Projections of modern Yakutian horses onto refYAK mostly lie around $\bar{w}(x) = 1$ (Figure 5.10), which corresponds to the expectation of a random mating population, including a slight departure for rare alleles (110). For three genomes, namely Yak1, Yak3 and notably Yak2, the projections rose above the $\bar{w}(x) = 1$ line for alleles found at a low frequency, suggesting they experienced reduced levels of admixture. All nine Yakutian horse genomes have similar projections relative to refFM, with minimum projection values (MPV) below one (mean MPV = 0.7724), and a standard deviation (sd) of 0.0092 (Table 5.5), suggesting that modern Yakutian and FM horses do not belong to the same breed.

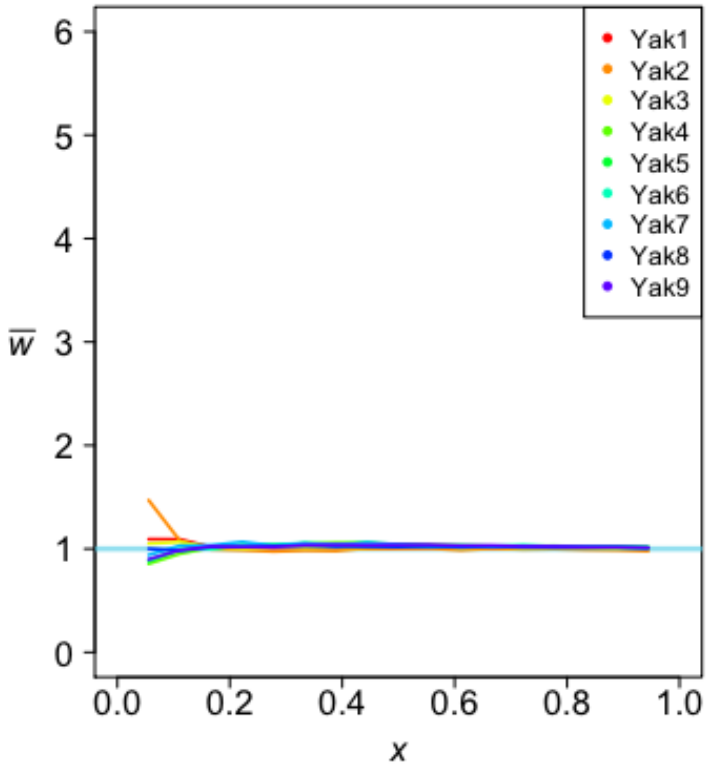


Figure 5.10: Projections of modern Yakutian horses onto the refYAK panel. The x-axis represents the categories of derived allele frequencies, while $\bar{w}$ represents the corresponding projection ($\bar{w}(x) > 1$ indicates that the test genome has more alleles at that frequency than the reference panel, and vice versa).

Projections of Przewalski's and Yakutian horse genomes onto refFM

When projected onto refFM, the mean MPV is lower for Przewalski's horses (mean MPV = 0.6762; $sd = 0.0109$) than for modern Yakutian horses (mean MPV = 0.7724; $sd = 0.0092$; Table 5.5), indicating that Franches-Montagnes horses are more closely related to modern Yakutian horses than to the Przewalski's horses (Figure 5.11 and 5.12). This is consistent with the phylogenetic position of modern Yakutian horses within the domesticated horse clade, and outside the monophyletic group of Przewalski's horses [41].

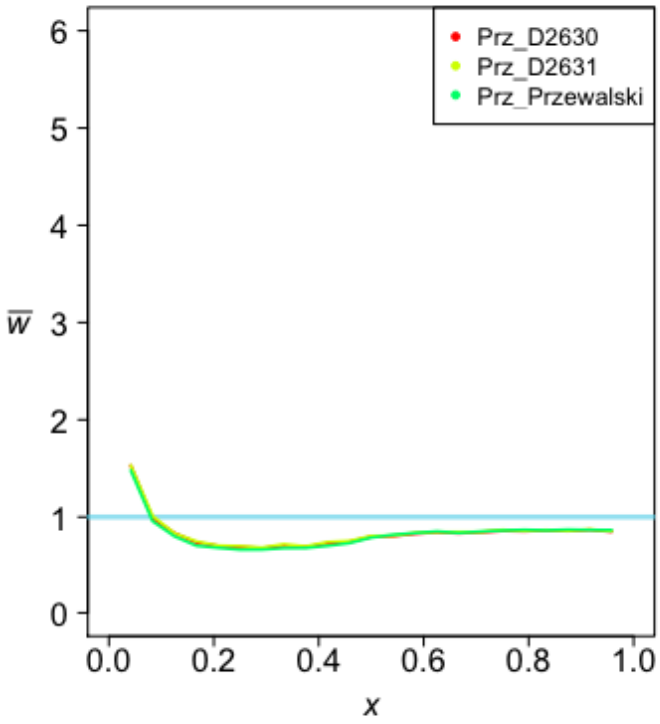


Figure 5.11: Projections of Przewalski's horses onto the refFM panel. See Figure 5.10 for captions.

Projections of non-Yakutian modern horse genomes onto refYAK

Further evidence that the modern Yakutian horses belong to the clade of domesticated horses is provided by the reciprocal projections onto refYAK (Figure 5.13), which report relatively high MPVs when testing either the Franches-Montagnes horse genomes (mean MPV = 0.8038; $sd = 0.0062$) or non-Franches-Montagnes domesticated horses (MPV = 0.8073; $sd = 0.0246$) (Table 5.5). These MPVs are higher than the MPVs observed when projecting the Przewalski's horse genomes onto refYAK (mean MPV = 0.6926; $sd = 0.0009$; Table 5.5),

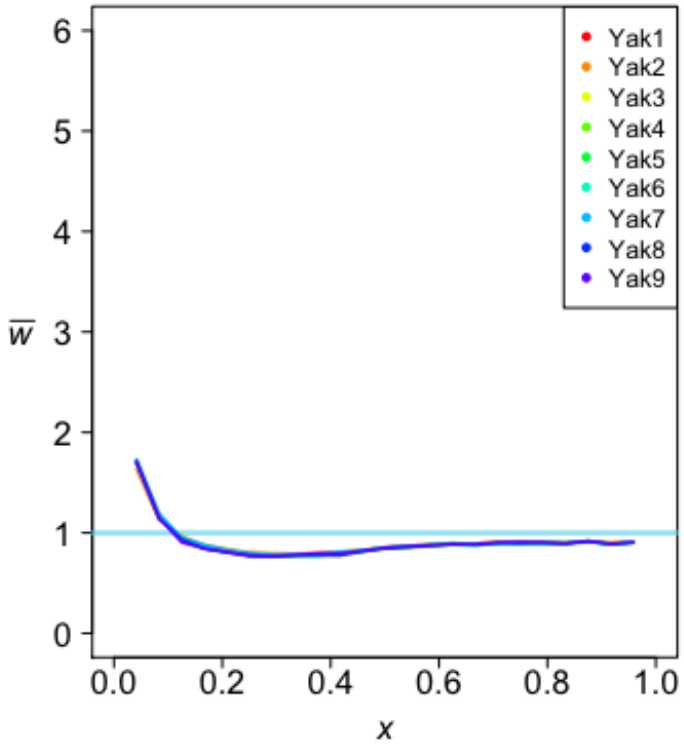


Figure 5.12: Projections of modern Yakutian horses onto the refFM panel. See Figure 5.10 for captions.

	refYAK		refFM	
Test genome(s)	**MPV**	**sd**	**MPV**	**sd**
Yakutian horses	*	*	0.7724	0.0092
Przewalski's horses	0.6926	0.0009	0.6762	0.0109
Non-Franches-Montagnes	0.8073	0.0246	0.8088	0.0283
Franches-Montagnes	0.8038	0.0062	*	*
CGG101397	0.8719	n/a	0.7637	n/a
Batagai	0.5967	n/a	0.5943	n/a
CGG10022	0.6067	n/a	0.6010	n/a
CGG10023	0.5898	n/a	0.5913	n/a
Ancient horses (excl. CGG101397)	0.5977	0.0085	0.5955	0.0050

"MPV": Minimum Projection Value; "sd": standard deviation.

Table 5.5: Minimum projection values.

in line with their early phylogenetic divergence [41]. Although the two mean MPVs obtained when testing Franches-Montagnes and non-Franches-Montagnes are comparable, the standard deviation is about four fold larger for non-Franches-Montagnes, nicely reflecting their

heterogeneous genetic background, consisting of a mixture from eight different domesticated breeds.

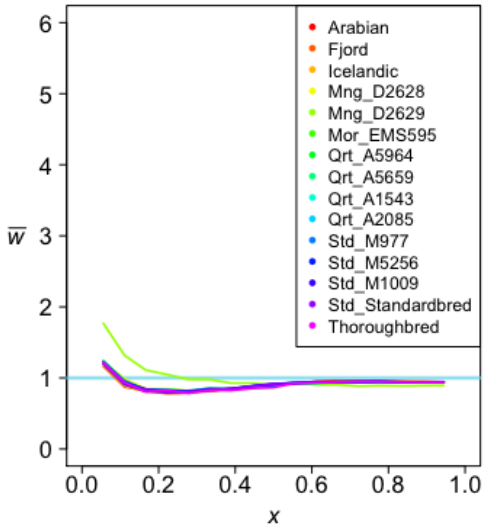


Figure 5.13: Projections of non-Yakutian and non-Franches-Montagnes modern horses onto the refYAK panel. See Figure S6.2 for captions.

Projections of ancient horse genomes onto the reference panels, refYAK and refFM

Projections of test ancient genomes onto refYAK show that CGG101397 on the one hand, and the other three surveyed ancient horses on the other hand (namely, CGG10022, CGG10023 and Batagai), have very different demographic histories (Figure 5.14). The genome of specimen CGG101397 shows mean MPVs of 0.8719 and 0.7637, when projected onto refYAK and refFM panels, respectively (Table 5.5). This suggests a closer relationship to modern Yakutian horses than to the Franches-Montagnes horses. These findings are in agreement with the results from phylogenetic analyses [41], where CGG101397 clustered within the modern Yakutian diversity, supporting genetic discontinuity in the horse population of Yakutia, with the ancient population represented by sample Batagai (5.2 kyr sample) being replaced by the population of present-day domesticated horses.

The projections of the CGG10022, CGG10023 and Batagai ancient horse genomes onto refYAK (mean MPV = 0.5977; $sd = 0.0085$) and refFM (mean MPV = 0.5955; $sd = 0.0050$) are very similar. The extremely reduced variance observed in the projections of these three genomes further suggests that they share a similar demographic history, despite spanning a 40 kyr-long temporal range. The mean MPV obtained for the projections of the CGG10022, CGG10023 and Batagai genomes onto refYAK (mean MPV = 0.5977) is much lower than

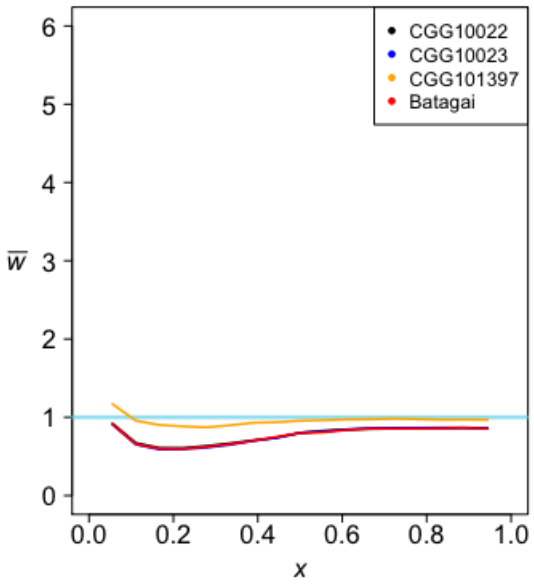


Figure 5.14: Projections of ancient horse genomes onto the refYAK panel. See Figure 5.10 for captions.

those obtained for the modern domesticated horses (mean MPV = 0.8038-0.8073), and even than those observed when projecting Przewalski's horses (mean MPV = 0.6926) (Table 5.5). This, again, supports that Late Pleistocene horses and the Batagai sample diverged prior to the most recent common ancestor of Przewalski's, Yakutian (including CGG101397), and other domesticated horses.

5.4 Conclusion

Using projection analysis, we showed that the Przewalski's horse and domesticated horse breeds likely had a period of high gene flow after divergence, followed by very low gene flow after around 23,200 yBP, roughly around the time of the Last Glacial Maximum. Furthermore, admixture into the Przewalski's horses sampled could also be captured using projection analysis. Furthermore, we showed that Yakut horses are more closely related to domesticated horse breeds than the Przewalski's horse and three of the four ancient horses sampled. Thus, they have adapted to extremely cold temperatures in just a few thousand years. These results are in agreement with other demographic analyses conducted on these data [56, 41], lending strong support to the demographic history described in this chapter.

Appendix A

Deriving the conditional derived *sfs* using coalescent theory

The conditional derived *sfs* in two closely related populations was derived by Chen et al. [8] and shown to be uniform using diffusion theory. Below, we derive this result again using coalescent theory.

Joint Site Frequency Spectrum in Two Closely Related Populations

Assume we have sampled n chromosomes from P_2 and one from P_1. Here we derive the conditional derived *sfs* in P_2, which is the *sfs* in P_2 given that the chromosome in P_1 carries the derived allele [8]. The two populations are assumed to split at time t_1, and there is no gene flow between them after they split. At time t_1, there are k lineages from P_2 and one from P_1. The probability distribution of k is given by Tavaré [66]:

$$\text{Prob}(k|n,t_1) = \frac{t}{\binom{k}{2}} \sum_{i=k}^{n} \binom{i}{2} e^{-\binom{i}{2}t_1} \prod_{j=k, j\neq i}^{n} \frac{\binom{j}{2}}{\binom{j}{2}-\binom{i}{2}}. \tag{A.1}$$

We further assume that the ancestral population of P_1 and P_2 was at mutation-drift equilibrium. Let m denote the number of ancestral lineages that carry the derived allele. We have

$$\text{Prob}(m) = \frac{\theta}{m}, \tag{A.2}$$

the equilibrium neutral spectrum. Thus, the joint probability that $j = m1$ lineages from P_2 and the lineage from P_1 carry the derived allele is

$$\text{Prob}(j, P_1 - \text{der}) = \frac{\theta}{m} \times \frac{m}{k+1} = \frac{\theta}{k+1}. \tag{A.3}$$

We denote H the hypergeometric distribution. $H(N, M, T)(K)$ is the probability to obtain K white balls when drawing T balls from an urn containing N balls, M of them being white. The probability that i lineages in P_2 today carry the derived allele, given j, k, and n is given by $\frac{k-1}{n-1} \times H(n-2, i-1, k-2)(j-1)$ [62]. Therefore, the joint probability that i lineages in P_2 and one lineage in P_1 carry the derived allele today is

$$\text{Prob}(i, P_1 - \text{der}|j, k, n) = \frac{\theta}{k+1} \frac{\binom{n-i-1}{k-j-1}\binom{i-1}{j-1}}{\binom{n-1}{k-1}}. \tag{A.4}$$

Averaging over j, we obtain

$$\text{Prob}(i, P_1 - \text{der}|k, n) = \frac{\theta}{k+1} \frac{k-1}{n-1}. \tag{A.5}$$

Therefore, the conditional derived spectrum is uniform on i. The exact density is obtained by averaging over the distribution of k.

Appendix B

ms simulations used in Chapter 2

The two main demographic models considered are a model of recent admixture and a model of ancient structure. The *ms* command for a model of recent admixture is of the form:

ms 12 1000000 -t θ -I 3 1 1 10 n 2 g n 3 g -m 3 2 $4Nm$ -m 2 3 $4Nm$ -es t_{GF} 3 1 -f -ej t_{GF} 4 1 -en t_b 3 bg -en $t_b + 0.025$ 3 g -ej t_H 3 2 en t_g 2 1 -ej t_N 2 1 (bottleneck older than time of admixture)

ms 12 1000000 -t θ -I 3 1 1 10 n 2 g n 3 g -m 3 2 $4Nm$ -m 2 3 $4Nm$ -en t_b 3 bg -en $t_b + 0.025$ 3 g -es t_{GF} 3 1 -f -ej t_{GF} 4 1 -ej t_H 3 2 en t_g 2 1 -ej t_N 2 1 (bottleneck younger than time of admixture)

For example, ms 12 1000000 -t 20 -I 3 1 1 10 n 2 100 n 3 100 -m 3 2 5 -m 2 3 5 -es 0.05 3 1 -f -ej 0.05 4 1 -en 0.1 3 1 -en 0.1025 3 100 -ej 0.1125 3 2 en 0.1150 2 1 -ej 0.3 2 1 (bottleneck older than time of admixture) means three populations were sampled, with recent symmetric gene flow of $4Nm = 5$ between populations 2 and 3. Populations 2 and 3 are 100 times the effective population size. A 10% admixture event occurred at time 0.05 from population 1 to population 3. A bottleneck older than the time of admixture occurred at time 0.1, reducing the population 100-fold. Populations 3 and 2 coalesce at time 0.1125, and a sudden expansion in the ancestral population occurred at time 0.1150 from the original effective population size. They and population 1 coalesce at time 0.3.

The ms command for a model of ancient structure is of the form:

ms 12 1000000 -t θ -I 3 1 1 10 n 2 g n 3 g -m 3 2 $4Nm$ -m 2 3 $4Nm$ -en t_b 3 bg -en $t_b + 0.025$ 3 g -em t_H 3 2 $4Nm_a$ -em t_H 2 3 $4Nm_a$ en t_g 2 1 en t_g 3 1 -ej t_N 1 3 -ej t_s 3 2.

For example, ms 12 1000000 -t 20 -I 3 1 1 10 n 2 1 n 3 1 -m 3 2 0 -m 2 3 0 -en 0.03 3 0.01 -en 0.0325 3 1 -em 0.1125 3 2 6 -em 0.1125 2 3 6 en 0.1150 2 1 en 0.1150 3 1 -ej 0.3 1 3 -ej 0.8 3 2 means three populations were sampled, with no recent gene flow and a population bottleneck occurring at time 0.03 reducing the population 100-fold. Ancient symmetric migration between Africans and Europeans occurred in the past starting at 0.1125, with a high gene flow of $4Nm_a = 6$, and Europeans and Neanderthals coalesce at time 0.3. Ancient gene flow ends at time 0.8. No pop- ulation growth occurred.

For no admixture, $f = 0$. For no gene flow, $4Nm = 0$, and for no bottleneck, $b = 1$. For no population growth, $g = 1$. The full range of parameters explored for each model can be

found in Table 2.1.

Bibliography

[1] Goncalo R. Abecasis et al. "A map of human genome variation from population-scale sequencing". In: *Nature* 467.7319 (2010), pp. 1061–73. DOI: 10.1038/nature09534.

[2] Adam Auton et al. "A global reference for human genetic variation". In: *Nature* 526.7571 (2015), pp. 68–74. DOI: 10.1038/nature15393.

[3] L. B. Barreiro et al. "The heritage of pathogen pressures and ancient demography in the human innate-immunity CD209/CD209L region". In: *American Journal of Human Genetics* 77.5 (2005), pp. 869–886. DOI: 10.1086/497613.

[4] P. Beerli. "Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations". In: *Molecular Ecology* 13.4 (2004), pp. 827–836. DOI: 10.1111/j.1365-294X.2004.02101.x.

[5] J. M. Braverman et al. "The hitchhiking effect on the site frequency-spectrum of DNA polymorphisms". In: *Genetics* 140.2 (1995), pp. 783–796.

[6] Adrian W. Briggs et al. "Patterns of damage in genomic DNA sequences from a Neandertal". In: *Proceedings of the National Academy of Sciences of the United States of America* 104.37 (2007), pp. 14616–14621. DOI: 10.1073/pnas.0704665104.

[7] Michael C. Campbell and Sarah A. Tishkoff. "African genetic diversity: Implications for human demographic history, modern human origins, and complex disease mapping". In: *Annual Review of Genomics and Human Genetics* 9 (2008), pp. 403–433. DOI: 10.1146/annurev.genom.9.081307.164258.

[8] Hua Chen et al. "The joint allele-frequency spectrum in closely related species". In: *Genetics* 177.1 (2007), pp. 387–398. DOI: 10.1534/genetics.107.070730.

[9] Radoje Drmanac et al. "Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays". In: *Science* 327.5961 (2010), pp. 78–81. DOI: 10.1126/science.1181498.

[10] C. Duarte et al. "The early Upper Paleolithic human skeleton from the Abrigo do Lagar Velho (Portugal) and modern human emergence in Iberia". In: *Proceedings of the National Academy of Sciences of the United States of America* 96.13 (1999), pp. 7604–7609. DOI: 10.1073/pnas.96.13.7604.

[11] Eric Y. Durand et al. "Testing for Ancient Admixture between Closely Related Populations". In: *Molecular Biology and Evolution* 28.8 (2011), pp. 2239–2252. DOI: 10.1093/molbev/msr048.

[12] Steven N. Evans, Yelena Shvets, and Montgomery Slatkin. "Non-equilibrium theory of the allele frequency spectrum". In: *Theoretical Population Biology* 71.1 (2007), pp. 109–119. DOI: 10.1016/j.tbp.2006.06.005.

[13] Laurent Excoffier et al. "Robust Demographic Inference from Genomic and SNP Data". In: *Plos Genetics* 9.10 (2013). DOI: 10.1371/journal.pgen.1003905.

[14] J. C. Fay, G. J. Wyckoff, and C. I. Wu. "Testing the neutral theory of molecular evolution with genomic data from Drosophila". In: *Nature* 415.6875 (2002), pp. 1024–1026. DOI: 10.1038/4151024a.

[15] Qiaomei Fu et al. "DNA analysis of an early modern human from Tianyuan Cave, China". In: *Proceedings of the National Academy of Sciences of the United States of America* 110.6 (2013), pp. 2223–2227. DOI: 10.1073/pnas.1221359110.

[16] Qiaomei Fu et al. "Genome sequence of a 45,000-year-old modern human from western Siberia". In: *Nature* 514.7523 (2014), pp. 445–+. DOI: 10.1038/nature13810.

[17] Y. X. Fu. "Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection". In: *Genetics* 147.2 (1997), pp. 915–925.

[18] M. Gallego Llorente et al. "Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent". In: *Science* 350.6262 (2015), pp. 820–2. DOI: 10.1126/science.aad2879.

[19] Simon Gravel et al. "Demographic history and rare allele sharing among human populations". In: *Proceedings of the National Academy of Sciences of the United States of America* 108.29 (2011), pp. 11983–11988. DOI: 10.1073/pnas.1019276108.

[20] Richard E. Green et al. "A complete neandertal mitochondrial genome sequence determined by high-throughput Sequencing". In: *Cell* 134.3 (2008), pp. 416–426. DOI: 10.1016/j.cell.2008.06.021.

[21] Richard E. Green et al. "A Draft Sequence of the Neandertal Genome". In: *Science* 328.5979 (2010), pp. 710–722. DOI: 10.1126/science.1188021.

[22] Richard E. Green et al. "Analysis of one million base pairs of Neanderthal DNA". In: *Nature* 444.7117 (2006), pp. 330–336. DOI: 10.1038/nature05336.

[23] R. C. Griffiths. "The frequency spectrum of a mutation, and its age, in a general diffusion model". In: *Theoretical Population Biology* 64.2 (2003), pp. 241–251. DOI: 10.1016/s0040-5809(03)00075-3.

[24] Ryan N. Gutenkunst et al. "Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data". In: *Plos Genetics* 5.10 (2009). DOI: 10.1371/journal.pgen.1000695.

[25] Wolfgang Haak et al. "Ancient DNA from European Early Neolithic Farmers Reveals Their Near Eastern Affinities". In: *Plos Biology* 8.11 (2010). DOI: 10.1371/journal.pbio.1000536.

[26] Wolfgang Haak et al. "Massive migration from the steppe was a source for Indo-European languages in Europe". In: *Nature* 522.7555 (2015), pp. 207–+. DOI: 10.1038/nature14317.

[27] Kelley Harris and Rasmus Nielsen. "Inferring Demographic History from a Spectrum of Shared Haplotype Lengths". In: *Plos Genetics* 9.6 (2013). DOI: 10.1371/journal.pgen.1003521.

[28] T. Hayakawa et al. "Fixation of the human-specific CMP-N-acetylneuraminic acid hydroxylase pseudogene and implications of haplotype diversity for human evolution". In: *Genetics* 172.2 (2006), pp. 1139–1146. DOI: 10.1534/genetics.105.046995.

[29] Ryan D. Hernandez, Scott H. Williamson, and Carlos D. Bustamante. "Context dependence, ancestral misidentification, and spurious signatures of natural selection". In: *Molecular Biology and Evolution* 24.8 (2007), pp. 1792–1800. DOI: 10.1093/molbev/msm108.

[30] R. R. Hudson. "Generating samples under a Wright-Fisher neutral model of genetic variation". In: *Bioinformatics* 18.2 (2002), pp. 337–338. DOI: 10.1093/bioinformatics/18.2.337.

[31] Philip L. F. Johnson and Montgomery Slatkin. "Accounting for bias from sequencing error in population genetic estimates". In: *Molecular Biology and Evolution* 25.1 (2008), pp. 199–206. DOI: 10.1093/molbev/msm239.

[32] Hakon Jónsson et al. "Speciation with gene flow in equids despite extensive chromosomal plasticity". In: *Proceedings of the National Academy of Sciences of the United States of America* 111.52 (2014), pp. 18655–18660. DOI: 10.1073/pnas.1412627111.

[33] Andreas Keller et al. "New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing". In: *Nature Communications* 3 (2012). DOI: 10.1038/ncomms1701.

[34] Y. Kim. "Allele frequency distribution under recurrent selective sweeps". In: *Genetics* 172.3 (2006), pp. 1967–1978. DOI: 10.1534/genetics.105.048447.

[35] Y. Kim and W. Stephan. "Detecting a local signature of genetic hitchhiking along a recombining chromosome". In: *Genetics* 160.2 (2002), pp. 765–777.

[36] Johannes Krause et al. "Neanderthals in central Asia and Siberia". In: *Nature* 449.7164 (2007), pp. 902–904. DOI: 10.1038/nature06193.

[37] M. Krings et al. "Neandertal DNA sequences and the origin of modern humans". In: *Cell* 90.1 (1997), pp. 19–30. DOI: 10.1016/s0092-8674(00)80310-4.

[38] Joseph Lachance et al. "Evolutionary History and Adaptation from High-Coverage Whole-Genome Sequences of Diverse African Hunter-Gatherers". In: *Cell* 150.3 (2012), pp. 457–469. DOI: 10.1016/j.cell.2012.07.009.

[39] Iosif Lazaridis et al. "Ancient human genomes suggest three ancestral populations for present-day Europeans". In: *Nature* 513.7518 (2014), pp. 409–+. DOI: 10.1038/nature13673.

[40] Heng Li and Richard Durbin. "Inference of human population history from individual whole-genome sequences". In: *Nature* 475.7357 (2011), 493–U84. DOI: 10.1038/nature10231.

[41] Pablo Librado et al. "Tracking the origins of Yakutian horses and the genetic basis for their fast adaptation to subarctic environments". In: *Proceedings of the National Academy of Sciences* (2015). DOI: 10.1073/pnas.1513696112.

[42] Molly E. McCue et al. "A High Density SNP Array for the Domestic Horse and Extant Perissodactyla: Utility for Association Mapping, Genetic Diversity, and Phylogeny Studies". In: *Plos Genetics* 8.1 (2012). DOI: 10.1371/journal.pgen.1002451.

[43] Matthias Meyer et al. "A High-Coverage Genome Sequence from an Archaic Denisovan Individual". In: *Science* 338.6104 (2012), pp. 222–226. DOI: 10.1126/science.1224344.

[44] Jose Luis Morales and Jorge Nocedal. "Remark on "Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound Constrained Optimization"". In: *Acm Transactions on Mathematical Software* 38.1 (2011). DOI: 10.1145/2049662.2049669.

[45] M. Nei, T. Maruyama, and R. Chakraborty. "Bottleneck effect and genetic-variability in populations". In: *Evolution* 29.1 (1975), pp. 1–10. DOI: 10.2307/2407137.

[46] M. Nordborg. "On the probability of Neanderthal ancestry". In: *American Journal of Human Genetics* 63.4 (1998), pp. 1237–1240. DOI: 10.1086/302052.

[47] Joseph K. Pickrell et al. "The genetic prehistory of southern Africa". In: *Nature Communications* 3 (2012). DOI: 10.1038/ncomms2140.

[48] Vincent Plagnol and Jeffrey D. Wall. "Possible ancestral structure in human populations". In: *Plos Genetics* 2.7 (2006), pp. 972–979. DOI: 10.1371/journal.pgen.0020105.

[49] Kay Prüfer et al. "The complete genome sequence of a Neanderthal from the Altai Mountains". In: *Nature* 505.7481 (2014), pp. 43–+. DOI: 10.1038/nature12886.

[50] Maanasa Raghavan et al. "Genomic evidence for the Pleistocene and recent population history of Native Americans". In: *Science* 349.6250 (2015).

[51] Maanasa Raghavan et al. "Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans". In: *Nature* 505.7481 (2014), pp. 87–+. DOI: 10.1038/nature12736.

[52] Morten Rasmussen et al. "An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia". In: *Science* 334.6052 (2011), pp. 94–98. DOI: 10.1126/science.1211177.

[53] Morten Rasmussen et al. "The genome of a Late Pleistocene human from a Clovis burial site in western Montana". In: *Nature* 506.7487 (2014), pp. 225–+. DOI: 10.1038/nature13025.

[54] David Reich et al. "Denisova Admixture and the First Modern Human Dispersals into Southeast Asia and Oceania". In: *American Journal of Human Genetics* 89.4 (2011), pp. 516–528. DOI: 10.1016/j.ajhg.2011.09.005.

[55] David Reich et al. "Genetic history of an archaic hominin group from Denisova Cave in Siberia". In: *Nature* 468.7327 (2010), pp. 1053–1060. DOI: 10.1038/nature09710.

[56] Clio Der Sarkissian et al. "Evolutionary Genomics and Conservation of the Endangered Przewalski's Horse". In: *Current Biology* 25.19 (2015), pp. 2577–2583.

[57] S. A. Sawyer and D. L. Hartl. "Population-genetics of polymorphism and divergence". In: *Genetics* 132.4 (1992), pp. 1161–1176.

[58] Stephan Schiffels and Richard Durbin. "Inferring human population size and separation history from multiple genome sequences". In: *Nature Genetics* 46.8 (2014), pp. 919–925. DOI: 10.1038/ng.3015.

[59] D. Serre et al. "No evidence of neandertal mtDNA contribution to early modern humans". In: *Plos Biology* 2.3 (2004), pp. 313–317. DOI: 10.1371/journal.pbio.0020057.

[60] Pontus Skoglund et al. "Genomic Diversity and Admixture Differs for Stone-Age Scandinavian Foragers and Farmers". In: *Science* 344.6185 (2014), pp. 747–750. DOI: 10.1126/science.1253448.

[61] Pontus Skoglund et al. "Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe". In: *Science* 336.6080 (2012), pp. 466–469. DOI: 10.1126/science.1216304.

[62] M. Slatkin. "Gene genealogies within mutant allelic classes". In: *Genetics* 143.1 (1996), pp. 579–587.

[63] Montgomery Slatkin and Joshua L. Pollack. "Subdivision in an ancestral species creates asymmetry in gene trees". In: *Molecular Biology and Evolution* 25.10 (2008), pp. 2241–2246. DOI: 10.1093/molbev/msn172.

[64] F. Tajima. "The effect of change in population-size on DNA polymorphism". In: *Genetics* 123.3 (1989), pp. 597–601.

[65] I. Tattersall and J. H. Schwartz. "Hominids and hybrids: The place of Neanderthals in human evolution". In: *Proceedings of the National Academy of Sciences of the United States of America* 96.13 (1999), pp. 7117–7119. DOI: 10.1073/pnas.96.13.7117.

[66] S. Tavare. "Line-of-descent and genealogical processes, and their applications in population-genetic models". In: *Theoretical Population Biology* 26.2 (1984), pp. 119–164. DOI: 10.1016/0040-5809(84)90027-3.

[67] Sarah A. Tishkoff et al. "The Genetic Structure and History of Africans and African Americans". In: *Science* 324.5930 (2009), pp. 1035–1044. DOI: 10.1126/science.1172257.

[68] J. Wakeley and J. Hey. "Estimating ancestral population parameters". In: *Genetics* 145.3 (1997), pp. 847–855.

[69] Jeffrey D. Wall et al. "Higher Levels of Neanderthal Ancestry in East Asians than in Europeans". In: *Genetics* 194.1 (2013), pp. 199–+. DOI: 10.1534/genetics.112.148213.

[70] Shuoguo Wang et al. "Apparent Variation in Neanderthal Admixture among African Populations is Consistent with Gene Flow from Non-African Populations". In: *Genome Biology and Evolution* 5.11 (2013), pp. 2075–2081. DOI: 10.1093/gbe/evt160.

[71] S. H. Williamson et al. "Simultaneous inference of selection and population growth from patterns of variation in the human genome". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.22 (2005), pp. 7882–7887. DOI: 10.1073/pnas.0502300102.

[72] Andreas Wollstein et al. "Demographic History of Oceania Inferred from Genome-wide Data". In: *Current Biology* 20.22 (2010), pp. 1983–1992. DOI: 10.1016/j.cub.2010.10.040.

[73] Melinda A. Yang, Kelley Harris, and Montgomery Slatkin. "The Projection of a Test Genome onto a Reference Population and Applications to Humans and Archaic Hominins". In: *Genetics* 198.4 (2014), pp. 1655–+. DOI: 10.1534/genetics.112.145359.

[74] Melinda A. Yang and Montgomery Slatkin. "Using Ancient Samples in Projection Analysis". In: *G3: Genes, Genetics, Genomics* (2015).

[75] Melinda A. Yang et al. "Ancient Structure in Africa Unlikely to Explain Neanderthal and Non-African Genetic Similarity". In: *Molecular Biology and Evolution* 29.10 (2012), pp. 2987–2995. DOI: 10.1093/molbev/mss117.